英伟达吸收 Groq 定义 AI 下半场！

华泰研究

2026 年 1 月 12 日 | 美国　　专题研究

Groq 交易是英伟达迄今披露的最大一笔交易，规模明显高于其 2019 年以 69 亿美元收购 Mellanox。我们认为，Groq 所掌握的低时延推理核心 IP 在战略层面的重要性，已与当年 Mellanox 的互连与网络技术处于同一量级。该交易进一步凸显英伟达对确定性、Batch Size = 1 推理的前瞻性布局，契合行业向 Agentic AI 演进的整体趋势。通过将 Groq 的确定性“反射式引擎”深度整合至 CUDA 与 GPU 技术栈，英伟达正加速推动 Agentic 经济走向主流，并在其已确立优势的 AI“上半场”基础上，逐步奠定低时延为核心特征的“下半场”的技术与规则框架。

Acqui-hire 模式锁定 Groq 的 LPU 人才与核心 IP

英伟达对价约 200 亿美元获得 Groq 推理技术的授权、收购部分知识产权，并引入 Groq 核心工程团队，包括创始人兼 CEO Jonathan Ross（原 TPU 架构师）与总裁 Sunny Madra。此次交易价格较 Groq 25 年 9 月最新私募融资估值的 69 亿美元隐含接近 3 倍溢价。从交易结构看，本次交易为 IP 授权叠加人才收购（Acqui-hire）的组合，而非完整的公司并购。GroqCloud 云服务将作为独立公司继续运营，由原 CFO Simon Edwards 出任 CEO。我们认为，这种“精准打击式”的交易结构，使英伟达能够在获取关键低时延推理 IP 的同时，有效规避整合硬件竞争对手带来的并购与监管不确定性。

英伟达以收购 Groq 定义 AI“下半场”规则

我们认为，该交易反映英伟达对 Agentic AI 时代需求结构变化的判断，即时延正成为继算力之后的关键约束因素。据 CNBC 报道，英伟达 CEO 黄仁勋在内部邮件中指出，此次交易的核心目标在于将 Groq 的低时延技术整合进英伟达的 AI 工厂。在此基础上，英伟达通过引入面向 Agentic AI 的低时延加速器，开始主动定义 AI“下半场”的技术标准。在 2025 年被普遍视为 Physical AI 元年之后，我们认为 2026 年有望成为 Agentic AI 元年，其核心特征在于，AI 工作负载将从以吞吐量为导向的训练阶段，转向为对时延高度敏感、执行过程具备确定性的实时应用阶段。我们认为，英伟达将把握这一关键时间节点，通过收购 Groq 为 Agentic 应用的规模化落地提供关键支撑，通过整合专用推理 IP 与其 CUDA 和 GPU 技术体系，英伟达得以在训练与实时推理两种核心范式下同时建立领先能力，并在一定程度上削弱云厂商依托自研芯片、从推理侧切入竞赛的潜在空间。

从 TPU 到 Dojo 与 Groq，计算架构趋同下的战略分化

我们认为，Groq、Tesla Dojo 以及谷歌 TPU 在底层均继承张量加速器的共同技术基因，但三者围绕 AI 工作负载的不同侧重点差异化设计。尽管 Dojo 与 Groq 同样依赖大规模片上 SRAM 与紧耦合的 Scale-up 互连，Tesla 选择将这一架构优势主要投向大规模、高吞吐的 FSD 训练场景，而非更适合发挥其低时延潜力的 Batch Size = 1 推理场景。相比之下，谷歌 TPU 虽起源于 Jonathan Ross 主导的“以推理为先”的设计理念，但其路线已演进为以 HBM 与 OCS 为核心的 Pod 级吞吐引擎，用于支撑大模型训练与推理。在体系结构上，TPU 与 Groq 的 Mega-Chip 理念存在呼应，但面向批处理的范式不同。 我们认为，Groq 或延续并强化“推理优先”的设计理念，通过确定性调度与片上 SRAM 带宽的协同优化，重点覆盖 Agentic 时代的低时延、交互式推理场景，并形成对科技巨头自研加速器的差异化优势。

风险提示：技术落地缓慢、需求不及预期等。

科技　　增持（维持）

何翩翩　　研究员
SAC No. S0570523020002　　purdyho@htsc.com
SFC No. ASI353　　+(852) 3658 6000

重点推荐

股票名称	股票代码	目标价（当地币种）	投资评级
英伟达(NVIDIA)	NVDA US	280.00	买入

资料来源：华泰研究预测

正文目录

问题 1：Groq 是什么？其架构在 AI 发展中有何战略意义？

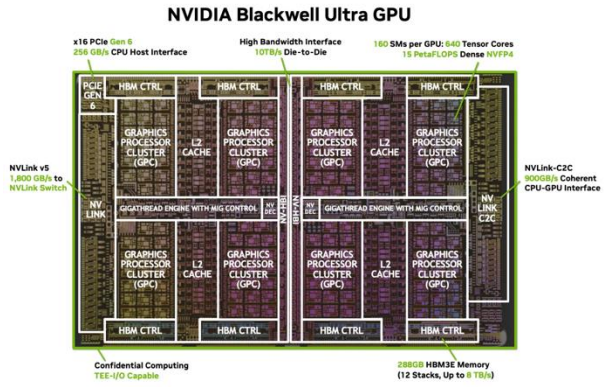
Groq 的核心产品是 Language Processing Unit（LPU），是面向推理计算阶段专门设计的 ASIC，其出发点并非追求更高的算力规模，而是解决通用 GPU 架构中长期存在的“时延-吞吐权衡（latency-throughput tradeoff）”问题。我们认为，Groq 本质上体现对交互式 Agentic AI 趋势未来主流化的押注：在这一趋势下，性能评价指标正从“每单位价格所能处理的总 token 数量”转向“单次请求的响应速度”。

与以训练和高吞吐批处理为核心优化目标的英伟达 GPU 不同，Groq 从设计开始即围绕实时、交互式推理场景进行设计，其核心价值主张在于 Determinism（确定性）。LPU 采用编译器驱动（compiler-driven）架构，在编译期对所有指令执行与内存访问进行预调度，从而消除动态调度所带来的不可预测的时延抖动（jitter）。本质上，Groq 以数学和逻辑可控的执行时序，取代传统硬件的概率性执行，从而压低 Batch Size = 1 场景下的“时延下限”。

我们认为，当前 AI 计算正在发生结构性分化，将逐步演化为以训练导向以及以部署导向的两条技术路径。其中，英伟达路线本质上是“吞吐优先”：依托大容量 HBM 与复杂的动态调度机制，最大化系统层面的批处理吞吐能力（即单位时间内处理的总 token 数量）。这一架构在模型训练及异步、批量推理场景中具备最优性。相对应地，Groq 路线则是“时延优先”：其目标客户为对“Time to First Token（第一个 token 的响应时间）”以及对 token 间时延高度敏感的实时、交互式的 Agentic AI 应用。通过移除动态硬件管理的系统开销，Groq 可实现小于 100ms 级的实时响应，满足自然人机交互对即时性的要求。

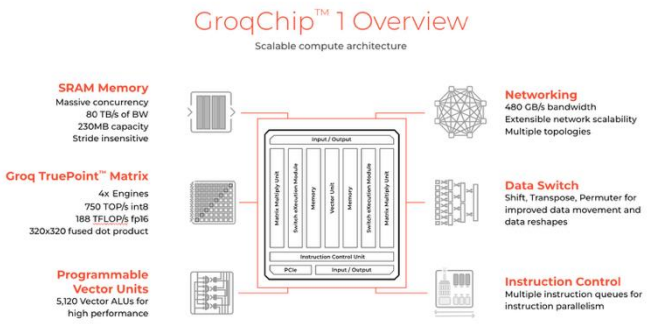
从产业分工角度看，我们认为 Groq 与英伟达并非替代关系，而是高度互补。Groq 更像是 AI 生命周期中推理阶段的专用计算层，服务于时延敏感型部署场景；而英伟达依旧是 AI 模型训练及高吞吐批量推理的通用标准，在大规模并行计算与内存密集型工作负载中具备不可替代的优势。我们认为英伟达架构在以超大内存容量与并行吞吐的场景中占据优势，而 Groq 正逐步成为时延敏感型推理的参考架构，为高性能的交互式部署提供支撑。

图表1： 英伟达 B300 搭载 288GB HBM3E



资料来源：英伟达官网，华泰研究

图表2： Groq LPU 搭载 230MB SRAM

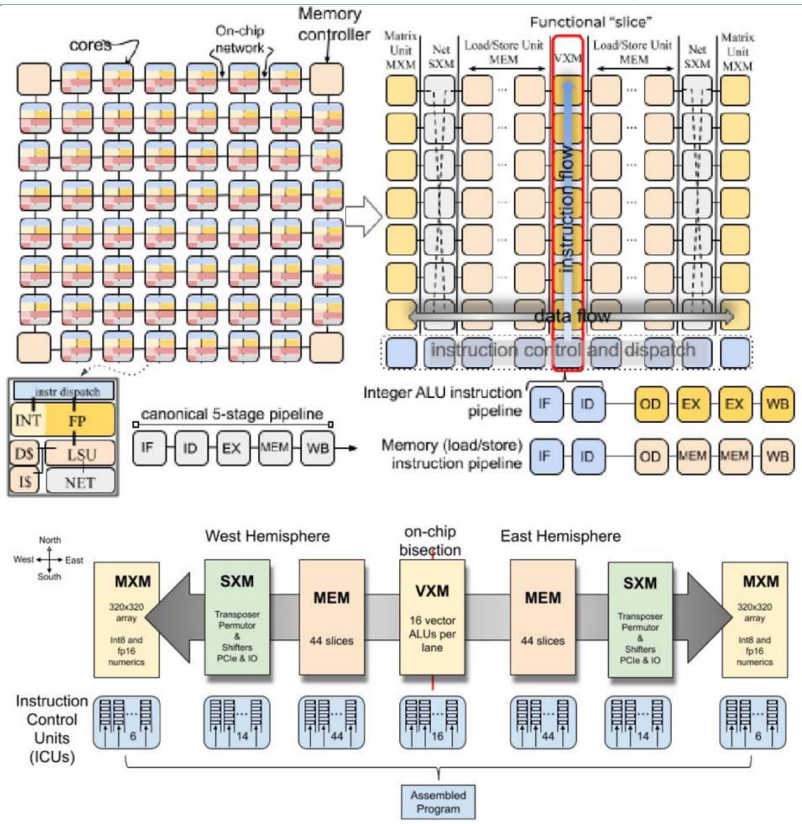


资料来源：Groq 官网，华泰研究

问题 2：Groq 架构如何区别于 GPU 范式，从而实现确定性的时延优势？

Groq 的性能优势源于其以编译器为先的设计理念，即将控制从硬件侧转移至软件侧。相较之下，英伟达 GPU 需要依赖运行时调度机制，在多任务并发过程中动态协调计算与存储资源。在 Groq 架构下，LPU 本质上仅负责严格执行预先生成的指令与访存计划，硬件层面不再引入缓存与动态仲裁等不确定性机制，从而有效消除运行时抖动（jitter）。基于这一确定性执行模型，Groq 构建可同步扩展的 Scale-up 计算域，最多可将 576 颗芯片整合为一个同步运行的单一逻辑处理器（Mega-Chip）。

图表3：LPU 的张量流处理器（TSP）的架构（右图）对比传统 GPU 采用多核布局（左图）



注：MXM 用于执行矩阵运算，SXM 用于向量的移位和旋转，MEM 用于内存读写，VXM 用于向量的运算；TSP 内部各切片（slice）之间的流式数据传输，数据流可以沿东西方向流动。

资料来源：Groq 官网，Groq 论文《Think Fast: A Tensor Streaming Processor (TSP) for Accelerating Deep Learning Workloads》，华泰研究

我们认为，Groq 的架构优势并非体现在“更快”的单一性能指标上，而是一种结构性差异。其核心取舍在于：主动放弃以 HBM 为核心、强调算力密度的 GPU 架构，转而采用以 SRAM 为核心的静态执行体系，以换取更低时延与更强的确定性。在以交互响应速度与一致性作为主要价值驱动的应用场景中，该取舍使 Groq 具备显著的系统级竞争优势。从架构层面看，Groq 相较传统 GPU 范式，主要体现在以下三项关键性的结构差异：

1）以 SRAM 为中心的存储架构（规避 HBM 瓶颈）

传统 GPU 普遍依赖外置 HBM 作为主存储，尽管具备较高容量（如 B300 约 288GB），其访问过程仍不可避免受到缓存未命中、内存控制器争用及刷新周期等因素影响，从而引入非确定性的时延抖动。相比之下，Groq 的 LPU 通过移除外部存储、在单芯片内集成约 230MB 高速 SRAM，将内存访问时延压缩至 10ns 以下，并实现 80TB/s 的确定性内存带宽，显著高于 HBM3E 约 8TB/s 的水平。该架构确保模型权重与激活数据可在计算所需时被精准、按时供给，从结构上削弱“内存墙”对推理场景的制约。但我们亦注意到，单颗 LPU 片上存储容量相对有限，大模型部署须依赖多芯片规模化扩展。例如，在 INT8 精度下部署一个 70B 参数模型（约需 70GB 内存），Groq 需配置约 576 颗芯片（系统通常由 8 个机架、每架 72 颗芯片构成）以满足 SRAM 容量需求。我们认为，这一显著的资本与系统规模投入，本质上反映以牺牲存储密度换取确定性低时延所需承担的成本。

2）编译期的确定性调度（“零抖动”模型）

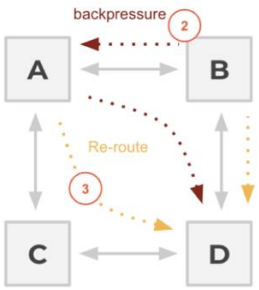
GPU 广泛采用运行时硬件调度机制（如 warp 调度器（warp schedulers,）、重排序缓冲区（reorder buffers）等），以在执行过程中动态管理数以千计的线程与指令流。当某一线程因等待 HBM 访问而阻塞时，调度器会切换至其他线程以提升整体吞吐率。该机制在高并发负载下有助于充分释放算力潜能，但也引入随机性的时延抖动。实际执行时间取决于运行时的缓存状态与资源争用情况。因此，在 Batch Size = 1 场景下，GPU 往往因内存时延与 kernel 启动开销而严重欠利用。

Groq 将系统控制权由硬件运行时调度前移至软件与编译阶段。其自研编译器 GroqWare 在模型部署前，对完整计算图进行静态解析与全局调度，提前确定每一条指令、每一次存储访问及数据传输在时序上的精确位置，从而消除运行时的不确定性（Zero Tail Latency）。在此基础上，Groq 实现严格的确定性执行特征，系统不存在长尾时延问题，P99 时延与中位时延基本一致。该能力在对话式智能体、实时推理等企业级应用场景中尤为关键：此类场景对响应一致性与时延可预测性要求极高，任何不可预期的卡顿都会直接影响用户体验。

3）软件定义的芯片互连（RealScale）

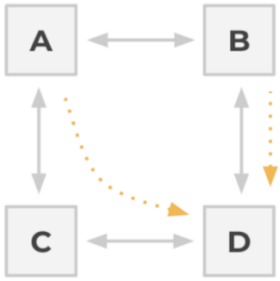
在 GPU 体系中，多卡扩展通常依赖 NVLink 或 InfiniBand 等网络互连方式，其底层仍涉及分组交换（packet switching）、握手（handshakes）等机制，因而不可避免地造成拥塞与不确定延迟。随着大模型参数持续扩大，单芯片已难以承载完整模型，而多 GPU 集群中的互联开销正逐步成为系统瓶颈。我们认为，Groq 的 RealScale 互连体系采用由编译器统一调度的芯片直连结构。由于编译器能够精确掌握数据在不同芯片间的发送与到达时间，系统可在无冲突、无缓冲的条件下完成数据传送。RealScale 使 Groq 能够在单一 Mega-Chip 中实现线性扩展，并协同多芯片系统同步运行。但我们认为，该同步系统的上限约为 576 颗芯片，超过该规模后仍需回退至标准以太网（Ethernet）互连。但在 576 颗芯片规模内，Groq 能够实现 GPU 架构难以达到的、低时延的甚至完全同步的并行推理。

图表4：传统非确定性网络架构



资料来源：Groq 官网，Groq 论文《A Software-defined Tensor Streaming Multiprocessor for Large-scale Machine Learning》，华泰研究

图表5：软件调度互联网络



资料来源：Groq 官网，Groq 论文《A Software-defined Tensor Streaming Multiprocessor for Large-scale Machine Learning》，华泰研究

图表6：Groq 的软件与编译生态

GroqChip 数量	峰值 INT8/FP16 性能	系统 SRAM（GB）	维度数量	网络直径（跳数）	端到端时延（μs）
1	750 TeraOps 189 TeraFlops	0.2	不适用	不适用	不适用
8 （1 个 GroqNode）	6 PetaOps 1.5 PetaFlops	1.76	0（单节点）	1	0.6
16	12 PetaOps 3 PetaFlops	3.5	1（2 个节点）	2	1.2
64 （1 个 GroqRack）	48 PetaOps 12 PetaFlops	14	1（8 个节点）	3	1.8

资料来源：Groq 官网，华泰研究

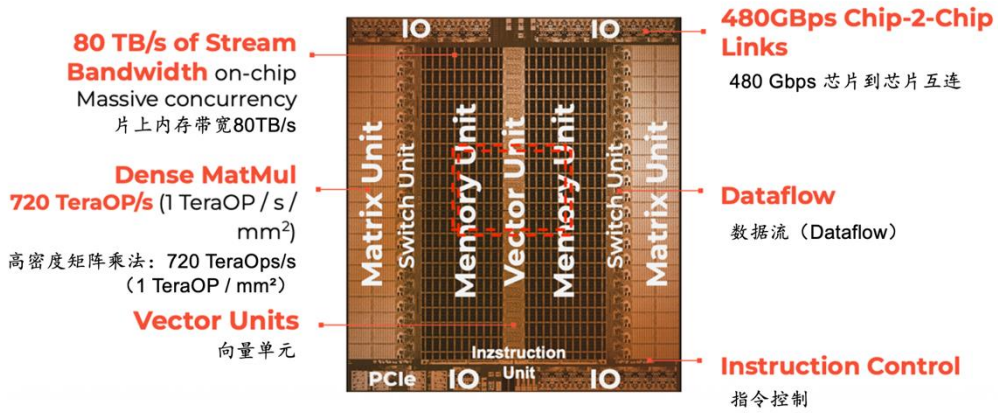
问题 3：Groq 的存储配置、互连（Scale-Up 与 Scale-Out）及软件架构如何支撑低时延推理？其设计选择在结构层面与英伟达 GPU 有何差异？

我们认为，Groq 的系统架构更接近一件为特定工作负载打造的“精密仪器”，其优化目标高度聚焦于对时延极度敏感的 Batch Size = 1 推理场景，并在由 576 颗芯片 Scale-up 的系统中表现最优。相比之下，英伟达 GPU 更接近于一套通用型算力引擎，目标是在不同规模、不同负载形态下最大化吞吐与容量，并依托成熟的软件生态，在 FP4/FP6/FP8 等硬件原生精度支持上具备更强的灵活性（如 B300 所体现的能力）。

1）存储配置：速度 vs. 容量间的取舍

我们认为，存储体系的结构性差异是 Groq 与 GPU 时延差距的最核心因素。Groq 采用 SRAM 设计，其 LPU 单芯片内集成约 230 MB 片上 SRAM，作为模型参数的主存储介质，存储带宽高达 80 TB/s。通过移除外置 HBM，Groq 避免任何跨芯片访存所带来的不可控时延，使权重访问可在<10 ns 的确定性窗口内完成；这一特性对于维持 Batch Size = 1 场景下的高利用率至关重要。相比之下，以 B300 为代表的英伟达 GPU 依赖 288 GB 外置 HBM3E，在提供较高容量的同时，其带宽规模约为 8 TB/s。GPU 的设计逻辑在于最大化容量密度（以更少芯片容纳更大模型），从而提升吞吐效率；而 Groq 则主动放弃内存容量以换取极低时延。这一取舍也意味着，Groq 在承载大模型时需通过多芯片系统（例如 576 颗芯片容纳一个 70B 模型），其扩展目的在于补足 SRAM 容量本身的物理限制。

图表7：GroqChip 可扩展架构



资料来源：Groq 官网，华泰研究

2）Scale-Up 互连：RealScale vs. NVLink

我们认为，若需要将多芯片组成一个同步的 Mega-Chip，需要一套高效的互联体系。Groq 采用 RealScale 互连，可最多支持 576 芯片（8 个 GroqRack）组成的同步系统。GroqWare 编译器将网络互联的收发单元视作“功能单元”，把数据传输编排在特定时钟周期内。由于数据传输计划在编译期已被完全确定，即便在跨数百芯片完成模型参数计算时，系统仍可维持亚微秒级时延（sub-microsecond latency）。相比之下，英伟达 B300 采用第五代 NVLink，单 GPU 提供约 1.8 TB/s 双向带宽。NVLink 的优势在于极高吞吐能力，但其调度依赖硬件仲裁机制，更适合大批量数据传输（尤其是训练场景）。从设计目标上看，NVLink 面向带宽优先的规模化计算，而 RealScale 则定位于时延更敏感的推理任务。

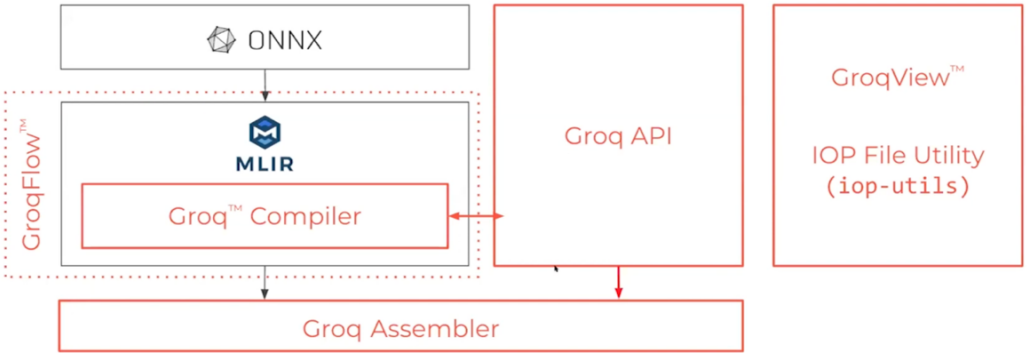
3）Scale-Out：Groq 的确定性扩展能力存在明确的物理边界（Determinism Cliff）

我们认为 Groq 的 Scale-Up 通常止步于 576 芯片互联；超过此规模，系统需退回至标准以太网进行扩展。我们认为，越过此“物理边界”后，Groq 不可避免地重新引入其原本试图规避的网络抖动和非确定性时延，限制其架构效率。相比之下，英伟达采用 InfiniBand 与 Spectrum-X 用于集群级扩展。以 B300 系统为例，其通过计算与通信重叠以及大规模批处理来容忍网络波动，从而在 Scale-out 的训练与批量推理工作负载中，以可接受的时延波动换取极高的吞吐能力。

4）软件生态与数值精度：GroqWare vs. CUDA

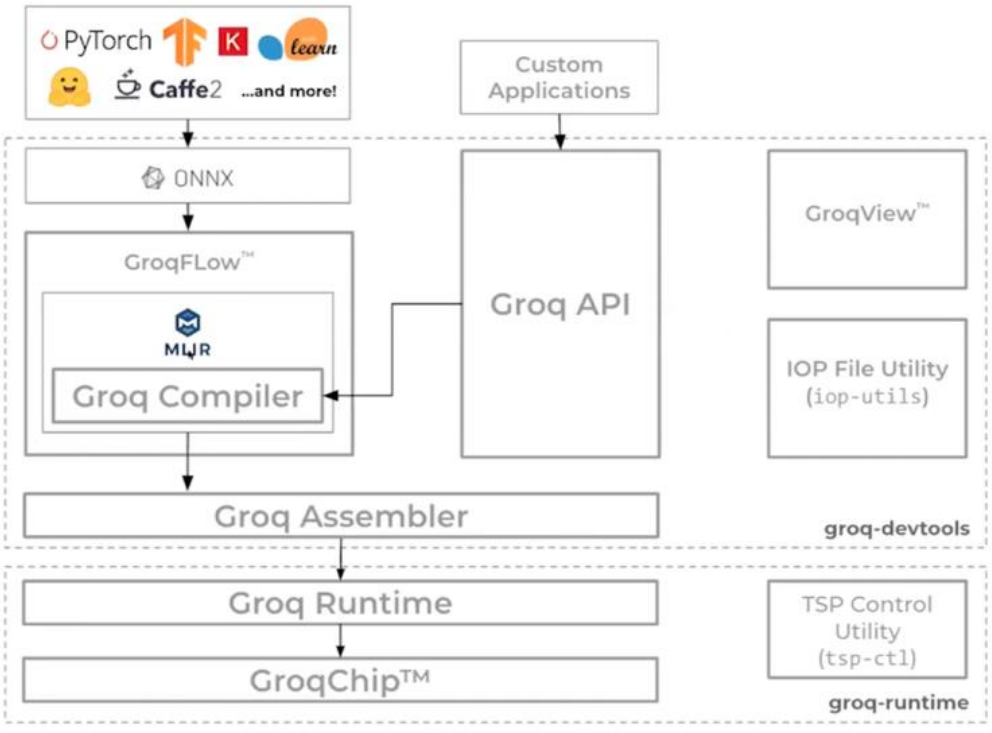
软件栈决定硬件精度能力在实际推理中的使用方式。GroqWare 并未对模型采取统一量化路径（如整体强制 INT8），而是基于算子与数值敏感度实施差异化的精度管理策略。例如，Attention logits（Softmax 输入）仍维持 FP32 精度，以避免微小数值误差在长序列中被放大；MoE 权重则采用 Block Floating Point 形式，在牺牲部分精度的同时保留量级尺度；同时，Groq 引入 TruePoint 数值体系，通过约 100 bits 的高精度中间累加抑制量化噪声。在权重与激活值层面，Groq 主要支持 INT8 与 FP16，并依托 TruePoint 的高精度累加机制缓解量化误差影响。当前，Groq 尚未采用 GPU 体系下的 FP8 硬件算子，而是通过这一混合精度路径，在维持模型精度的前提下，相较 BF16 实现约 2-4 倍的性能提升。而英伟达在硬件层面原生支持 FP4、FP6，并同时覆盖 FP8、BF16 与 FP32 等多种数值格式，其软件生态（CUDA、TensorRT-LLM）成熟且高度灵活，开发者可在完善的库与工具链支持下，自主选择并调优不同精度组合。与之相比，Groq 软件生态更为封闭，精度控制在更大程度上由编译器侧（如 TruePoint 体系）统一管理，开发者手动调节空间相对有限。

图表8：GroqWare 生态



资料来源：Groq 官网，华泰研究

图表9：Groq 开发者工具



资料来源：Groq 官网，华泰研究

问题 4：Groq 架构的主要结构性约束与经济性限制是什么？

我们认为，Groq 的 LPU 架构面临两项核心约束：其一是限制同步扩展能力的明确物理边界（Determinism Cliff），其二是由 SRAM 带来的昂贵资本与运行开支（SRAM Tax）；其在初始资本开支层面显著高于英伟达平台，但在以交互速度作为核心价值的 Agentic 经济中，更胜一筹。在此类场景下，Groq 在 Batch Size = 1 条件下仍能维持较高算力利用率，使其在部分对实时性要求极高的应用场景中，具备相对可竞争的总体拥有成本（TCO）。从本质上看，Groq 并非一项“算力最大化”投资，而是一项以客户体验为核心的系统性投入。尽管 SRAM 架构在模型规模与物理部署上存在天然约束，但其低时延的特性，为高实时要求的应用提供一条可验证的商业化路径，从而在特定场景下合理化较高的前期资本投入。

1）“确定性孤岛”（Determinism Island）边界：Scale-Out 的结构性约束

Groq 的核心技术优势（指令级确定性执行）在物理上受限于其互连体系。我们认为，Groq 的 RealScale 互连在 576 颗芯片规模内，构成一个近乎理想的同步执行环境：所有计算与通信时序均在编译期被精确规划。然而，当模型规模进一步扩大（如万亿参数级别），系统不可避免地需要将多个 576 芯片域通过标准以太网进行连接。此时，网络拥塞、数据缓冲与不可预测的抖动重新出现，其在超大模型场景下的核心价值也随之被削弱。

2）SRAM 带来的高资本开支强度（SRAM Tax）

我们认为，Groq 选择 SRAM 而非 HBM 作为主存储介质，在时延上带来优势的同时，也在前期资本投入具备较高代价。以 70B 参数模型（INT8）为例，Groq 需要部署 576 颗 LPU、共 8 个机架，仅用于提供约 70 GB 的 SRAM 容量；相比之下，同一模型在英伟达平台仅需 1-2 张 B300 GPU（单卡 288 GB HBM3E）。从资本开支角度测算，70B 模型下，Groq 集群的硬件投入约 300 万美元（约 5 千美元单颗芯片）；而英伟达双 B300 卡配置仅需约 8 万美元（即便考虑以 8 卡构成的完整服务器节点，成本亦约 40 万美元）。在 1T 参数模型场景下（1000GB SRAM 需求），这一差距进一步放大：Groq 需约 3,000 万美元的部署成本，而英伟达仍可在单节点内完成（8 卡服务器），资本投入维持在 40 万美元量级。需要强调的是，上述测算仅用于说明数量级差异，而非精确成本对比。但可以明确的是，该差异同时意味着 Groq 在功耗、散热、布线以及数据中心占地面积等方面承担显著更高的系统性开销。

图表10：Groq API 调用价格

大语言模型			
AI 模型	当前速度（每秒 tokens）	输入 Token 价格（每百万 tokens）	输出 Token 价格（每百万 tokens）
GPT OSS 20B 128k	1,000 TPS	$0.075	$0.30
GPT OSS Safeguard 20B	1,000 TPS	$0.075	$0.30
GPT OSS 120B 128k	500 TPS	$0.15	$0.60
Kimi K2-0905 1T 256k	200 TPS	$1.00	$3.00
Llama 4 Scout (17Bx16E) 128k	594 TPS	$0.11	$0.34
Llama 4 Maverick (17Bx128E) 128k	562 TPS	$0.20	$0.60
Llama Guard 4 12B 128k	325 TPS	$0.20	$0.20
Qwen3 32B 131k	662 TPS	$0.29	$0.59
Llama 3.3 70B Versatile 128k	394 TPS	$0.59	$0.79
Llama 3.1 8B Instant 128k	840 TPS	$0.05	$0.08
文本转语音模型			
AI 模型	速度（每秒字符数）	价格（每百万字符，美元）	
Canopy Labs Orpheus English	100	22	
Canopy Labs Orpheus Arabic Saudi	100	40	
自动语音识别（ASR）模型			
AI 模型	速度系数	价格（每小时转录，美元）	
Whisper V3 Large	217x	0.111	
Whisper Large v3 Turbo	228x	0.04	
Prompt caching			
模型	未缓存输入（每百万美元）	已缓存输入（每百万美元）	输出 Token（每百万美元）
moonshotai/kimi-k2-instruct-0905	1	0.5	3
openai/gpt-oss-120b	0.15	0.075	0.6
openai/gpt-oss-20b	0.075	0.0375	0.3
内置工具（Compound）			
工具	价格	参数	
Basic Search	$5 / 1000 requests	web_search	
Advanced Search	$8 / 1000 requests	web_search	
Visit Website	$1 / 1000 requests	visit_website	
Code Execution	$0.18 / hour	code_interpreter	
Browser Automation	$0.08 / hour	browser_automation	
内置工具（GPT-OSS）			
工具	价格	参数	
Browser Search – Basic Search	$5 / 1000 requests	browser_search - browser.search	
Browser Search – Visit Website	$1 / 1000 requests	browser_search - browser.open	
Code Execution – Python	$0.18 / hour	code_interpreter - python	

资料来源：Groq 官网，华泰研究

3）经济可行性：Token 效率 vs. 容量规模

尽管前期资本开支明显，我们认为 Groq 在特定运行条件下仍具备经济可行性，其核心在于利用率结构的差异。GPU 的成本效率高度依赖高并发负载，只有在被成千上万并发请求“填满”时，才能有效摊薄 HBM 访存时延。对于流量波动大、请求不可预测的应用，长期维持高负载 GPU 集群反而可能效率偏低。在 Batch Size = 1 情境下，B300 GPU 往往因等待 HBM3E 数据而处于低利用状态；而 Groq 能够在单请求条件下维持较高算力占用，其单位 token 能耗显著更低（约 1-3 焦耳，而 GPU 通常为 10-30 焦耳）。在交互型、实时型业务中，这一差异可转化为更具竞争力的 token 运营成本。

4）混合部署范式：

我们认为，行业已逐步形成分工明确的混合部署策略，在 Groq 的速度优势与英伟达的容量优势之间取得平衡。英伟达 GPU 仍将作为高吞吐训练与大批量推理的基础设施，其大容量 HBM 以及统一的 InfiniBand / Spectrum-X 互连，使其成为“AI Factory”的底座。而 Groq 的 LPU 更适合作为面向用户的“接口层”，承担对时延高度敏感的“最后一公里”任务，在该层面上，速度本身即是产品。

问题 5：哪些市场细分能够支撑 Groq 的前期资本投入？为何“时延敏感型推理”正在从小众需求变为主流？

我们认为，Groq 的价值已不再局限于少数“特殊场景”。随着交互式、实时 Agentic AI 逐步走向主流，时延敏感型推理正从“特定需求”转变为“基础设施级需求”，其可服务市场正在显著扩张。我们认为，此结构性变化为 Groq 较高的前期较高的资本投入与“Token 经济学”提供合理性，使其在 AI 基础设施栈中占据关键节点位置。

当交互响应速度成为产品的核心价值主张时，Batch Size=1 不再是少数状态，而是系统的常态运行模式。在这一运行范式下，系统无法通过等待更多请求形成批处理（否则将引入额外排队时延），而必须对单次请求即时执行；而单请求执行效率并非 GPU 的主要优化目标。此时，单一用户或对话正在等待模型生成下一个 token，而端到端时延与尾时延（P99）将直接影响用户体验与转化效果。在标准 GPU 环境中，Batch Size = 1 在经济上效率较低，因为计算核心往往需要等待从 HBM 中取回权重数据而处于空闲状态。

对于绝大多数 AI 工作负载（如摘要、翻译、分类），单位 token 成本仍是唯一关键指标。在这些场景中，英伟达的 GPU 通过请求批处理实现极高吞吐，从而具备经济优势。因此，对于 Groq 而言，经济可行性最为明确、置信度最高的细分市场，集中在那些时延可以被直接变现，或时延本身即产品价值的应用中。对于交互式应用而言，人类大脑对超过 200ms 的延迟即可感知为对话或思维上的卡顿；若 AI 需要执行多步推理，例如 Chain-of-Thought（CoT）智能体，则将叠加为数秒级等待，从而打断用户的认知连续性。

例如，在实时语音/电话推理场景中，时延是关乎用户体验的核心变量：人类对对话轮次切换的延迟具有即时感知，而尾时延尤为关键，因为“偶发性的卡顿”会破坏整体体验，并直接影响用户留存。同样，在具备严格 Time-to-First-Token 约束的交互式聊天场景（如客服助手、面向消费者的对话产品），若产品明确以“响应速度与即时交互”为核心卖点，而非最低的单位 token 成本，则客户在成本与体验权衡下，具备为更低且更稳定的时延支付溢价的合理性。

尽管大量 Agentic 工作流 以异步方式运行，对时延具备一定容忍度，但在引入人类参与闭环（Human-in-the-Loop，HITL）的交互式智能体场景中，往往存在更为刚性的时延约束。以吞吐为核心设计目标的 GPU，在架构层面难以系统性满足低时延需求。因此，有必要将“多步骤/Agentic 工作流”与“时延敏感型工作负载”明确区分。具体而言，后台自动化任务（如发票处理、业务流程管理）以及多数研究型智能体，通常以吞吐效率或单位成本为核心指标，而非 P99 时延（低时延要求）；另一方面，追求极致低时延的高频交易系统则属于纳秒级基础设施范畴，同样处于不同的技术区间。

即便 Batch Size = 1 在战略层面具有重要性，我们认为 Groq 的经济性仍受到其 SRAM 容量的制约：超大模型需要更高的芯片数量，从而增加系统体量与综合开销。SemiAnalysis 认为，一旦将平台级成本纳入考量，“高速 token” 在 TCO 层面的相对优势在多数部署场景中并不显著。因此，对 Groq 需求的合理评估，关键在于：究竟有多少具备规模收入的产品对 Batch Size = 1 下的尾时延存在刚性要求，并愿意为此支付溢价？若这一数量显著上升（例如语音与实时交互式 AI 成为默认 UI），Groq 的确定性将更具经济相关性；反之，GPU 仍将是最优选择，其批处理实现吞吐效率与成本效益的更优平衡。

图表11：Groq 与英伟达推理系统运行成本对比

指标	单位	GPU 系统（低时延优化）	GPU 系统（吞吐量优化）	Groq 机柜（成本计价）	Groq 机柜（60% GPM 计价）
系统资本成本		**8xH100**	**8xH100**	**8 Rack**	**8 Rack System**
前期系统资本支出	USD	350000	350000	2520000	6350000
使用寿命	Years	5	5	5	5
月均摊销资本支出	USD/mth	3638	3638	26191	65998
资本成本 / 最低预期回报率	%	18%	18%	18%	18%
月均资本成本	USD/mth	5250	5250	37800	95250
月均系统总资本成本	USD/mth	8888	8888	63991	161248
系统托管成本					
电网电价	USD/kWh	0.087	0.087	0.087	0.087
每月时长	Hours	730	730	730	730
利用率	%	80%	80%	80%	80%
电源使用效率（PUE）	Ratio	1.25	1.25	1.25	1.25
每千瓦月均有效电费	USD/kW/mth	63.5	63.5	63.5	63.5
托管服务费	USD/mth	190	190	190	190
每千瓦月均总托管成本	**USD/kW/mth**	**253.5**	**253.5**	**253.5**	**253.5**
系统功耗	kW	10.2	10.2	230.4	230.4
月均系统总托管成本	**USD/mth**	**2586**	**2586**	**58409**	**58409**
月均资本成本占系统总成本比例	**%**	**77%**	**77%**	**52%**	**52%**
月均资本 + 托管总成本	**USD/mth**	**11474**	**11474**	**122400**	**219657**
资本 + 托管小时总成本	**USD/hour**	**15.7**	**15.7**	**167.7**	**300.9**
单系统芯片数量	Chips	8	8	576	576
单推理单元芯片数量	Chips	8	2	576	576
单系统推理单元数量	Units	1	4	1	1
单推理单元小时租赁成本	**USD/hour**	**15.7**	**3.9**	**167.7**	**300.9**
单用户每秒Token数	Tok/s/user	420	30	500	500
批处理大小	-	2	64	3	3
流水线并行度	-	1	1	16	16
单推理单元并发用户数	Users	2	64	48	48
单推理单元每秒总处理令牌数	Tok/s	840	1920	24000	24000
单推理单元每小时总处理令牌数	Tok/hour	3024000	6912000	86400000	86400000
每百万Token成本	**USD / 1M Tok**	**5.20**	**0.57**	**1.94**	**3.48**

资料来源：Groq 官网，Semi Analysis，华泰研究

问题 6：如何理解英伟达 25 年 12 月收购 Groq 的战略动因？

我们认为，英伟达以约 200 亿美元收购 Groq，本质上是一项前瞻性战略布局，旨在引入一类专门面向实时 Agentic 推理的超低时延 AI 加速器架构。该举措使英伟达得以在 AI 产业“下半场”（以 Agentic 推理为核心）率先确立技术标准，并补齐低时延推理的短板。

GPU 架构以 HBM 与动态硬件调度为核心，在高吞吐任务中效率极高，但在 Agent 所需的多步骤推理中，仍可能出现时延抖动与尾时延放大。通过引入 Groq 的片上 SRAM 架构与确定性执行机制，英伟达有望在现有计算体系中补齐低时延短板，为复杂 Agentic AI 场景提供稳定、可预测的推理性能。我们认为，该举措有助于英伟达在 Agentic 时代率先构建更为完整的 AI 加速平台能力，并进一步巩固其平台级竞争壁垒。

从更宏观的视角看，英伟达在巩固训练端主导地位（AI 上半场）之后，已开始前瞻性布局以 Agentic 推理为核心的 AI 下半场，提升实时交互能力。通过将确定性计算的先行者 Groq 纳入自身体系，英伟达旨在确保面向实时交互的 Agentic 工作负载（如交互式 AI 助手）仍持续运行于其 GPU + CUDA 生态之上。我们认为，英伟达正推动 Agentic AI 走向规模化部署，并演进至一种统一的“反射式（reflex）”系统架构：即便工作负载从大规模训练迁移至低 batch、强时效的实时推理阶段，其平台级中心地位仍可维持。其战略核心在于构建一套异构的 Agentic 技术栈，由 GPU 承担高吞吐训练与批量推理，而 Groq 的确定性技术则作为“反射层”，专门服务于对时延高度敏感的实时 Agentic 推理场景。

我们认为，此次交易并非意在“封堵竞争对手”的防守动作，而是一次将确定性计算 DNA 主动注入 CUDA 生态的进攻型整合。通过架构融合，英伟达希望确保其平台仍是 Agentic AI 时代的核心，且开发者能够在 CUDA 环境内构建更复杂、更自主的智能体系统。从执行层面看，交易的重要组成部分在于引入 Groq 创始人 Jonathan Ross 及其系统架构团队，以加速英伟达 Rubin 平台及后续路线图的推进。Groq 的 RealScale Scale-up 互联技术提供确定性的通信结构，使整个集群能够在逻辑上作为一个低时延的 Mega-Chip 协同运行。我们认为，在多数竞争对手仍聚焦追赶英伟达训练性能之际，英伟达已将竞争焦点前移至实时 Agentic 推理，在战略层面削弱科技巨头为实时 Agentic 场景自研、整合 ASIC 的威胁。

问题 7：Groq 的 LPU 与英伟达 GPU 如何在训练与推理环节形成互补，共同支撑 Agentic AI 时代？

我们认为， GPU 与 LPU 的混合协同，为 Agentic AI 时代提供关键基础设施，而 Agentic AI 有望在 2026 年成为 AI 应用演进的主线。英伟达在训练环节的长期主导地位，与 Groq 在推理侧所具备的速度优势相结合，构成 Agentic 经济中极具吸引力的技术组合：一方面，英伟达 GPU 仍是模型训练阶段不可替代的 AI 工厂；另一方面，Groq 的 LPU 架构则在实时自主智能体场景中，充当专用的“推理引擎”。在用户交互这一关键环节，系统以牺牲 GPU 的部分通用性为代价，换取 LPU 所提供的速度与确定性。我们预计，这种分工将建立 Agentic AI 时代的全生命周期服务，从训练阶段的模型生成，到面向用户的实时推理与决策。

1）从 Chatbot 向 Agent 的转变

我们认为，2026 年的核心变化在于 AI 形态从被动响应式 Chatbot，转向具备主动性的 Agent。与传统 Chatbot 不同，AI Agent 并非仅对提示作出回应，而是能够自主拆解目标、规划执行路径，并通过内部的思维链（CoT）完成多步骤推理，且可在多智能体协同的体系中运行。该模式下，单一用户请求往往会触发数万 token 规模的内部推理、规划与反思过程。在推理执行层面，Groq 的 LPU 通过推测式解码（speculative decoding）可实现约 1,000 -1,600+ tokens/秒的生成速度，使智能体能够运行较长的内部 CoT 推理流程，同时在用户体验层面仍保持“即时响应”的感知。相比之下，英伟达 GPU 依然在基础模型训练、微调以及高吞吐推理方面具备不可替代的优势。

在此背景下，我们认为 2025 年 12 月 Meta 对 Manus 的收购具有重要的行业信号意义。该交易为 Meta 历史上规模第三大的并购，我们判断，这一举动象征着“Chatbot 时代的结束”与“Agentic 时代的开启”。 Meta 通过收购推动 Agent 在 WhatsApp、Instagram 等核心产品中的落地，旨在验证多步骤、持续运行的智能体将成为主流产品形态的战略判断。我们亦认为，这进一步强化 GPU+LPU 混合算力架构作为底层基础设施的合理性：由 GPU 负责生成、训练与持续更新智能体的能力边界，由 LPU 负责在交互端支撑智能体以极高速度完成“思考与推理”，且不会造成用户侧的可感知时延。这一分工模式，或将成为 Agentic AI 走向规模化落地的自然选择。

2）Agentic 推理闭环：当“速度”本身成为智能

我们认为，在 Agentic 经济中，系统性能的核心瓶颈已不再是人类的阅读或理解速度，而是智能体内部的推理与决策速度。新一代交互式智能体要求底层硬件能够以“机器速度”完成思考，才能在用户侧维持连贯、自然的体验。对于运行在用户屏幕上的实时智能体（如 Manus），其“思考-行动”闭环必须接近即时完成。若智能体在实时界面中为决定下一步操作需要生成 2,000 个 token 的内部推理，Groq LPU（约 1,600+ token/秒）的推理可在约 1.2 秒内完成，从而使实时自主交互成为可能。同样，为避免打断对话节奏，在语音交互场景中，系统对整体时延的容忍度更为严格（总体时延通常需 500ms 以内）。我们认为，在该混合架构下，GPU 的作用更接近“大脑皮层”：依托高密度 HBM 承载海量参数、提供基础智能；而 Groq LPU 的作用更像“反射系统”，以 80 TB/s 速度调度模型权重（约为 HBM3E 的 10 倍），以亚毫秒级精度执行推理闭环。

3）多智能体工作流的确定性扩展

随着 Agentic AI 的普及，借助 RealScale 互连将数千颗芯片同步为 Mega-Chip 的能力，正逐步演化为关键竞争优势。在多智能体工作流中，任务需要频繁交接；一旦交接环节出现抖动或非确定性时延，整体自主流程即可能发生失步，进而影响系统稳定性与执行效率。在体系分工上，英伟达提供面向全局 AI 工厂的 Scale-out 基础设施（InfiniBand/Spectrum-X），而 Groq 的 RealScale 为本地推理集群提供 Scale-up 的确定性执行。由此，Agent 闭环在能力上实现分工协同：在英伟达平台上完成高质量训练，确保“足够聪明”；在 Groq 平台上进行确定性、低时延执行，体现“足够主动”。我们认为，具备确定性行为能力，是企业级自主智能体在专业化、实时场景中满足严格用户交互要求的关键前提。

4）Groq 式推理能力如何纳入英伟达路线图

我们认为，交易完成后的核心问题在于，英伟达如何将 Groq 的确定性推理编织进整体平台以承载 Agentic 负载。我们预计，Rubin、Feynman 及后续架构将在传统吞吐模式之外，引入明确面向智能体的“Agent 优化”运行模式。从落地路径看，整合体现在三层，1）硬件层：Groq 的编译期互连调度与定时张量并行，将影响未来系统在长 CoT 序列下的运行方式，显著降低抖动；2）软件层：GroqWare 式静态调度与混合精度可吸收进 CUDA/TensorRT，在不改变开发者编程模型的前提下优化 Batch Size =1；3）部署层：客户可进行分层部署，GPU 层承担训练与后台任务，LPU 衍生层支撑在线需求、增强用户的智能体体验。

图表12： Manus 2025 年进展、架构和工作流

资料来源：Manus 官网，华泰研究

问题 8：Groq 与 Tesla Dojo 在定位、架构与存储配置上有何差异？其战略结果为何出现分化？

我们认为，Groq 与 Tesla Dojo 在架构层面具有相似性：二者均采取对片上 SRAM 的高度依赖，以绕开传统基于 HBM 的 GPU 所面临的“存储墙”瓶颈。然而，在设计目标与落地执行上，二者却呈现出根本性的分化。Dojo 试图作为面向自动驾驶的高吞吐训练工厂，但受制于制造复杂度高，以及英伟达 GPU 在训练领域所形成的压倒性优势，最终未能取得成功；相比之下，Groq 则通过聚焦确定性、速度，成功将自身定位为服务于交互式 AI 的专用“推理引擎”。尽管 Tesla 在 2025 年末对其硬件路线图进行整合调整，我们认为，这一对比仍然构成一个典型案例，展示同样基于片上 SRAM 的架构，如何被用于两种几乎完全相反的目标。我们认为，Dojo 更像是一项面向晶圆级训练硬件的“登月式尝试”，其失败主要源于制造复杂度过高以及产品迭代节奏滞后；而 Groq 的成功，则来自其对确定性推理引擎的专注，该架构恰好满足 Agentic AI 对实时交互能力的核心要求。

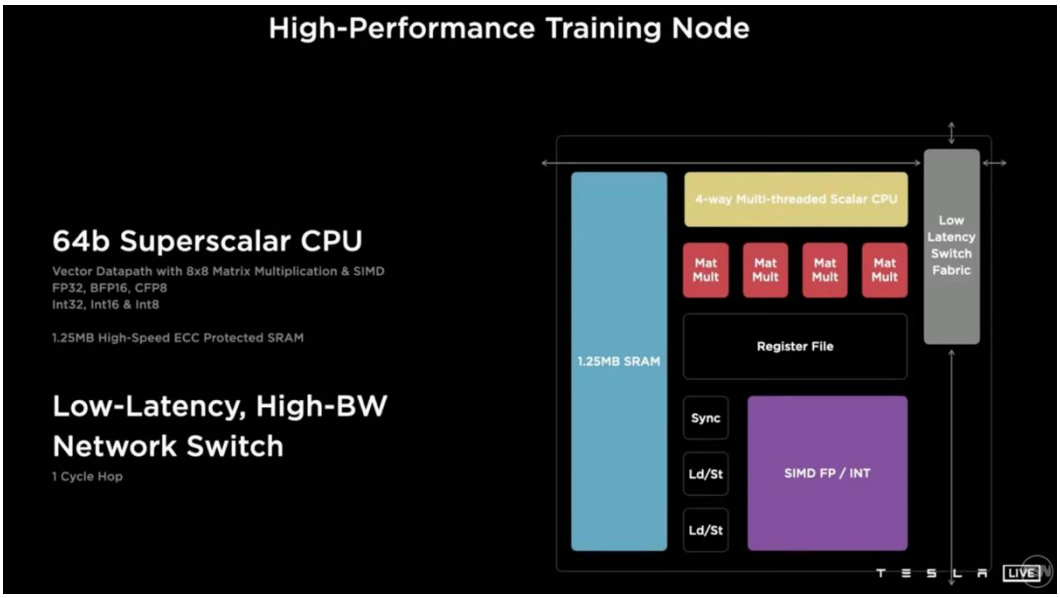
1）定位差异：训练工厂 vs. 推理工具

Tesla Dojo 的设计目标是一套高吞吐训练型超级计算系统，用于处理数十亿帧视频数据，以支撑特斯拉 FSD 模型的训练需求。相比之下，Groq 则是一款专注于推理的 ASIC，其核心目标是在大模型与 Agentic AI 的推理闭环中，提供确定性、实时的响应能力。

Dojo 为何未能成功？我们认为是其试图解决的是训练阶段所面临的大规模数据并行问题。我们认为，其受挫主要源于若干基础性硬件限制：包括晶圆级封装所带来的极高系统复杂度、HBM 与计算核心之间物理距离过远所导致的存储层级效率问题，以及在产品迭代节奏上难以跟上英伟达的快速更新周期。

Groq 为何能够成功？我们认为是其并未选择在 AI 训练市场与英伟达正面竞争，而是将重心放在一个技术上可行且边界清晰的细分场景，即 Batch Size =1 的推理任务。这一聚焦使其在独立的大模型推理基准测试中取得领先表现（可达 1,600+ tokens/秒），并将自身定位为 Agentic 经济中承担快速响应职责的“反射系统”。

图表13： Tesla Dojo 单节点配置 1.25MB SRAM



资料来源：Semi Analysis, Tesla AI Day 2021，华泰研究

2）架构差异：动态 Mesh vs. 静态确定性

Dojo 采用的是高性能动态 Mesh 架构，其本质是一套以硬件为中心的训练型超级计算机，通过定制化的片上网络（NoC）路由器，在二维 Mesh 中动态调度通信流量，以最大化满足 Physical AI 与 FSD 训练所需的超大规模数据流吞吐。同时，Dojo 依赖硬件级流控机制来缓解大规模训练过程中的网络拥塞问题。

Groq 则建立在纯粹的静态确定性之上。其 LPU 采用“软件定义硬件”架构，彻底移除运行时控制逻辑，包括分支预测器与硬件调度器等。GroqWare 编译器在执行前即对所有数据流动与指令执行进行逐时钟周期的预计算与排程。这种“指令级确定性”是多智能体工作流的必要条件。从功能定位看，Dojo 的动态 Mesh 更适合承担大批量训练中的“皮层（cortex）”角色；而 Groq 的静态确定性架构则更适合作为面向实时推理的横向标准。

3）存储配置：分布式 SRAM vs.层级化存储

两种系统均选择以 SRAM 的超高带宽，替代 HBM 的高容量，但在存储层级的组织方式上采取不同路径。Dojo 的存储层级成为其关键掣肘。尽管每个计算核心配备 1.25MB 的 SRAM（单 D1 芯片 SRAM 容量达 440MB），但 HBM 与计算核心之间的物理距离过远，内存访问请求需要穿越复杂的片上互连网络，带来较高的访问时延，从而抵消本地 SRAM 的带宽优势。相比之下，Groq 将 230 MB 的 SRAM 直接集成于芯片之上，并作为主参数存储，以 80TB/s 的带宽向计算单元供数。该设计使模型能够在单用户请求场景下以极高速度在处理器中流动，实质上缓解推理场景中由存储时延所形成的时延问题。

4）数值精度：可配置 FP8 vs.策略性混合精度

Dojo 采用的是可配置 FP8（CFP8）数值格式，旨在最大化梯度计算中的向量处理效率；相比之下，Groq 使用 TruePoint 数值体系，通过 100bits 高精度累加，确保以 INT8 存储的权重不会受到噪声的影响。同时，Groq 对关键的 Attention logits 专门维持 FP32 精度，以避免误差在计算过程中发生传播，从而满足 Agentic AI 中复杂 CoT 推理所需的高精度要求。

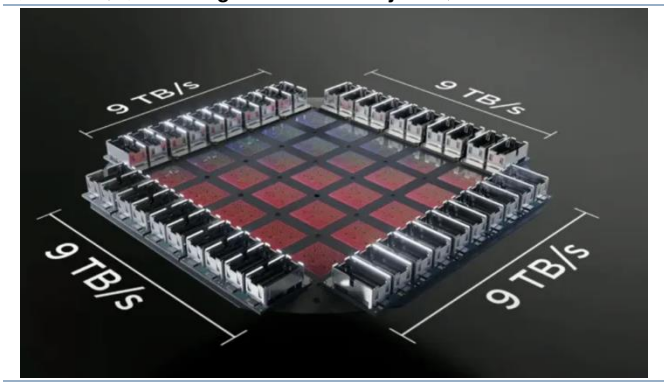
5）互连方式：晶圆级封装 vs.软件定义互连

Dojo 的受挫在很大程度上源于封装层面的失败。其依赖台积电的 System-on-Wafer（InFO_SoW）技术，将 25 颗芯片键合为单一的“训练 Tile”。这种极端复杂的封装方案导致良率偏低，并带来显著的散热管理问题；其中，与芯片共同封装的 HBM 尤其容易因热膨胀失配而发生故障。相比之下，Groq 通过 RealScale 避开晶圆级封装所带来的风险。由于编译器能够精确掌握芯片间数据包的传输时序，Groq 得以将最多 576 芯片（8 机架）同步为一个 Mega-Chip，并实现零网络拥塞的协同运行。

图表14：354 个功能单元构成的一颗 Dojo 芯片

资料来源：Semi Analysis, Tesla AI Day 2021，华泰研究

图表15：每个 Training Tile 由 25 颗 Dojo 芯片组成



资料来源：Semi Analysis, Tesla AI Day 2021，华泰研究

问题 9：Groq 与谷歌最新一代 TPU v7 如何对比？Jonathan Ross 的设计理念如何从 TPU v1 演进至 LPU？

我们认为，Groq 与谷歌最初的 TPU v1（由 Jonathan Ross 任职谷歌期间主导设计）在理念上具有一致性，二者均是以推理为优先、专门用于加速矩阵计算的 ASIC。TPU v1 于 2015 年投入部署，其设计目标是在满足严格的在线时延 SLA 的同时，实现相较当时 CPU/GPU 的吞吐效率（在代表性负载下，P99 响应时间经验证可达 7ms）。从技术脉络看，Groq 架构延续并强化 Ross 提出的“软件定义硬件”思想，通过将调度与控制完全前移至编译期，将早期 TPU 中残留的运行时不确定性，系统性演化为逐周期可预测的确定性执行流水线。

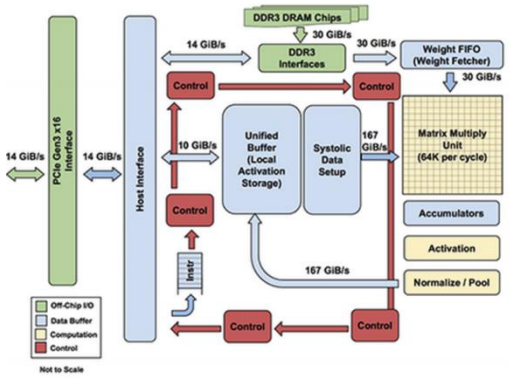
长期以来，谷歌依托纵向一体化的 TPU 体系及配套自研互连，在 Search、YouTube、Gemini 等核心内部业务中持续验证，其在特定工作负载下的性能与性价比可超越通用 GPU，从而对英伟达的市场主导地位形成实质性挑战。其成功的关键在于，谷歌对 TPU 在不同推理负载下的适配性与性价比边界具有高度清晰的判断。尽管 TPU 的目标已从最初的推理，演进为面向超大规模训练与服务的平台，但 Groq 则将“推理优先”基因进一步演化为面向 Agentic AI 的确定性计算工具。我们认为，这种战略目标的分化使得 TPU 在超大规模吞吐方面表现突出，而 Groq 则在自主智能体所需的亚毫秒级推理闭环上，保持标杆地位。

我们认为，谷歌 TPU 纵向整合模式的战略威胁，是推动英伟达与 Groq 达成交易的重要动因之一。将深谙谷歌 TPU 路线图的 Jonathan Ross 引入英伟达，本质上是面对 TPU 竞争的反制措施。通过内化 Ross 的经验，英伟达不仅推动其 Rubin 及其后续架构能够吸收确定性、低时延的核心优势，也将成为应对 TPU 竞争的关键手段，在事实上补齐“推理侧缺口”的短板。通过此次战略协同，英伟达有望确保下一代低时延推理理念仍在其 CUDA 生态中完成演进，确立 Agentic 经济下的全球性技术标准。

1）从 TPUv1 到 Groq LPU

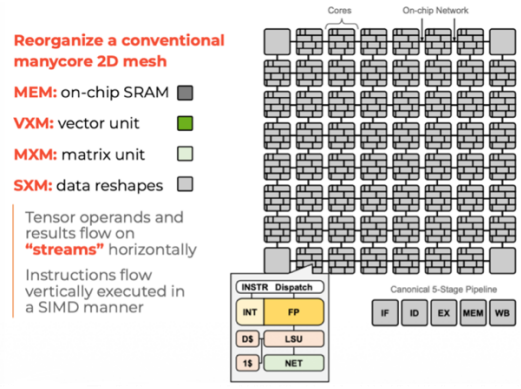
我们认为，Groq 与最初的 TPU v1 在结构层面的相似性，源于 Jonathan Ross 对传统动态硬件调度范式的反思，均体现以通过高度专用化的张量引擎绕开“冯·诺依曼瓶颈”的设计思路。最初 TPU v1 是一款专用于 AI 推理的 ASIC，起源于 Ross 在谷歌任职期间的“20% 项目”。Ross 认为，脉动阵列通过在计算单元之间实现数据的直接、有序流动，使得仅需一套精简的矩阵计算引擎，便可在推理场景下较当时主流 CPU/GPU 架构实现约 15-30 倍的性能提升。Ross 在 Groq 中延续并强化这一设计理念，从硬件主导的执行模式转向由编译器统一编排的执行模式。Groq 不再依赖动态硬件调度，而是几乎完全移除运行时控制复杂度与指令集开销，使芯片成为一个被动执行单元，仅按照编译器生成的、精确到时钟周期的预计算执行计划运行。相较之下，现代 TPU 为管理大规模训练 Pod 与混合负载，重新引入较为复杂的硬件控制逻辑。我们认为，Groq 的 LPU 有意保持高度架构刚性，通过 GroqWare 在编译期对每个时钟周期进行预调度，实现指令级确定性，这也成为 Ross 继 TPU 之后架构演进的标志性特征。

图表16：谷歌 TPUv1 架构



资料来源：谷歌官网，华泰研究

图表17：Groq LPU 芯片架构



资料来源：Groq 官网，华泰研究

2）架构取向：脉动阵列集群 vs. 确定性数据流

尽管两类芯片同样源自矩阵计算加速的技术谱系，但我们认为其定位已分化。谷歌的 TPU 为面向超大规模部署的“吞吐型引擎”，其优化目标在于批处理规模，以全局集群的单位时间的 token 处理量为最核心的经济指标；而 Groq 则面向交互场景的专用计算，重点服务于新兴 Agentic 经济中对极低时延的需求。

TPU 的架构核心围绕大规模矩阵乘单元（MXU）与脉动阵列展开，主要面向 AI 训练相关的计算。以最新一代 TPU v7（Ironwood）为例，其核心为一套 256×256 的脉动阵列，单芯片理论峰值性能达 4,614 TFLOPS。该架构通过硬件管理的调度机制来掩盖存储时延，以最大化数据的整体吞吐能力，从而成为高并发批量推理与基础模型训练的理想引擎。

相比之下，Groq LPU 是一款专为 Batch Size = 1 与实时 Agentic 推理打造的计算工具；采用张量流（Tensor Streaming）设计，将矩阵、向量与存储等功能单元交织于同一条同步流水线中。通过移除内核调用、硬件缓冲区与上下文切换，GroqWare 编译器以指令级确定性对每一次操作进行预先排程。这使得 LPU 的 token 生成速度可超过 1,600 t/s，从而在复杂的 CoT 推理场景中显著压缩 Time to First Token，并确保实时的用户智能体使用感知。

3）存储配置：HBM 容量 vs. SRAM 带宽

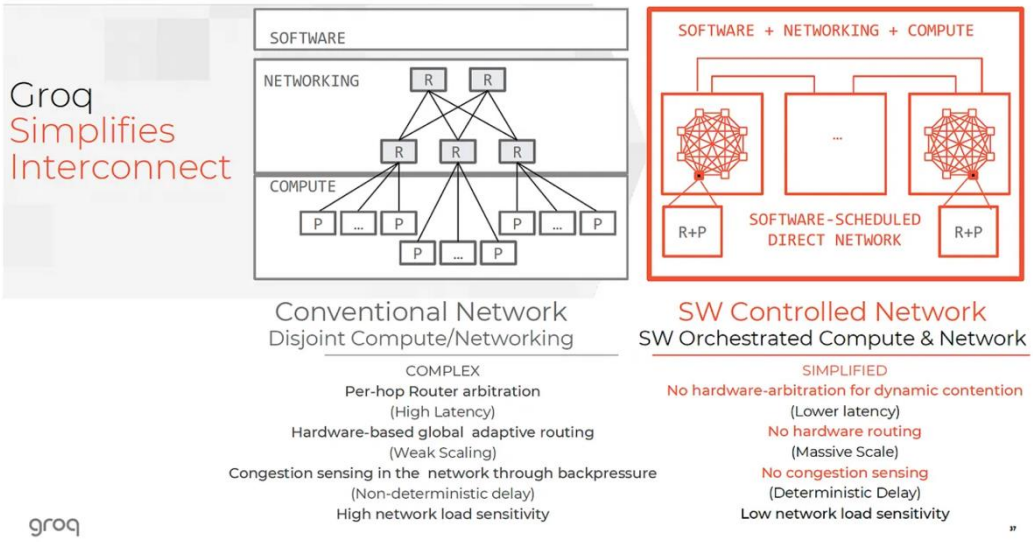
两种架构通过不同的存储层级设计来应对“存储墙”问题。以 TPU v7 为例，其面向万亿参数级模型（如 Gemini 3）的承载需求进行设计，单芯片配备 192 GB HBM3E，内存带宽约 7.4 TB/s。这一以 HBM 为中心的方案侧重于容量，以便在 Pod 级规模（单个 Pod 可达 9,216 颗芯片）内运行模型。相比之下，Groq 延续 Ross 尽量减少外部存储访问的设计理念，完全移除外部 HBM，以片上 SRAM 作为主存储介质，从而规避“存储墙”带来的时延。每颗 LPU 集成 230 MB 片上 SRAM，内部带宽达 80 TB/s，约为 TPU v7 HBM 带宽的 10 倍。这一设计使 Groq 成为 Batch Size = 1 推理的理想引擎。

图表18： 主流 AI 芯片参数对比

特点	Groq LPU (TSP)	Nvidia B300	Google TPU v7p	Cerebras CS-3
核心侧重点	Inference (Latency)	Training & Inference	Training & Inference	Training & Inference
内存架构	On-chip SRAM	Off-chip HBM3	Off-chip HBM	On-Wafer SRAM
内存容量	230 MB	288 GB	192 GB	44 GB
内存带宽	80 TB/s (Internal)	8.0 TB/s (External)	7.4 TB/s (External)	21 PB/s (Internal)
控制逻辑	Software (Compiler)	Hardware (Scheduler)	Hybrid (XLA)	Software (Compiler)
网络连接	RealScale (Switchless)	NVLink + InfiniBand	ICI (Torus)	SwarmX
单批次效率	Extremely High	Low (Memory Bound)	Medium	High

资料来源：Medium，Groq 官网，英伟达官网，谷歌官网，Cerebras 官网，华泰研究

图表19： RealScale 互联与传统互联对比



资料来源：Groq 官网，华泰研究

4）互连与扩展：OCS Pod vs. RealScale 确定性互连

随着 AI 工作负载向多智能体协同演进，扩展策略也成为区分两种架构的重要维度。TPU v7 采用光路交换（OCS）与三维环面（3D torus）互连，单 Pod 最多集成 9,216 颗芯片，聚合算力可达 42.5EFLOPS。Groq 则通过 RealScale 互连，侧重于 Scale-up 层面的确定性执行。不同于依赖硬件仲裁的 TPU 网络，Groq 的编译器将芯片间链路视作功能单元，并在时钟级别对每一次数据包传输进行预先排程。由此，最多 576 颗芯片可作为一个同步运行的 Mega-Chip 协同工作，实现零网络抖动，这对于对话式 AI 与自主 Agentic 工作所需的“反射式响应”至关重要。在 Pod/Mega-Chip 边界之外，两种体系均回退至标准以太网。Groq 的时延优势也将随之消失；而 TPU 由于其面向可容忍网络抖动的超大规模训练负载进行优化，仍能保持效率优势。

5）软件生态：谷歌的纵向体系 vs. 英伟达–Groq 的横向体系

我们预计，英伟达与 Groq 的合作模式将成为行业新范式。该交易将 Ross 及其技术经验引入英伟达生态中，使 Groq 的低时延“反射系统”能力被整合进全球领先的 AI 软件栈（CUDA）之中。相比之下，TPU 的软件生态相对闭塞，其围绕 JAX、TensorFlow 与 XLA 打造。

问题 10：并入英伟达体系后，Groq“下一代”芯片将呈现哪些特征？

我们认为，英伟达与 Groq 的交易本质上是“授权+人才并购（license +acqui-hire）”，其核心目的在于将确定性推理能力整合进全球领先的 AI 计算平台，并顺应交互式 AI 走向主流的趋势。通过将 Groq 的确定性调度机制与 TruePoint 数值体系（混合精度计算）纳入 CUDA / TensorRT 技术栈，英伟达得以确保其平台在 Agentic AI 时代仍保持领先。

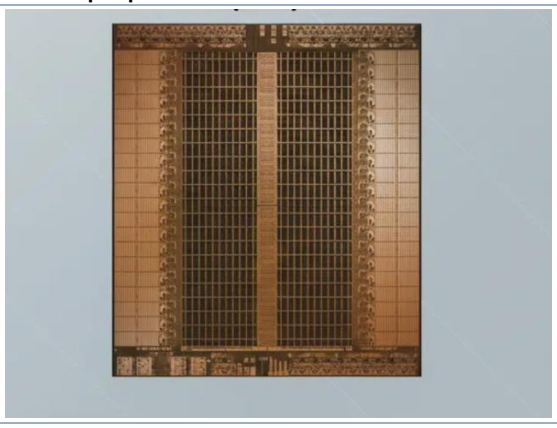
英伟达已获得 Groq LPU 技术的非独家授权，并吸纳包括创始人 Jonathan Ross 在内的核心工程团队，将相关 IP 纳入自身硬件路线图加速演进。我们认为，其战略意图在于，同时掌握 AI 训练侧的吞吐能力与 Agentic 推理侧的速度优势，并构建一套异构计算体系，由 GPU 承担高容量的“认知皮层”角色，由 LPU 担当高速、低时延的“反射系统”。随着这类“软件定义”的关键 IP 被融入 Rubin 及其后续架构，英伟达有望在 Agentic 时代提供明确面向智能体优化的运行模式，在保留 GPU 原算力优势的同时，实现确定性、零抖动的响应。

从 Groq 的独立硬件路线图来看，我们认为其具备三大核心方向：**1）制程演进：**GroqChip（LPU v1）由 GlobalFoundries 采用 14nm 工艺制造。下一代芯片或与 Samsung 合作，基于其 4nm（SF4X）制程开发。虽然最初预计于 2025 年末投产，但截至 2025 年底，尚无报道可确认下一代芯片实现规模化出货。**2）存储配置：**当前 LPU v1 在单芯片上集成 230 MB 的片上 SRAM。向 4nm 制程迁移，旨在提升晶体管密度，从而显著增加单芯片 SRAM 容量，以提高单 die 的有效存储规模，降低单一模型所需的芯片数量。目前，下一代芯片的具体 SRAM 容量尚未公开披露。**3）互连与扩展能力：**我们认为，下一代 Groq 技术的核心目标之一，在于扩大同步计算域的规模。在 LPU v1 中，系统最多仅维持 576 颗芯片组成的同步域（即 Mega-Chip）。我们认为，RealScale 2.0 或将提升这一同步扩展上限。

图表20：GroqChip v1

资料来源：Groq 官网，华泰研究

图表21：GroqChip v1



资料来源：Groq 官网，华泰研究

投资逻辑：英伟达布局 AI“下半场”，奠定 Agentic AI 时代技术标准

英伟达正通过推出专为实时 Agentic AI 优化的新型低延迟 AI 加速器，主动定义 AI 竞赛下半场的技术标准。继 2025 年被视为“物理 AI 元年”之后，我们认为 2026 年或为“Agentic AI 元年”；核心特征是向“低延迟、确定性执行”的转变。我们认为，英伟达此番布局，直指“延迟即瓶颈”的新战场。凭借此次整合，当市场从“重吞吐的训练阶段”转向“延迟关键的实时推理阶段”时，英伟达在训练与推理两大范式中均握有顶尖架构，提前化解科技巨头以定制芯片强攻“推理鸿沟”的战略威胁。

图表22：英伟达 2025 年投融资

序号	公司名称	交易日期	交易类型	交易金额（百万美元）	所属领域
1	Hippocratic AI	2025/1/9	早期风投	141	人工智能与机器学习、数字医疗
2	Synthesia	2025/1/15	后期风投	184	人工智能与机器学习、软件即服务
3	Generalist	2025/1/27	早期风投	128	人工智能与机器学习、机器人与无人机、软件即服务
4	Baskit	2025/1/27	加速器 / 孵化器孵化	—	人工智能与机器学习、大数据、金融科技、软件即服务、供应链技术
5	seqSight	2025/2/2	加速器 / 孵化器孵化	—	人工智能与机器学习、医疗科技、生命科学、软件即服务
6	Playbox	2025/2/4	加速器 / 孵化器孵化	—	人工智能与机器学习
7	MyWorker AI	2025/2/7	加速器 / 孵化器孵化	—	人工智能与机器学习、软件即服务
8	Lambda	2025/2/19	后期风投	480	人工智能与机器学习、云技术与开发运维、软件即服务
9	Gamerboom	2025/2/19	早期风投	9	人工智能与机器学习、加密货币 / 区块链、游戏
10	Together AI	2025/2/20	早期风投	305	人工智能与机器学习、云技术与开发运维、软件即服务
11	Achira	2025/2/21	种子轮融资	33	医疗科技、生命科学
12	Ubitus	2025/2/26	后期风投	30	游戏、移动互联网、软件即服务
13	Gretel	2025/3/19	兼并 / 收购	320	人工智能与机器学习、大数据
14	Lepton AI	2025/3/27	兼并 / 收购	—	人工智能与机器学习
15	Hibit	2025/3/31	早期风投	5	人工智能与机器学习、加密货币 / 区块链、金融科技
16	Runway AI	2025/4/3	后期风投	307	人工智能与机器学习、大数据、软件即服务
17	SandboxAQ	2025/4/4	后期风投	450	人工智能与机器学习、软件即服务
18	nEye Systems	2025/4/10	后期风投	58	人工智能与机器学习
19	StrateSea Technology	2025/4/10	加速器 / 孵化器孵化	—	人工智能与机器学习、软件即服务
20	Safe Superintelligence（与谷歌、绿橡树资本等联合投资）	2025/4/11	早期风投	2000	人工智能与机器学习
21	Nvidia (Manufacturing Space)	2025/4/14	企业资产收购	—	—
22	Utilidata	2025/4/29	后期风投	60	先进制造业、人工智能与机器学习、大数据、清洁能源技术、气候技术、工业技术数据
23	Confidios	2025/4/30	加速器 / 孵化器孵化	—	人工智能与机器学习、大数据、软件即服务
24	AI21 Labs	2025/5/11	后期风投	300	先进制造业、人工智能与机器学习、大数据、制造业、机器人与无人机
25	Skild AI	2025/5/16	早期风投	500	人工智能与机器学习
26	JV (Nvidia / Mistral AI / Bpifrance / MGX)	2025/5/19	合资企业	—	先进制造业、人工智能与机器学习、制造业、机器人与无人机
27	Figure AI（Parkway Venture Capital领投，英特尔等联合出资	2025/5/22	后期风投	1500	人工智能与机器学习、大数据
28	Lyzr	2025/5/22	加速器 / 孵化器孵化	—	人工智能与机器学习、移动互联网、软件即服务
29	Perplexity AI	2025/6/4	后期风投	600	人工智能与机器学习、网络安全、软件即服务
30	Repello AI	2025/6/12	种子轮融资	1	人工智能与机器学习、软件即服务
31	CentML	2025/6/13	兼并 / 收购	400	人工智能与机器学习、大数据、软件即服务
32	Cohere	2025/6/17	后期风投	—	人工智能与机器学习、软件即服务
33	Thinking Machines Lab（与AMD、Cisco等联合投资20亿）	2025/6/20	种子轮融资	2000	人工智能与机器学习、软件即服务
34	Commonwealth Fusion Systems	2025/6/26	后期风投	863	清洁能源技术、气候技术
35	xAI （与安德森·霍洛维茨、Blackrock等多家企业联合出资）	2025/7/10	后期风投	20000	人工智能与机器学习、移动互联网、软件即服务
36	Reka	2025/7/22	早期风投	110	人工智能与机器学习、软件即服务
37	Factory AI	2025/7/25	早期风投	50	人工智能与机器学习、大数据、软件即服务
38	FieldAI	2025/7/29	早期风投	315	人工智能与机器学习、机器人与无人机
39	DeepAware AI	2025/8/1	加速器 / 孵化器孵化	—	人工智能与机器学习、机器人与无人机、软件即服务
40	PhysicsX	2025/8/3	后期风投	209	先进制造业、人工智能与机器学习、大数据
41	Uber Freight	2025/8/13	后期风投	—	先进制造业、人工智能与机器学习、软件即服务、供应链技术
42	Nuro	2025/8/21	后期风投	200	人工智能与机器学习、自动驾驶汽车、移动出行技术、机器人与无人机
43	Sferical AI	2025/8/21	合资企业	—	—
44	Scintil Photonics	2025/8/29	后期风投	58	人工智能与机器学习
45	CHARM Therapeutics	2025/9/2	早期风投	80	人工智能与机器学习、生命科学、肿瘤学
46	Solver	2025/9/2	兼并 / 收购	—	人工智能与机器学习
47	Periodic Labs	2025/9/3	种子轮融资	300	人工智能与机器学习
48	Mistral AI（ASML领投，英伟达参与）	2025/9/9	后期风投	1518	人工智能与机器学习、移动互联网、软件即服务
49	PsiQuantum	2025/9/10	后期风投	750	先进制造业
50	A.A.A C(H+A)RM	2025/9/13	种子轮融资	4	人工智能与机器学习
51	AAA C(H+A)RM	2025/9/15	种子轮融资	4	人工智能与机器学习
52	Intel (NAS: INTC)	2025/9/18	私募股权投资	5000	人工智能与机器学习、物联网、制造业
53	Blue Water Autonomy	2025/9/23	加速器 / 孵化器孵化	—	—
54	Cohere	2025/9/24	后期风投	700	人工智能与机器学习、大数据、软件即服务
55	Nscale（与Blue Owl、戴尔、诺基亚等联合投资）	2025/9/25	后期风投	1487	人工智能与机器学习、大数据、云技术与开发运维
56	Phaidra	2025/10/1	后期风投	50	先进制造业、人工智能与机器学习
57	Nscale	2025/10/1	私募股权成长 / 扩张融资	1	人工智能与机器学习、大数据、云技术与开发运维
58	VAST Data	2025/10/5	后期风投	—	人工智能与机器学习、大数据、软件即服务
59	David AI (Delaware)	2025/10/8	早期风投	50	人工智能与机器学习
60	Bonvago	2025/10/16	加速器 / 孵化器孵化	—	人工智能与机器学习、加密货币 / 区块链、电子商务、金融科技
61	Uniphore	2025/10/22	后期风投	296	人工智能与机器学习、大数据、软件即服务
62	EMCOOL	2025/10/23	加速器 / 孵化器孵化	—	人工智能与机器学习、电子商务、制造业、纳米技术
63	Crusoe（Valor Equity Partners、阿布扎比主权财富基金 Mubadala Capital领投，英伟达参与）	2025/10/24	后期风投	1375	人工智能与机器学习、云技术与开发运维、软件即服务
64	Cassava Technologies	2025/10/24	早期风投	—	人工智能与机器学习、清洁能源技术、云技术与开发运维、网络安全、金融科技
65	Nokia	2025/10/28	私募股权投资	1003	—
66	Cartesia	2025/10/29	早期风投	100	人工智能与机器学习
67	Emerald AI	2025/10/30	种子轮融资	52	人工智能与机器学习、软件即服务
68	Quantinuum	2025/11/5	早期风投	839	人工智能与机器学习
69	Risorius	2025/11/6	加速器 / 孵化器孵化	—	医疗科技
70	Reflection AI（与花旗、红杉资本等联合投资）	2025/11/12	早期风投	2000	人工智能与机器学习、大数据、软件即服务
71	Anysphere（与谷歌等联合出资）	2025/11/13	后期风投	2300	人工智能与机器学习
72	Firmus Technologies	2025/11/14	后期风投	541	人工智能与机器学习、云技术与开发运维
73	Anthropic（与微软联合投资）	2025/11/19	后期风投	15000	人工智能与机器学习、大数据、软件即服务
74	Aligned Data Centers（与微软、xAI、BlackRock联合出资）	2025/11/26	收购 / 杠杆收购	40000	—
75	Synopsys	2025/12/1	私募股权投资	2000	—
76	Black Forest Labs	2025/12/1	早期风投	300	人工智能与机器学习、软件即服务
77	SchedMD	2025/12/15	兼并 / 收购	—	—
78	Adaptive	2025/12/16	早期风投	81	人工智能与机器学习、网络安全、软件即服务
79	Groq	2025/12/26	兼并 / 收购	20000	人工智能与机器学习、大数据、云技术与开发运维、软件即服务

资料来源：Pitchbook，华泰研究

图表23：英伟达 2024 年投融资

序号	公司名称	交易日期	交易类型	交易金额（百万美元）	所属行业
1	FieldAI	2024/1/1	早期风险投资	91	人工智能与机器学习、机器人与无人机
2	Dynamo	2024/1/1	早期风险投资	-	人工智能与机器学习、清洁能源科技、气候科技、移动技术
3	Cohesity	2024/2/7	后期风险投资	150	人工智能与机器学习、大数据、云技术与开发运维、网络安全、软件即服务
4	LIMAA Technologies	2024/2/13	加速器 / 孵化器孵化	-	人工智能与机器学习、生命科学
5	PT Blink	2024/2/22	加速器 / 孵化器孵化	-	气候科技、建筑科技、房地产科技、软件即服务
6	Mistral AI	2024/2/26	早期风险投资	431	人工智能与机器学习、移动技术、软件即服务
7	Talus Network	2024/2/26	早期风险投资	3	人工智能与机器学习、大数据、加密货币 / 区块链
8	Figure AI	2024/2/29	早期风险投资	675	先进制造业、人工智能与机器学习、制造业、机器人与无人机
9	Medvise	2024/3/10	加速器 / 孵化器孵化	-	人工智能与机器学习、健康科技、软件即服务
10	Union	2024/3/13	加速器 / 孵化器孵化	-	人工智能与机器学习、大数据、软件即服务
11	Hippocratic AI	2024/3/18	早期风险投资	72	人工智能与机器学习、数字医疗
12	DATS Project	2024/3/23	政府 / 机构补贴	0	加密货币 / 区块链、网络安全
13	Perplexity AI	2024/3/25	早期风险投资	135	人工智能与机器学习、移动技术、软件即服务
14	Bright Machines	2024/4/19	后期风险投资	126	先进制造业、人工智能与机器学习、机器人与无人机
15	Deci	2024/5/2	并购	300	人工智能与机器学习、大数据、软件即服务
16	Wayve（软银、Balderton等联合投资）	2024/5/6	后期风险投资	1027	人工智能与机器学习、自动驾驶汽车、移动技术、软件即服务
17	WEKA	2024/5/15	后期风险投资	140	人工智能与机器学习、大数据、金融科技、生命科学、软件即服务、科技媒体通信
18	PolyAI	2024/5/16	后期风险投资	51	人工智能与机器学习、大数据、软件即服务
19	Mistral AI	2024/6/11	早期风险投资	651	人工智能与机器学习、移动技术、软件即服务
20	Arrcus	2024/6/11	后期风险投资	30	云技术与开发运维、软件即服务
21	Waabi	2024/6/16	早期风险投资	200	人工智能与机器学习、自动驾驶汽车、大数据、移动技术、软件即服务
22	Factory AI	2024/6/18	早期风险投资	15	人工智能与机器学习、大数据、软件即服务
23	Shoreline.io	2024/7/1	并购	100	云技术与开发运维
24	Hayden AI	2024/7/3	后期风险投资	95	人工智能与机器学习、自动驾驶汽车、大数据、移动技术
25	SimProBot	2024/7/12	加速器 / 孵化器孵化	-	人工智能与机器学习、大数据、云技术与开发运维、移动技术、软件即服务
26	CytoReason	2024/7/15	后期风险投资	80	人工智能与机器学习、大数据、数字医疗、健康科技、生命科学、肿瘤学
27	Mazing	2024/7/16	加速器 / 孵化器孵化	-	人工智能与机器学习、增强现实、软件即服务、虚拟现实
28	Brev.Dev	2024/7/17	并购	-	云技术与开发运维、软件即服务
29	Accuknox	2024/7/19	加速器 / 孵化器孵化	-	人工智能与机器学习、网络安全、软件即服务
30	Cohere	2024/7/22	后期风险投资	500	人工智能与机器学习、大数据、软件即服务
31	Odyssey	2024/8/1	早期风险投资	-	人工智能与机器学习、软件即服务、航天技术
32	Fireworks AI	2024/8/7	早期风险投资	52	人工智能与机器学习、软件即服务
33	Safe Superintelligence（与Andreessen Horowitz、Sequoia Capital等联合投资）	2024/9/4	早期风险投资	1000	人工智能与机器学习
34	you.com	2024/9/4	早期风险投资	54	人工智能与机器学习、软件即服务、科技媒体通信
35	Xscape Photonics	2024/9/4	早期风险投资	47	人工智能与机器学习
36	W.AI	2024/9/4	种子轮融资	9	人工智能与机器学习、移动技术
37	Applied Digital (NAS: APLD)	2024/9/6	私募股权投资	160	人工智能与机器学习
38	OctoAI	2024/9/10	并购	250	人工智能与机器学习、云技术与开发运维、软件即服务
39	World Labs	2024/9/13	早期风险投资	230	人工智能与机器学习、增强现实、软件即服务、虚拟现实
40	Achira Labs	2024/9/16	后期风险投资	-	人工智能与机器学习、数字医疗、健康科技、生命科学
41	Sakana AI	2024/9/17	早期风险投资	214	人工智能与机器学习
42	AiMerch	2024/9/17	种子轮融资	0	人工智能与机器学习、制造业
43	OpenAI（Thrive Capital领投，与微软、软银等联合投资）	2024/10/2	后期风险投资	6600	人工智能与机器学习、大数据、软件即服务
44	Poolside (Software Development Applications)	2024/10/2	早期风险投资	500	人工智能与机器学习、软件即服务
45	Artisight	2024/10/23	后期风险投资	42	人工智能与机器学习、健康科技、物联网
46	Warburg AI	2024/10/26	加速器 / 孵化器孵化	-	人工智能与机器学习、大数据、金融科技
47	1910	2024/10/31	后期风险投资	-	人工智能与机器学习、大数据、健康科技、生命科学
48	Blomso	2024/11/1	加速器 / 孵化器孵化	-	农业科技、人工智能与机器学习、软件即服务
49	CentML	2024/11/6	加速器 / 孵化器孵化	-	人工智能与机器学习、软件即服务
50	SuperAnnotate	2024/11/18	后期风险投资	36	人工智能与机器学习、软件即服务
51	xAI（与A16Z、Blackrock等联合投资）	2024/11/20	后期风险投资	6000	人工智能与机器学习、移动技术、软件即服务
52	Ainovis	2024/11/20	加速器 / 孵化器孵化	-	人工智能与机器学习、健康科技
53	Nebulon	2024/11/20	并购	-	云技术与开发运维、软件即服务
54	Varjo	2024/11/25	后期风险投资	-	增强现实、游戏、科技媒体通信、虚拟现实
55	Black Forest Labs	2024/11/26	早期风险投资	126	人工智能与机器学习、软件即服务
56	Augtera Networks	2024/12/1	并购	-	人工智能与机器学习、大数据、软件即服务
57	Nebius Group	2024/12/2	私募股权投资	700	人工智能与机器学习、大数据、科技媒体通信
58	Yandex	2024/12/2	私募股权投资	700	大数据
59	Primech AI	2024/12/5	加速器 / 孵化器孵化	-	制造业、机器人与无人机
60	VinBrain	2024/12/6	并购	-	人工智能与机器学习、大数据、健康科技、肿瘤学、科技媒体通信
61	Lightning AI	2024/12/10	后期风险投资	50	人工智能与机器学习、大数据、云技术与开发运维、软件即服务
62	Ayar Labs	2024/12/11	后期风险投资	155	先进制造业、人工智能与机器学习、金融科技
63	Crusoe	2024/12/12	后期风险投资	686	人工智能与机器学习、云技术与开发运维、软件即服务
64	Perplexity AI	2024/12/16	后期风险投资	500	人工智能与机器学习、移动技术、软件即服务
65	Pleno	2024/12/24	加速器 / 孵化器孵化	-	人工智能与机器学习、机器人与无人机、软件即服务
66	Run:AI	2024/12/30	并购	700	人工智能与机器学习、软件即服务

资料来源：Pitchbook，华泰研究

风险提示

技术落地缓慢：公司的生产技术推进和产品落地可能达不到预期，或影响营收及利润。

芯片需求不及预期：市场的芯片需求规模可能不及预期，影响行业营收及利润。

图表24：重点公司推荐一览表

			最新收盘价	目标价	市值 (百万)	EPS (元)				PE (倍)			
股票名称	股票代码	投资评级	(当地币种)	(当地币种)	(当地币种)	2024	2025E	2026E	2027E	2024	2025E	2026E	2027E
英伟达(NVIDIA)	NVDA US	买入	184.86	280.00	4,491,913	1.33	2.99	4.96	8.48	139.02	61.74	37.27	21.79

资料来源：Bloomberg，华泰研究预测

图表25：重点推荐公司最新观点

股票名称	最新观点
英伟达(NVIDIA) (NVDA US)	路透社美东 12 月 8 日报道，美国政府已批准英伟达对华出口 H200 芯片。受此利好提振，英伟达股价盘后最高涨约 3.0%。美国政府将对每块出口芯片征收 25%费用，且特朗普表示，新政策同样适用于 AMD、英特尔等厂商。我们认为，此举反映政策预期有望保持温和改善。我们在 7 月《H20 恢复对华出口》报告中指出，英伟达 H20 与 AMD MI308 已恢复对华出口。但受限于芯片性能，市场反响平平。本次获准出口的 H200 在性能上更具优势，但仍属于上一代 Hopper 架构产品。鉴于此次政策缓和或可被视作以 H200 替代 Blackwell 合规产品出口的信号，政策仍具进一步改善空间，后续 Blackwell 合规版的出口或仍可期待。相比之下，B30A 在中国市场的竞争力更强，我们预计将延续 B20 规格，支持 FP4/FP6 算力，并搭载大容量的 HBM3e。此外，财富 12 月 2 日报道，英伟达 CFO 指出，与 OpenAI 的合作尚未计入当前约 5,000 亿美元的 Blackwell 与 Rubin 订单，英伟达中长期收入或仍具上行空间。重申“买入”。 展望 26 年，若以台积电 CoWoS 年产能约 100-120 万片计，我们预计英伟达可获得其约 60%晶圆份额，叠加台积电外溢 CoWoS 约 5 万片，并假设 GPU ASP 约 3.0-3.5 万美元；此次 H200 对华出口或新增约 200 亿美元收入；总计对应数据中心业务收入可超过 3300 亿美元。我们维持 FY26 营收净利预测，考虑到新增 Blackwell 和 Rubin 的订单将于明年放量，上调 FY27-28E 营收 5.6/28.7%到 3629/5020 亿美元，上调 FY27-28E Non-GAAP 净利润 5.6/28.7%到 2062/ 2811 亿美元。考虑到公司历史上增速中等的年份 PE 估值在 30x 附近，维持给予 33x FY27E PE。重申“买入”。 风险提示：技术落地缓慢、中美贸易摩擦、需求不及预期等。 报告发布日期：2025 年 12 月 10 日 **点击下载全文：英伟达(NVIDIA)(NVDA US,买入)：政策缓和推动 H200 对华出口**

资料来源：Bloomberg，华泰研究预测

免责声明

分析师声明

本人，何翩翩，兹证明本报告所表达的观点准确地反映了分析师对标的证券或发行人的个人意见；彼以往、现在或未来并无就其研究报告所提供的具体建议或所表达的意见直接或间接收取任何报酬。

一般声明及披露

本报告由华泰证券股份有限公司或其关联机构制作，华泰证券股份有限公司和其关联机构统称为“华泰证券”(华泰证券股份有限公司已具备中国证监会批准的证券投资咨询业务资格) 。本报告所载资料是仅供接收人的严格保密资料。本报告仅供华泰证券及其客户和其关联机构使用。华泰证券不因接收人收到本报告而视其为客户。

本报告基于华泰证券认为可靠的、已公开的信息编制，但华泰证券对该等信息的准确性及完整性不作任何保证。

本报告所载的意见、评估及预测仅反映报告发布当日的观点和判断。在不同时期，华泰证券可能会发出与本报告所载意见、评估及预测不一致的研究报告。同时，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。以往表现并不能指引未来，未来回报并不能得到保证，并存在损失本金的可能。华泰证券不保证本报告所含信息保持在最新状态。华泰证券对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

华泰证券（华泰证券（美国）有限公司除外）不是 FINRA 的注册会员，其研究分析师亦没有注册为 FINRA 的研究分析师/不具有 FINRA 分析师的注册资格。

华泰证券力求报告内容客观、公正，但本报告所载的观点、结论和建议仅供参考，不构成购买或出售所述证券的要约或招揽。该等观点、建议并未考虑到个别投资者的具体投资目的、财务状况以及特定需求，在任何时候均不构成对客户私人投资建议。投资者应当充分考虑自身特定状况，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。对依据或者使用本报告所造成的一切后果，华泰证券及作者均不承担任何法律责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

除非另行说明，本报告中所引用的关于业绩的数据代表过往表现，过往的业绩表现不应作为日后回报的预示。华泰证券不承诺也不保证任何预示的回报会得以实现，分析中所做的预测可能是基于相应的假设，任何假设的变化可能会显著影响所预测的回报。

华泰证券及作者在自身所知情的范围内，与本报告所指的证券或投资标的不存在法律禁止的利害关系。在法律许可的情况下，华泰证券可能会持有报告中提到的公司所发行的证券头寸并进行交易，为该公司提供投资银行、财务顾问或者金融产品等相关服务或向该公司招揽业务。

华泰证券的销售人员、交易人员或其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。华泰证券没有将此意见及建议向报告所有接收者进行更新的义务。华泰证券的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。投资者应当考虑到华泰证券及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突。投资者请勿将本报告视为投资或其他决定的唯一信赖依据。有关该方面的具体披露请参照本报告尾部。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布的机构或人员，也并非意图发送、发布给因可得到、使用本报告的行为而使华泰证券违反或受制于当地法律或监管规则的机构或人员。

中国香港

本报告由华泰证券股份有限公司或其关联机构制作,在香港由华泰金融控股（香港）有限公司向符合《证券及期货条例》及其附属法律规定的机构投资者和专业投资者的客户进行分发。华泰金融控股（香港）有限公司受香港证券及期货事务监察委员会监管，是华泰国际金融控股有限公司的全资子公司，后者为华泰证券股份有限公司的全资子公司。在香港获得本报告的人员若有任何有关本报告的问题,请与华泰金融控股（香港）有限公司联系。

香港-重要监管披露

- 华泰金融控股（香港）有限公司的雇员或其关联人士没有担任本报告中提及的公司或发行人的高级人员。
- 有关重要的披露信息，请参华泰金融控股（香港）有限公司的网页 https://www.htsc.com.hk/stock_disclosure

其他信息请参见下方 **“美国-重要监管披露”**。

美国

在美国本报告由华泰证券（美国）有限公司向符合美国监管规定的机构投资者进行发表与分发。华泰证券（美国）有限公司是美国注册经纪商和美国金融业监管局（FINRA）的注册会员。对于其在美国分发的研究报告，华泰证券（美国）有限公司根据《1934 年证券交易法》（修订版）第 15a-6 条规定以及美国证券交易委员会人员解释，对本研究报告内容负责。华泰证券（美国）有限公司联营公司的分析师不具有美国金融监管（FINRA）分析师的注册资格，可能不属于华泰证券（美国）有限公司的关联人员，因此可能不受 FINRA 关于分析师与标的公司沟通、公开露面和所持交易证券的限制。华泰证券（美国）有限公司是华泰国际金融控股有限公司的全资子公司，后者为华泰证券股份有限公司的全资子公司。任何直接从华泰证券（美国）有限公司收到此报告并希望就本报告所述任何证券进行交易的人士，应通过华泰证券（美国）有限公司进行交易。

美国-重要监管披露

- 分析师何翩翩本人及相关人士并不担任本报告所提及的标的证券或发行人的高级人员、董事或顾问。分析师及相关人士与本报告所提及的标的证券或发行人并无任何相关财务利益。本披露中所提及的“相关人士”包括 FINRA 定义下分析师的家庭成员。分析师根据华泰证券的整体收入和盈利能力获得薪酬，包括源自公司投资银行业务的收入。
- 华泰证券股份有限公司、其子公司和/或其联营公司，及/或不时会以自身或代理形式向客户出售及购买华泰证券研究所覆盖公司的证券/衍生工具，包括股票及债券（包括衍生品）华泰证券研究所覆盖公司的证券/衍生工具，包括股票及债券（包括衍生品）。
- 华泰证券股份有限公司、其子公司和/或其联营公司，及/或其高级管理层、董事和雇员可能会持有本报告中所提到的任何证券（或任何相关投资）头寸，并可能不时进行增持或减持该证券（或投资）。因此，投资者应该意识到可能存在利益冲突。

新加坡

华泰证券（新加坡）有限公司持有新加坡金融管理局颁发的资本市场服务许可证，可从事资本市场产品交易，包括证券、集体投资计划中的单位、交易所交易的衍生品合约和场外衍生品合约，并且是《财务顾问法》规定的豁免财务顾问，就投资产品向他人提供建议，包括发布或公布研究分析或研究报告。华泰证券（新加坡）有限公司可能会根据《财务顾问条例》第 32C 条的规定分发其在华泰证券内的外国附属公司各自制作的信息/研究。本报告仅供认可投资者、专家投资者或机构投资者使用，华泰证券（新加坡）有限公司不对本报告内容承担法律责任。如果您是非预期接收者，请您立即通知并直接将本报告返回给华泰证券（新加坡）有限公司。本报告的新加坡接收者应联系您的华泰证券（新加坡）有限公司关系经理或客户主管，了解来自或与所分发的信息相关的事宜。

评级说明

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力（含此期间的股息回报）相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，台湾市场基准为台湾加权指数，日本市场基准为日经 225 指数，新加坡市场基准为海峡时报指数，韩国市场基准为韩国有价证券指数，英国市场基准为富时 100 指数，德国市场基准为 DAX 指数），具体如下：

行业评级

增持：预计行业股票指数超越基准
中性：预计行业股票指数基本与基准持平
减持：预计行业股票指数明显弱于基准

公司评级

买入：预计股价超越基准 15%以上
增持：预计股价超越基准 5%~15%
持有：预计股价相对基准波动在-15%~5%之间
卖出：预计股价弱于基准 15%以上
暂停评级：已暂停评级、目标价及预测，以遵守适用法规及/或公司政策
无评级：股票不在常规研究覆盖范围内。投资者不应期待华泰提供该等证券及/或公司相关的持续或补充信息

法律实体披露

中国：华泰证券股份有限公司具有中国证监会核准的“证券投资咨询”业务资格，经营许可证编号为：91320000704041011J

香港：华泰金融控股（香港）有限公司具有香港证监会核准的“就证券提供意见”业务资格，经营许可证编号为：AOK809

美国：华泰证券（美国）有限公司为美国金融业监管局（FINRA）成员，具有在美国开展经纪交易商业务的资格，经营业务许可编号为：CRD#:298809/SEC#:8-70231

新加坡：华泰证券（新加坡）有限公司具有新加坡金融管理局颁发的资本市场服务许可证，并且是豁免财务顾问，经营许可证编号为：202233398E

华泰证券股份有限公司

南京
南京市建邺区江东中路 228 号华泰证券广场 1 号楼/邮政编码：210019

电话：86 25 83389999/传真：86 25 83387521
电子邮件：ht-rd@htsc.com

北京
北京市西城区太平桥大街丰盛胡同 28 号太平洋保险大厦 A 座 18 层/邮政编码：100032
电话：86 10 63211166/传真：86 10 63211275
电子邮件：ht-rd@htsc.com

深圳
深圳市福田区益田路 5999 号基金大厦 10 楼/邮政编码：518017
电话：86 755 82493932/传真：86 755 82492062
电子邮件：ht-rd@htsc.com

上海
上海市浦东新区东方路 18 号保利广场 E 栋 23 楼/邮政编码：200120
电话：86 21 28972098/传真：86 21 28972068
电子邮件：ht-rd@htsc.com

华泰金融控股（香港）有限公司
香港中环皇后大道中 99 号中环中心 53 楼
电话：+852-3658-6000/传真：+852-2567-6123
电子邮件：research@htsc.com
http://www.htsc.com.hk

华泰证券（美国）有限公司
美国纽约公园大道 280 号 21 楼东（纽约 10017）
电话：+212-763-8160/传真：+917-725-9702
电子邮件: Huatai@htsc-us.com
http://www.htsc-us.com

华泰证券（新加坡）有限公司
滨海湾金融中心 1 号大厦, #08-02, 新加坡 018981
电话：+65 68603600
传真：+65 65091183
https://www.htsc.com.sg