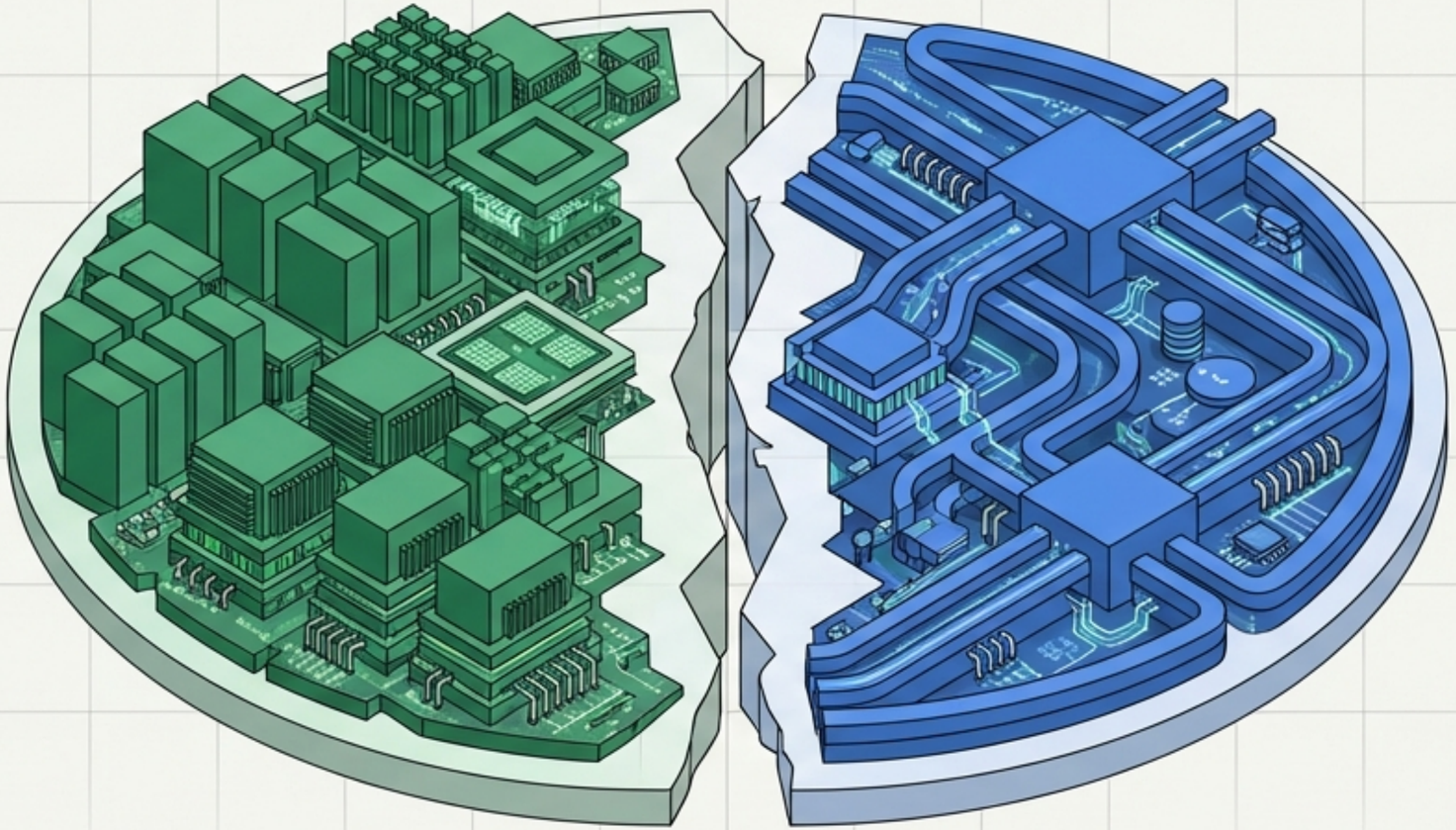
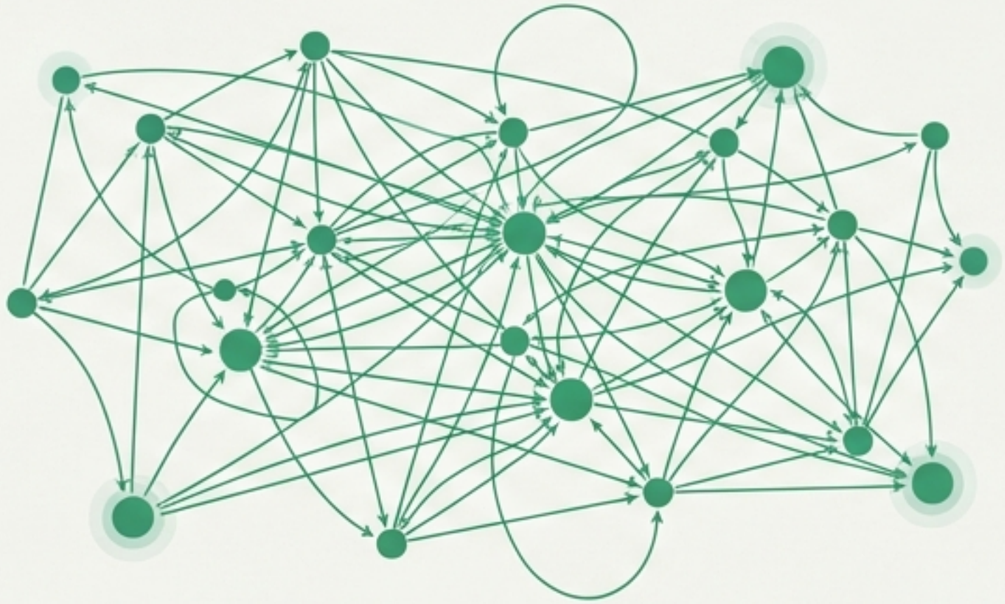


算力蓝图：这不是一场硬件替代战，而是生态的彻底裂变

基于前谷歌核心工程师访谈的底层架构与商业逻辑重构。

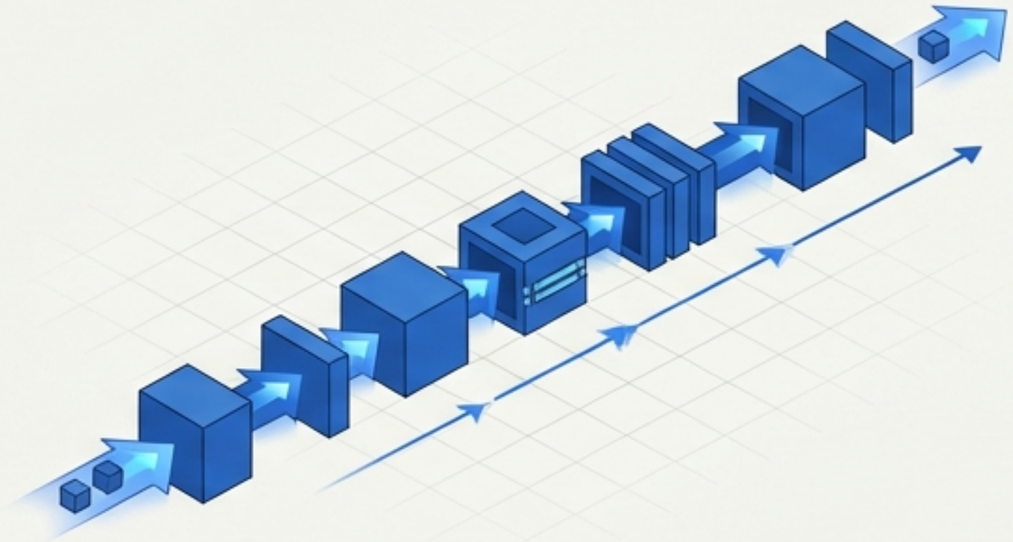
摒弃单维性能跑分，理解两种截然不同的底层计算哲学

通用性之王



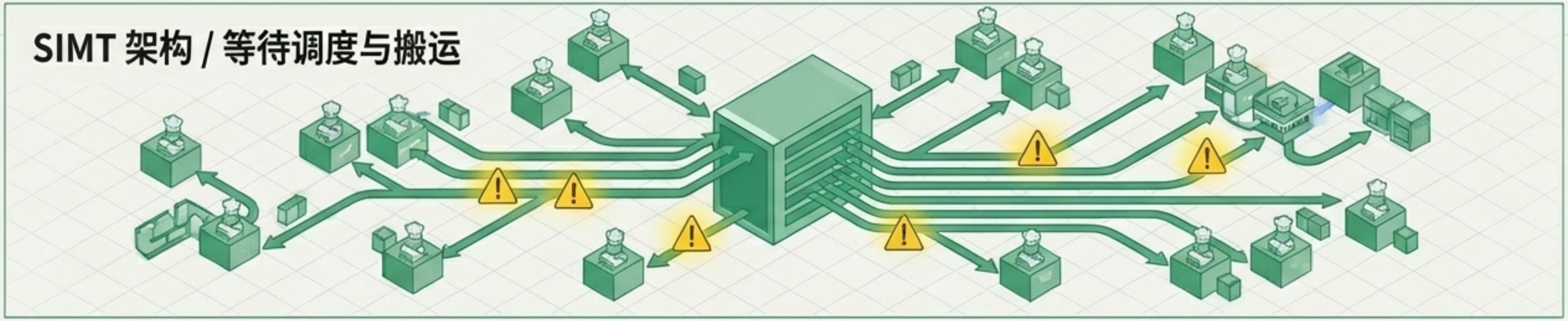
单兵作战极强，时刻准备适应未知算法。

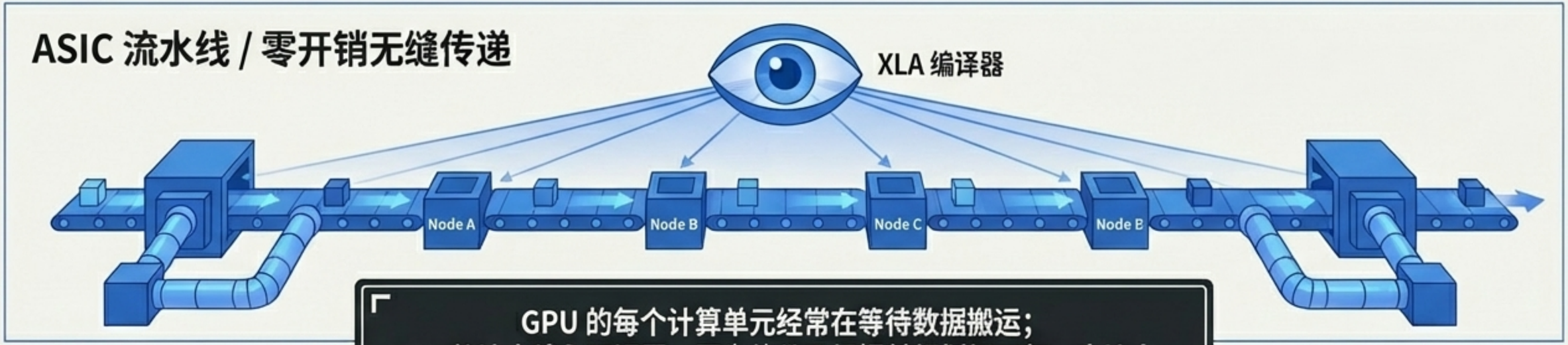
定制化之刃



全局统筹极高，为确定性工作负载压榨极限效率。

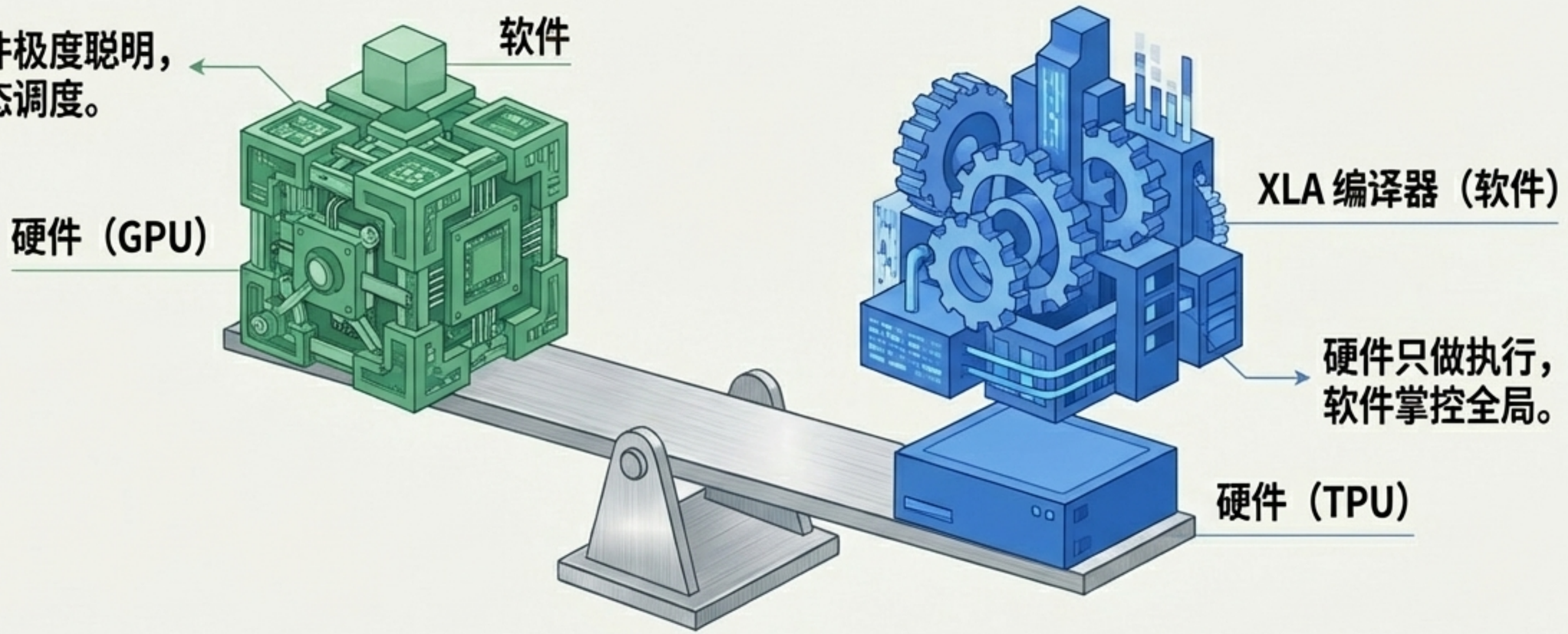
架构的本质差异：一千个独立大厨对决一条精密流水线





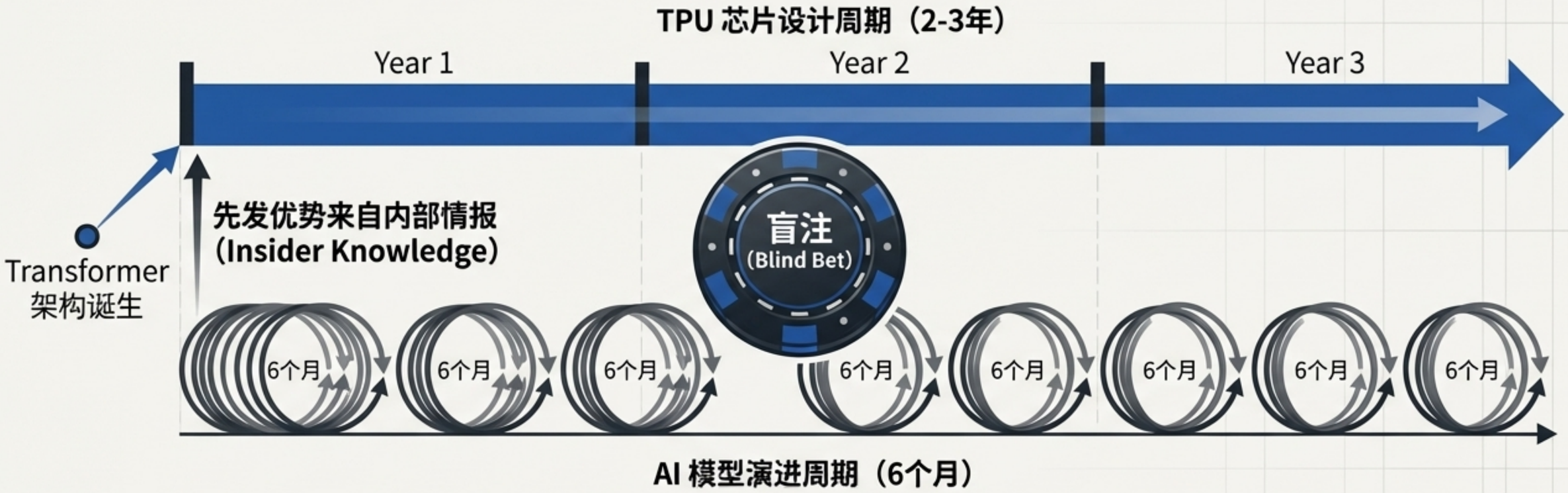
GPU 的每个计算单元经常在等待数据搬运；
TPU 的流水线永不闲置，因为软件已经提前规划好了每一步流向。

计算效率的秘密：硬件的“愚钝”需要依靠软件的“全知”来弥补



XLA 编译器具备全局上帝视角。它在系统层面执行算子融合、内存分配与数据搬运规划。
TPU 极高的理论利用率，完全建立在编译器完美解题的前提之上。

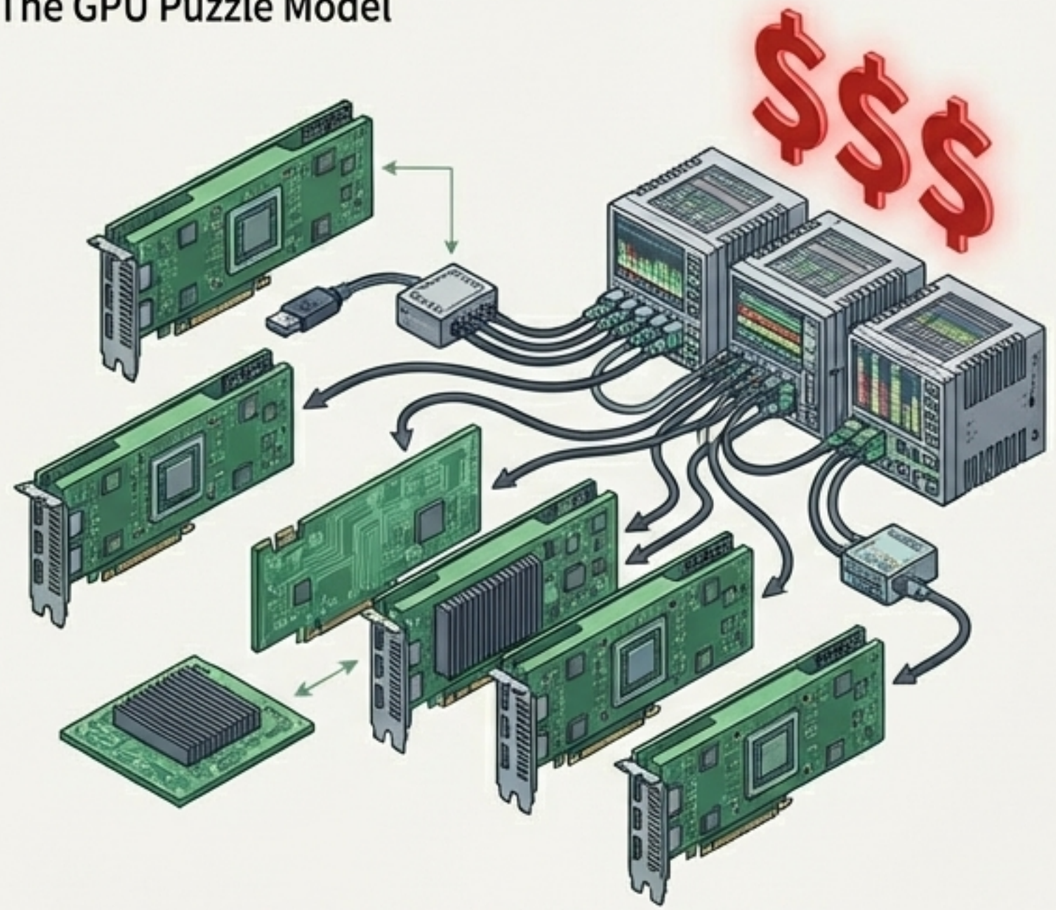
押注定制芯片的阿喀琉斯之踵：以年计的硬件研发追赶以月计的算法迭代



谷歌凭借发明 Transformer 的先知优势，在 V6/V7 时代迅速调转船头重仓大模型。但如果下一个 AI 范式不再是 Transformer？通用性极强的 GPU 将瞬间拉开身位。

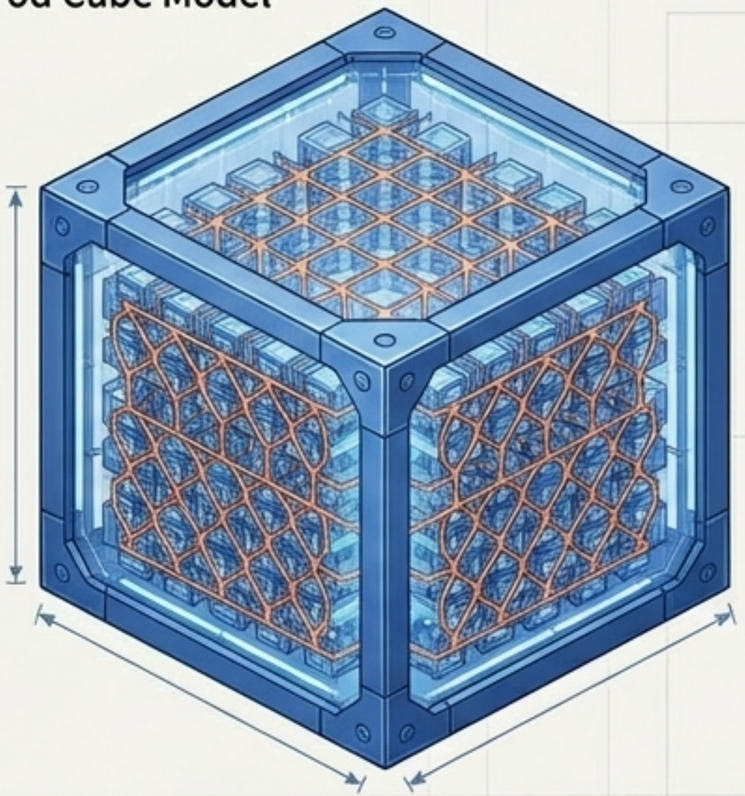
算力壁垒的升维：真正的产品不是单张加速卡，而是整个物理集群

The GPU Puzzle Model



高昂的网络基建成本

The TPU Pod Cube Model

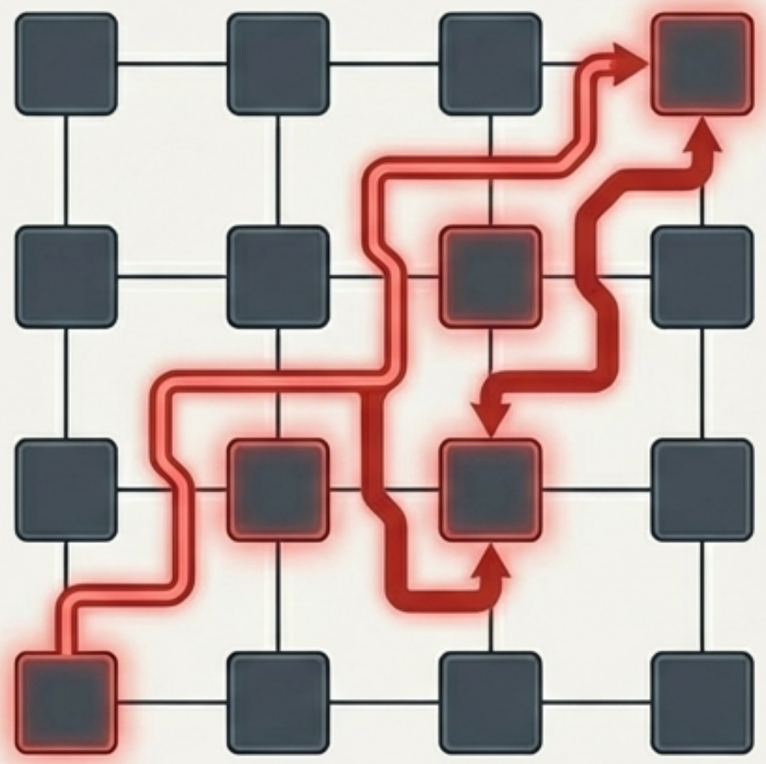


通信成本骤降：摒弃昂贵的外置交换机，集群内部采用纯铜线直连。

TCO 优势显著：在训练 Gemini 级别的超大模型时，高度协同的 TPU 集群总拥有成本大幅低于拼装式 GPU 集群。

硬件追着算法跑：MoE 模型如何逼迫系统网络拓扑实现三维升维

Before



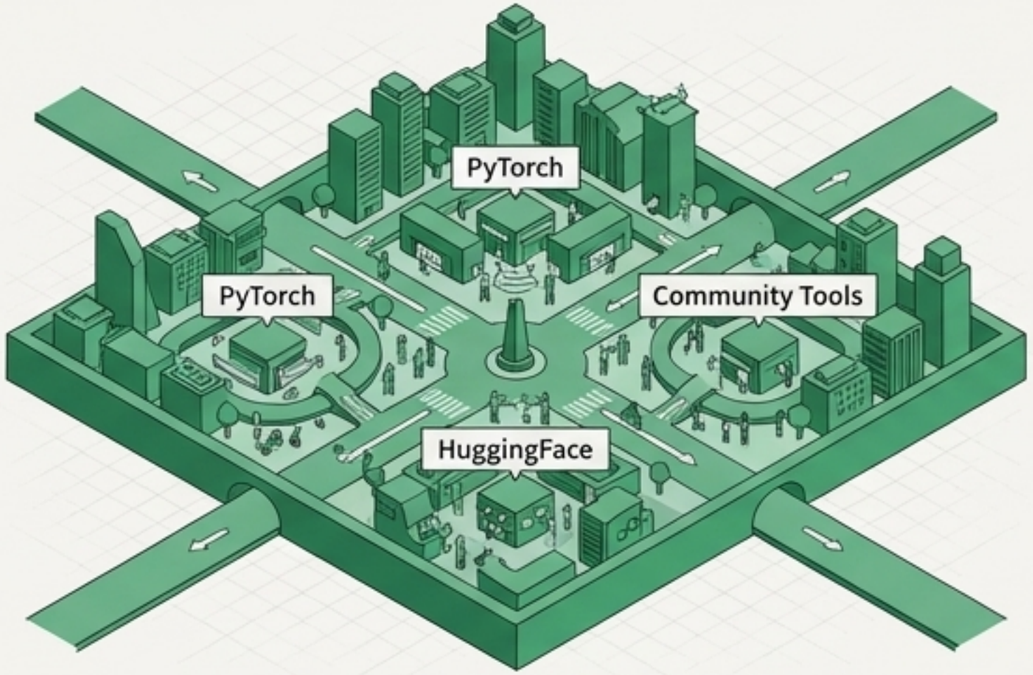
V4 之前（2D Torus）：跨芯片寻找专家模型，通信严重拥堵

After

V4 时代（3D Torus + OCS）：单步空间直达

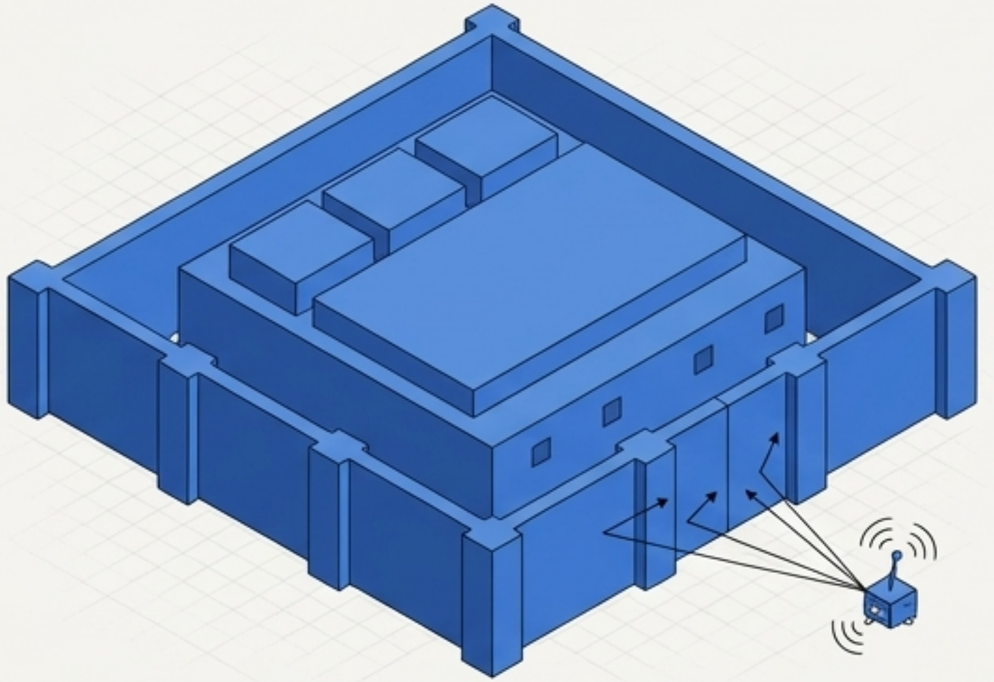
引入可编程光纤交换机（OCS）后，TPU 系统可以根据模型需求动态重构物理通信路径，成功解锁 MoE 架构的高效训练。

看不见的生态高墙：开放的 CUDA 帝国对决封闭的 XLA 黑盒



Nvidia CUDA

默认的行业标准，极低的学习与调试门槛。

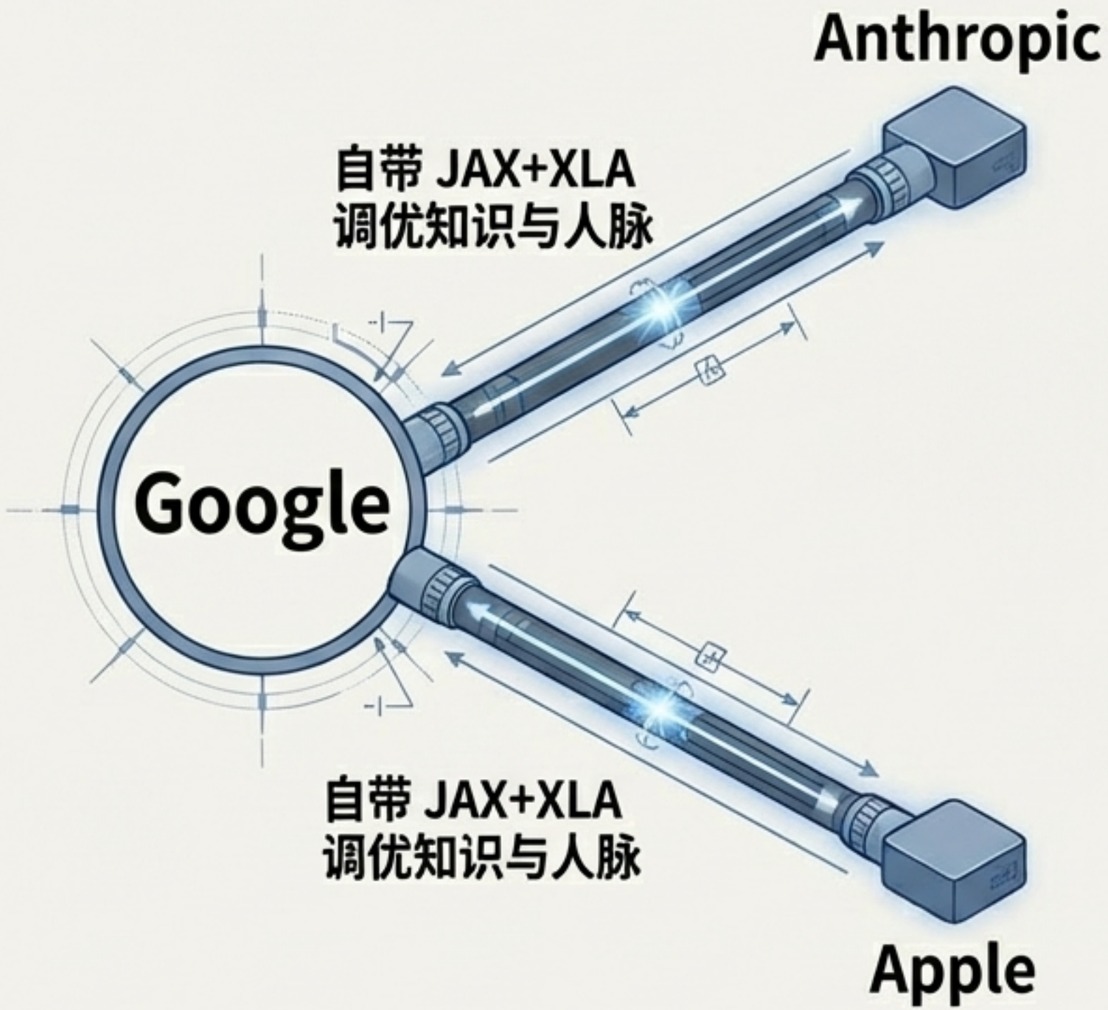


Google JAX+XLA

静态编译器的全局调优黑盒，外部开发者极难独立 Debug。

CUDA 的护城河不仅是代码，是全世界工程师的肌肉记忆。
而面对 XLA，出了 bug 你通常只能去求助谷歌的工程师。

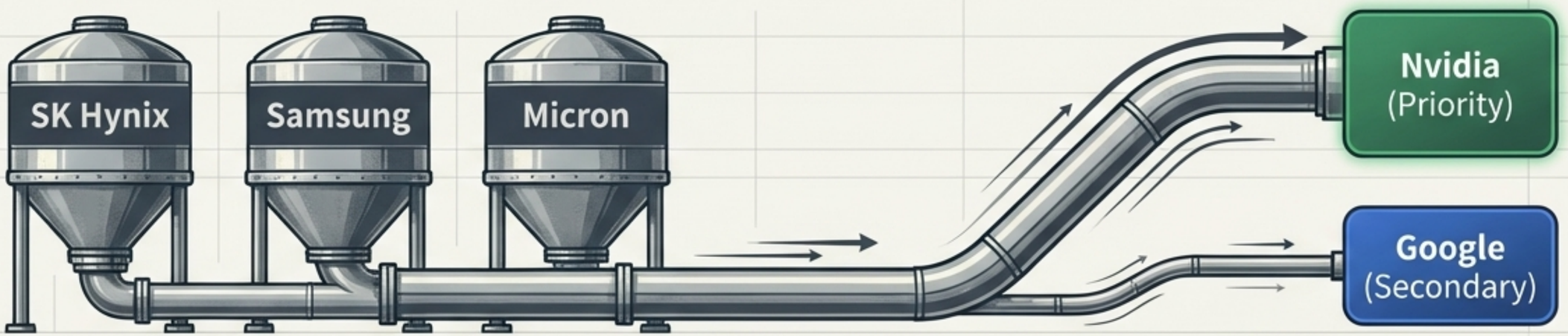
核心人才即 API：外部公司驾驭 TPU 的关键在于“溢出的谷歌基因”



公有云利用率陷阱：普通外部客户直接在谷歌云调用 TPU，因缺乏底层调优能力，实际算力利用率往往仅有 50-60%，却需支付 100% 的费用。性价比甚至不及 GPU。

绕不开的物理法则：主导未来算力格局的底层供应链双雄

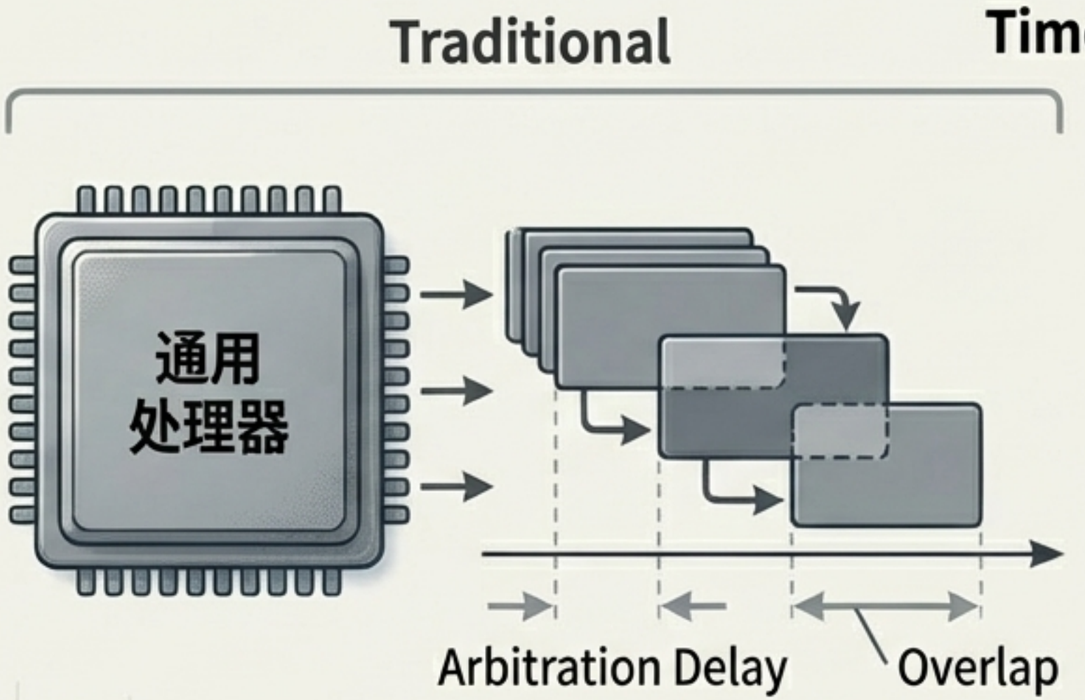
存储墙（HBM 产能）

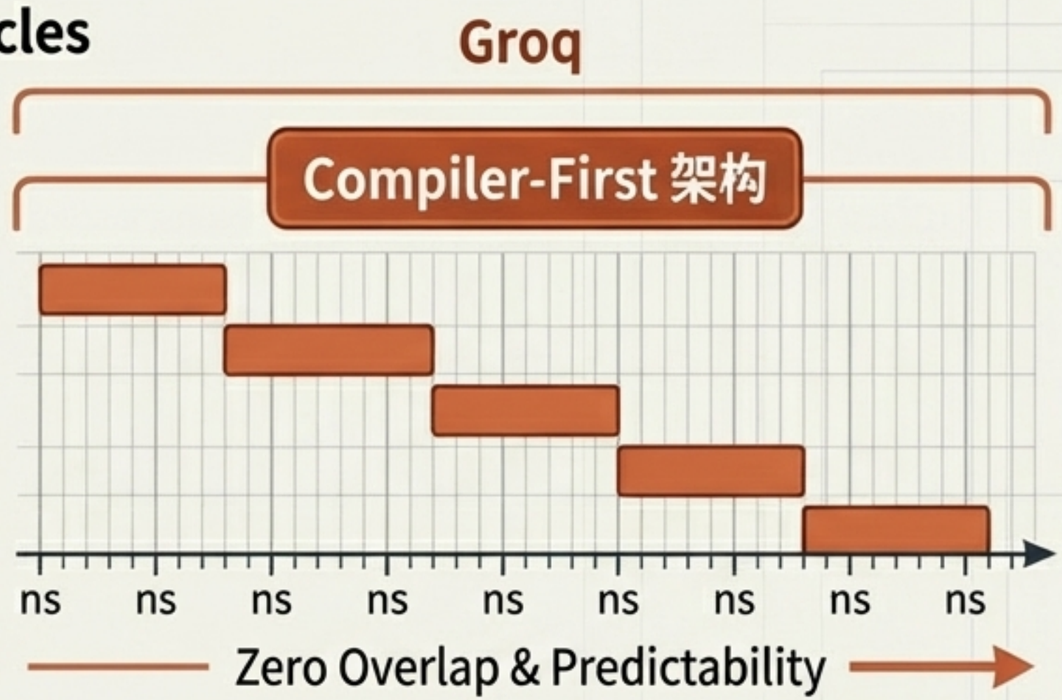


封装与良率（CoWoS & Yield）

再完美的芯片设计，卡在台积电 CoWoS 封装和存储巨头产能上就是零。Nvidia 凭借庞大的订单量，提前 1-2 年锁死了全球绝大部分关键供应链。

极速响应的破局者：Groq 在低延迟推理与 Agent 时代的生态位





底层逻辑

由前 TPU 核心成员创立，本质是“编译器公司而非芯片公司”。硬件比 TPU 更极致的“蠢”，全靠软件精确排布每一个时钟周期。

场景契合

精准踩准 Agent 元年。在多步思考调用的 Agent 链条中，毫秒级延迟的消除决定了产品的生死，这正是 Groq 的绝对主场。

终局推演：AI 基础设施将彻底走向多元分层，告别一统天下

泛用型霸主
Nvidia GPU

- 灵活性：高
- 软件生态：开放标准
- 最佳场景：未知工作负载 / 极高通用需求
- 规模化 TCO：偏高

确定性怪兽
Google TPU Pods

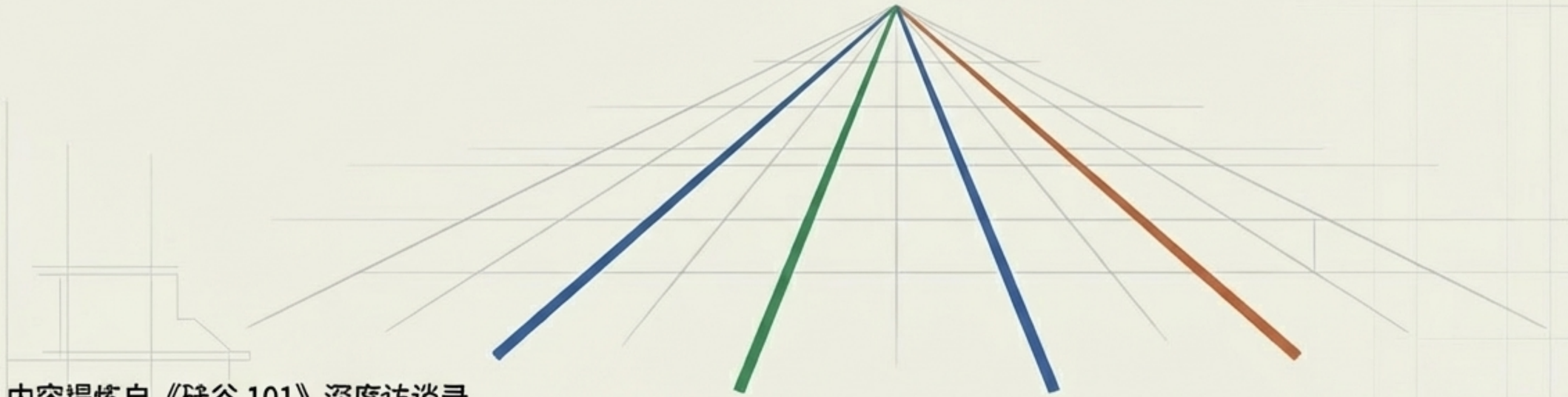
- 灵活性：低
- 软件生态：封闭黑盒
- 最佳场景：已知大模型训练 / 需精英团队支撑
- 规模化 TCO：极低

毫秒级利刃
Groq LPU

- 灵活性：极低
- 软件生态：定制化
- 最佳场景：超低延迟推理 / Agentic Workflows
- 延迟表现：行业最优

核心启示 | 应用层开发者与投资者必须抛弃“All-in 单一硬件平台”的执念。未来的算力蓝图，是依据 “训练 vs 推理”及“吞吐量 vs 延迟”划定的分层割据战。

“技术往往为了解决一个问题，而创造一种全新的生态。”



内容提炼自《硅谷 101》深度访谈录
嘉宾：Henry（前谷歌 V7/V8 TPU 核心研发工程师）
视觉与架构梳理：The Compute Blueprint 研报系列