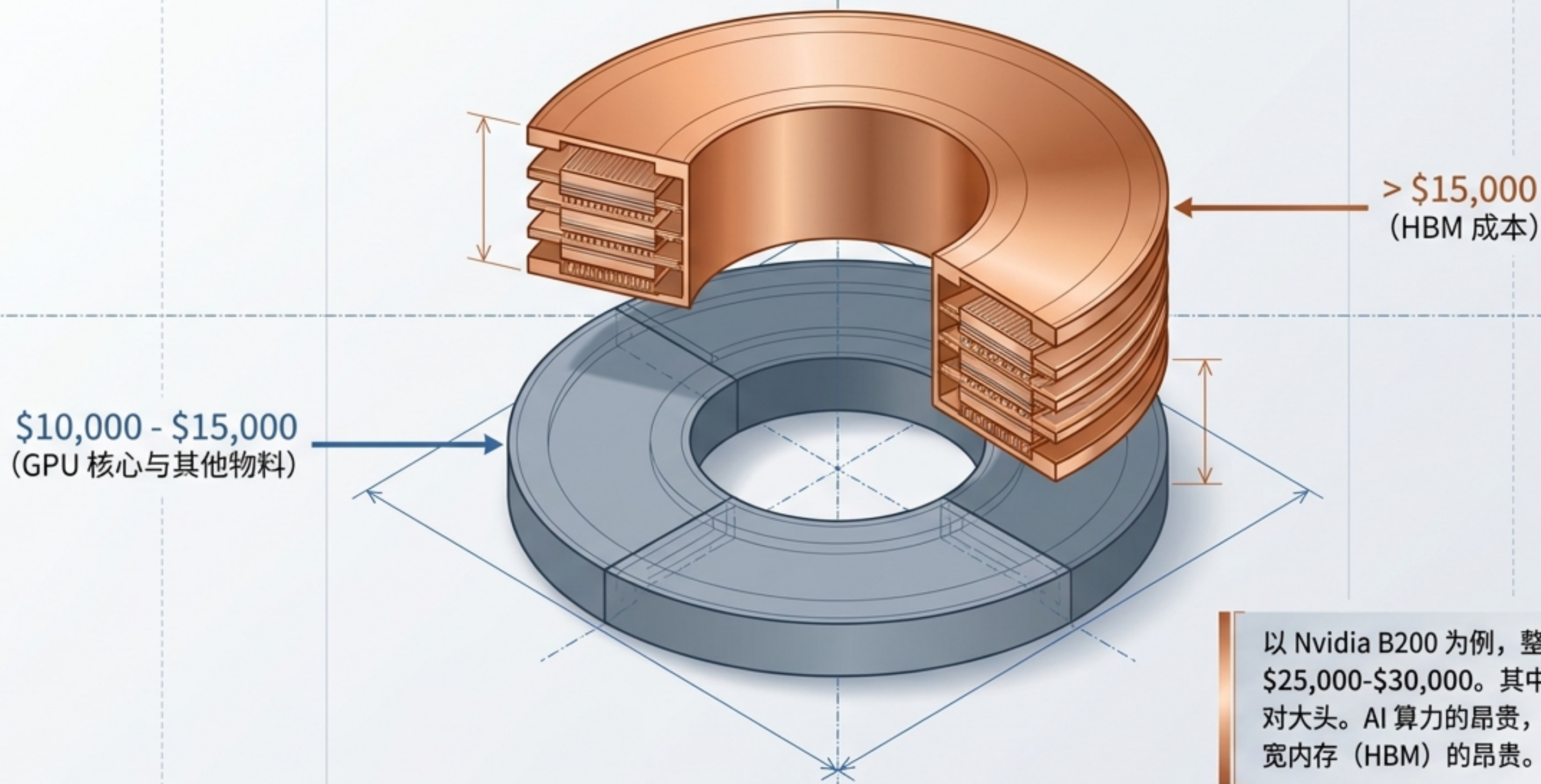


HBM：悬在 AI 算力头顶的达摩克利斯之剑

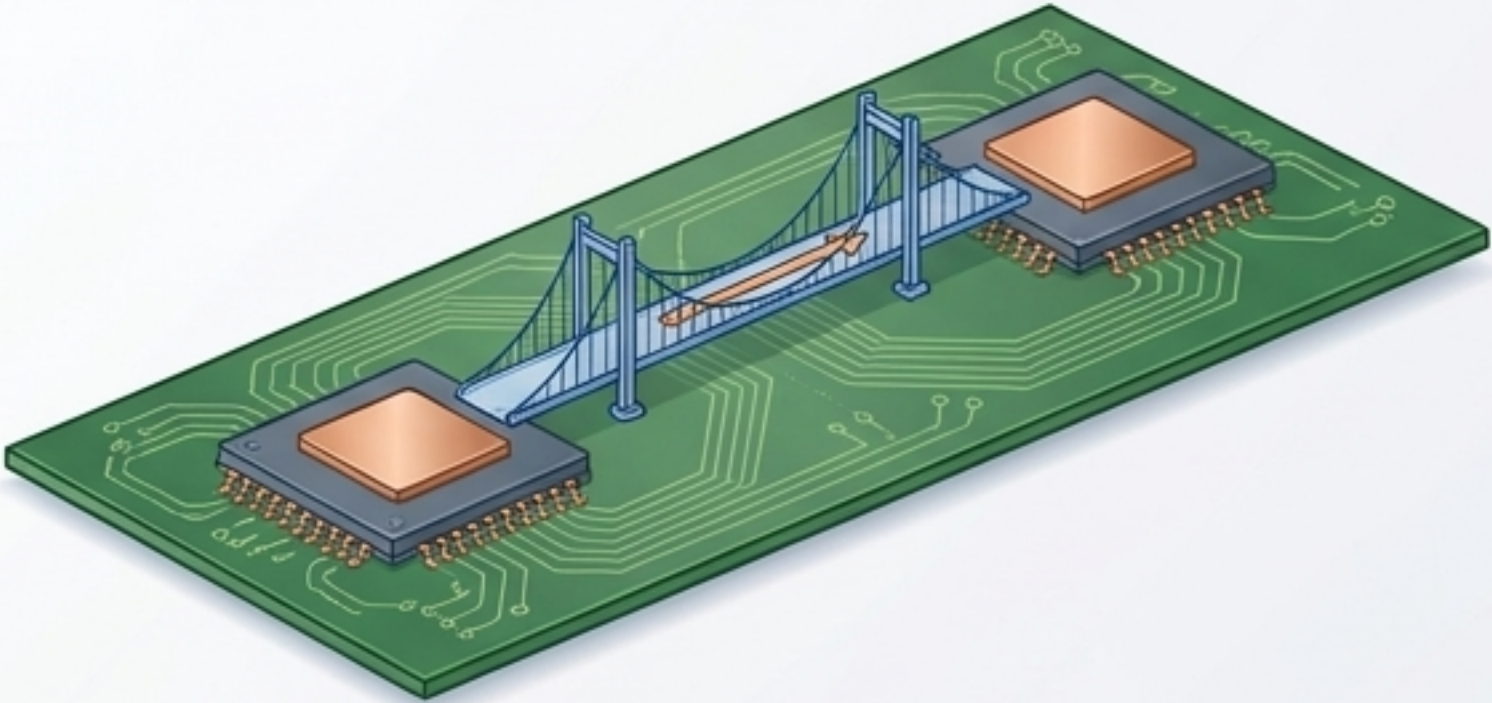
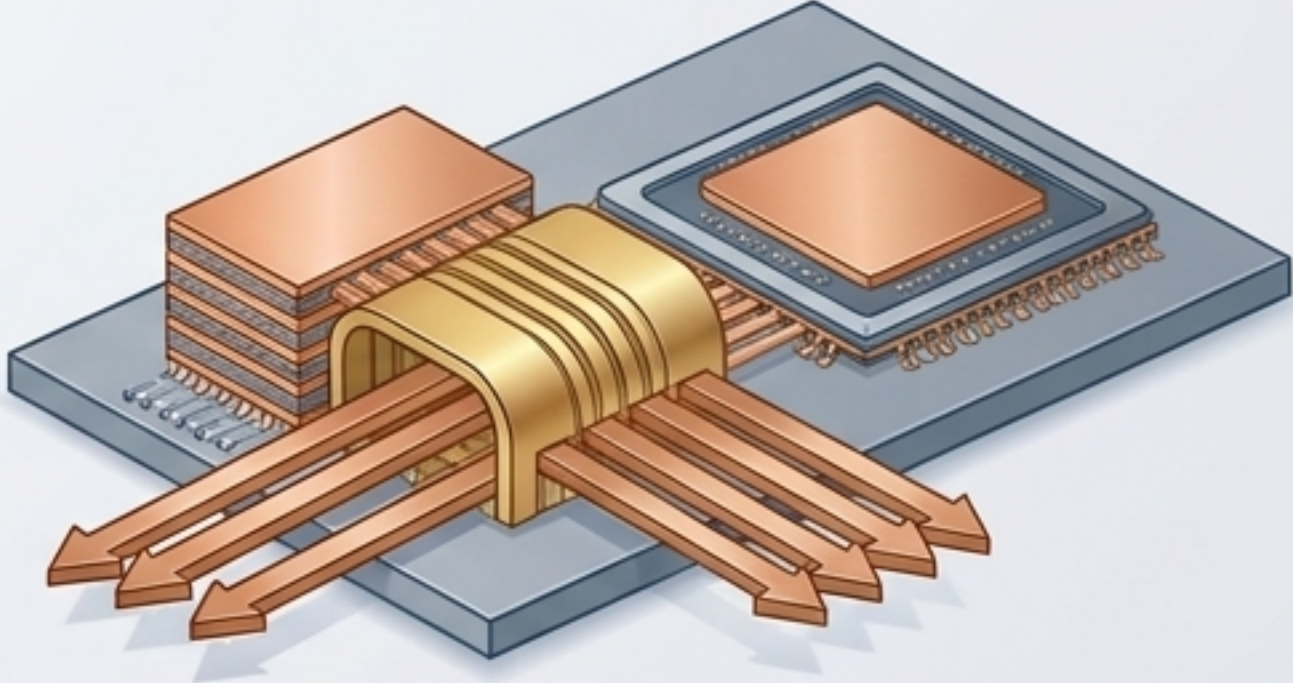
解析高带宽内存的物理极限、成本真相与 SK 海力士的垄断帝国

为什么你买的不是 GPU，而是附赠 GPU 的内存条。

旗舰 GPU 成本拆解：超过 50% 是“内存税”

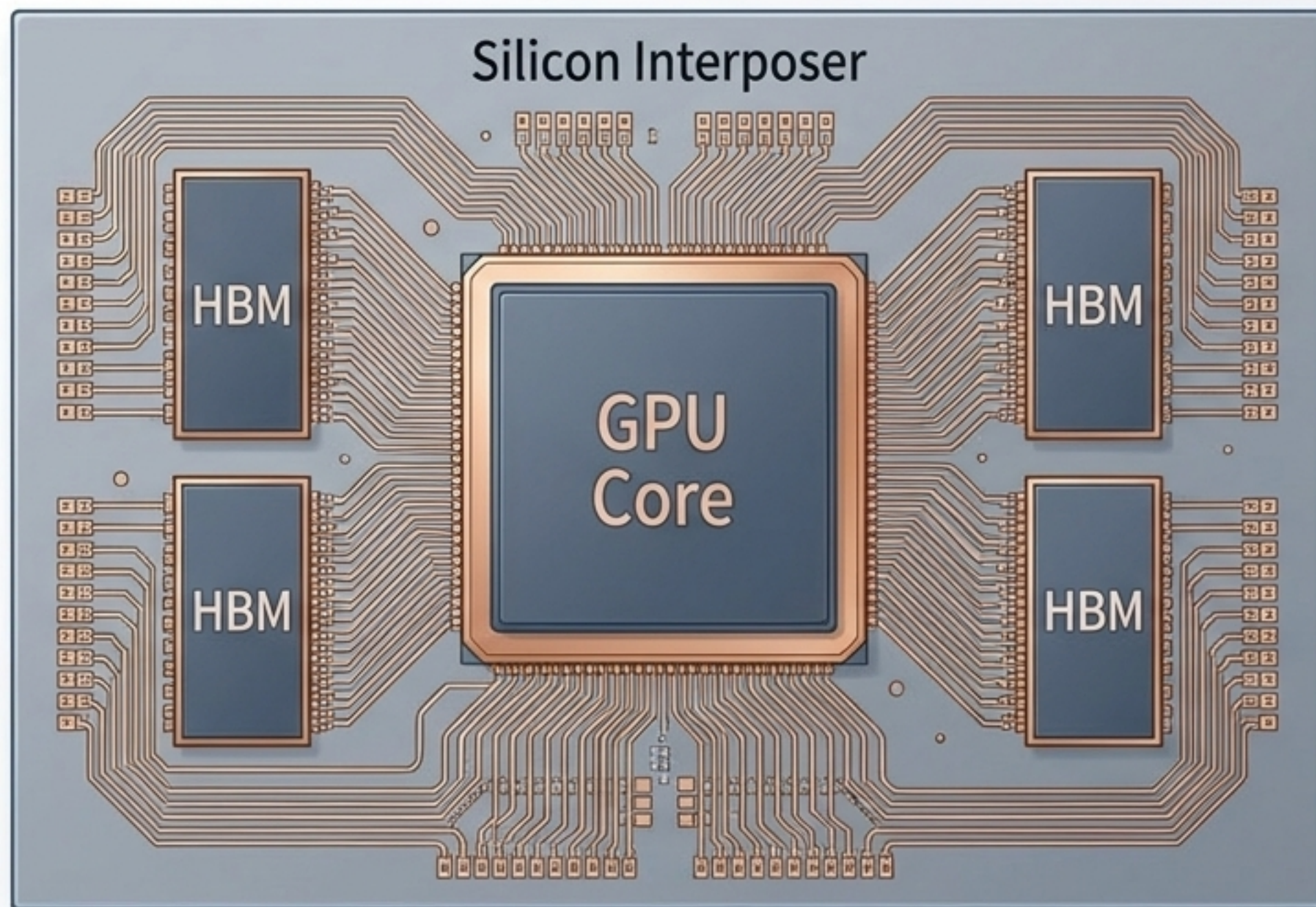


从 2D 铺平到 3D 堆叠：打破带宽之墙

GDDR (2D 铺平)		HBM (3D 堆叠)	
			
物理距离：	主板 PCB，几厘米	物理距离：	同一硅中介层，几毫米
数据总线：	32 位宽	数据总线：	1024 位宽
信号介质：	PCB 树脂材料	信号介质：	硅晶圆级介电性能

物理距离缩短至毫米级，总线位宽拉升 32 倍。带宽直接起飞。

硅中介层 (Silicon Interposer): 百美元级别的昂贵地基



← $>2000 \text{ mm}^2$ (Nvidia H100 中介层面积) →

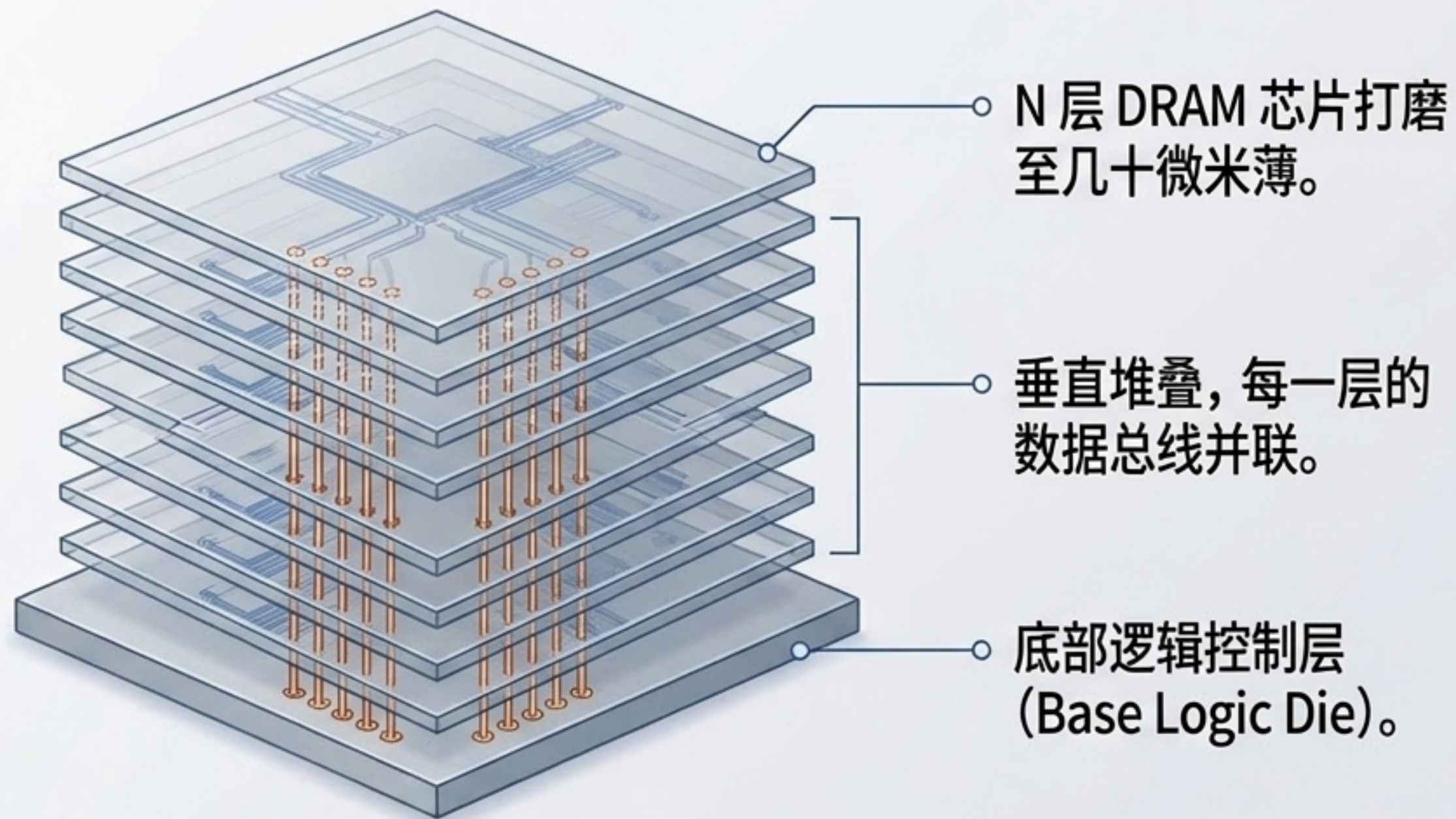
不是传统的 PCB 绿板，而是一整片昂贵的硅晶圆。

内部蚀刻微米级金属互连线，连接 GPU 与 HBM 的 I/O 接口。

极佳的信号完整性是实现 1024 位宽的物理前提。

按晶圆成本和良率计算，仅这块“地基”的成本就高达上百美元。

Z 轴架构：将内存叠成“千层饼”



HBM 的极高带宽不是靠单片芯片压榨算力，而是通过垂直堆叠实现物理并联。N 层堆叠 = N 倍单片带宽。

TSV（硅通孔）工艺：微米级“穿针引线”

1. 减薄：
晶圆打磨至极限厚度。



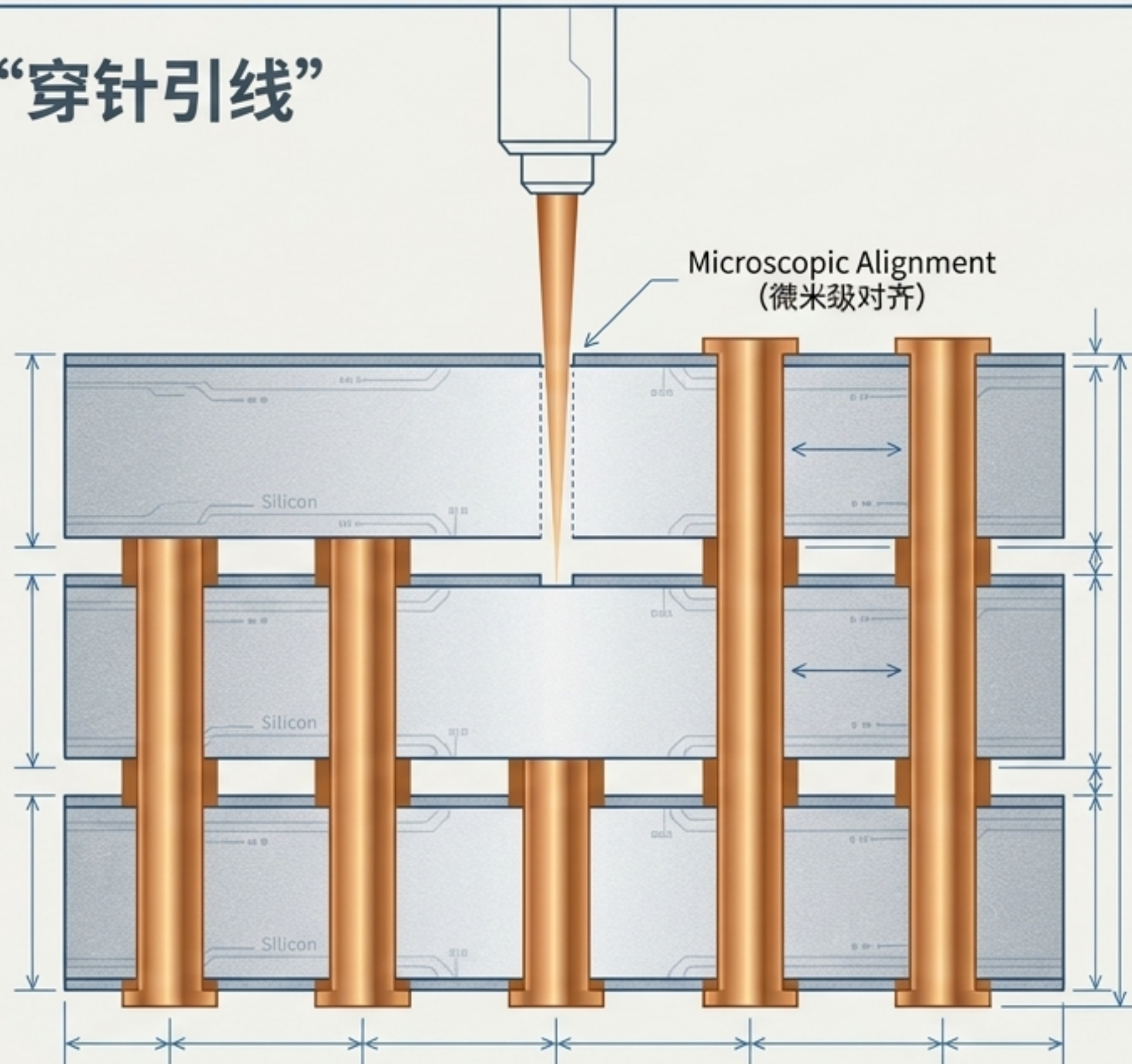
2. 激光穿孔：
在极小面积内打出数千个垂直通孔。



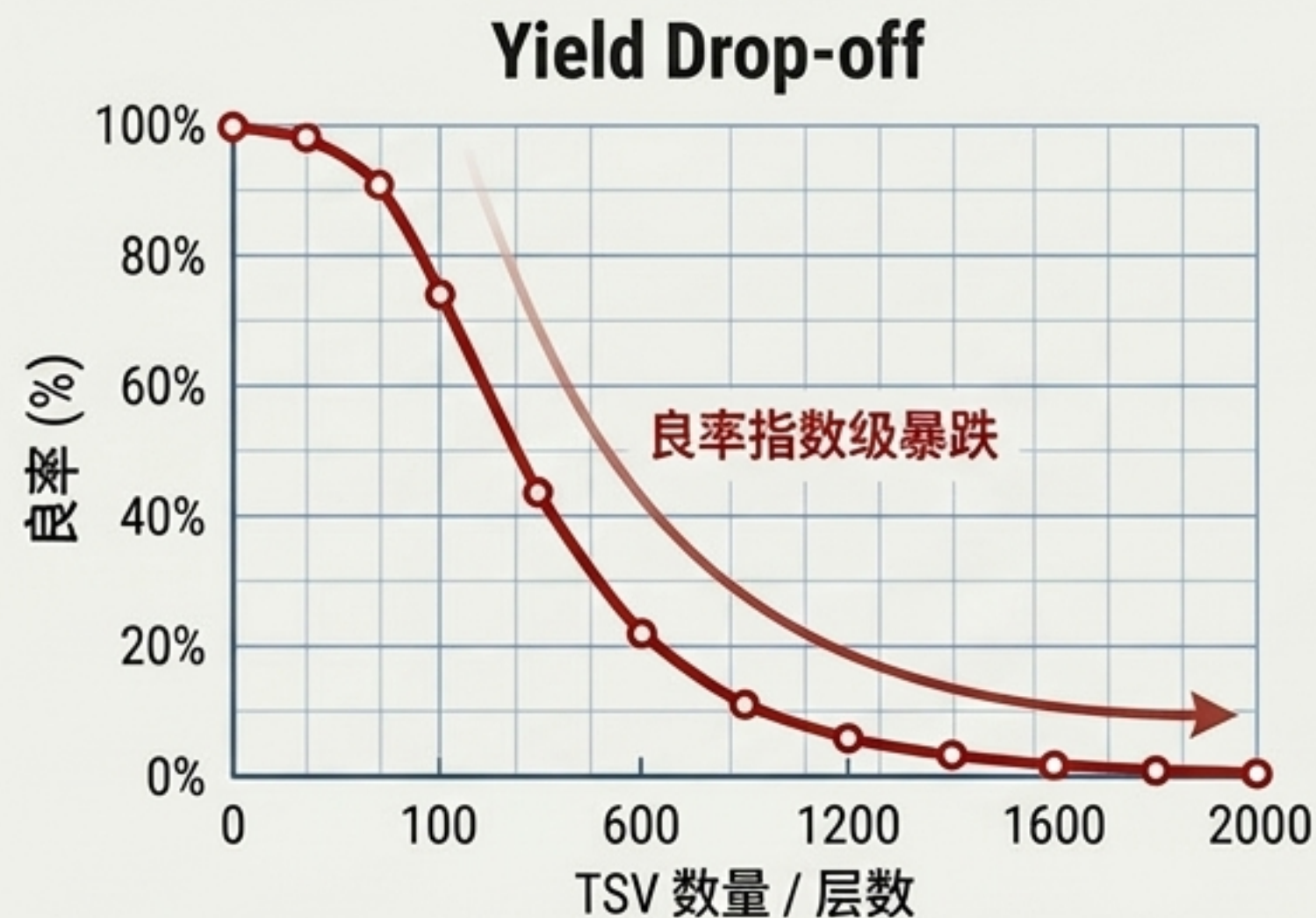
3. 铜填充：
注入导电铜柱，形成垂直数据总线。



对齐精度要求达到微米级
(Micrometer Alignment)。8
层堆叠意味着数万个铜柱必须
严丝合缝。

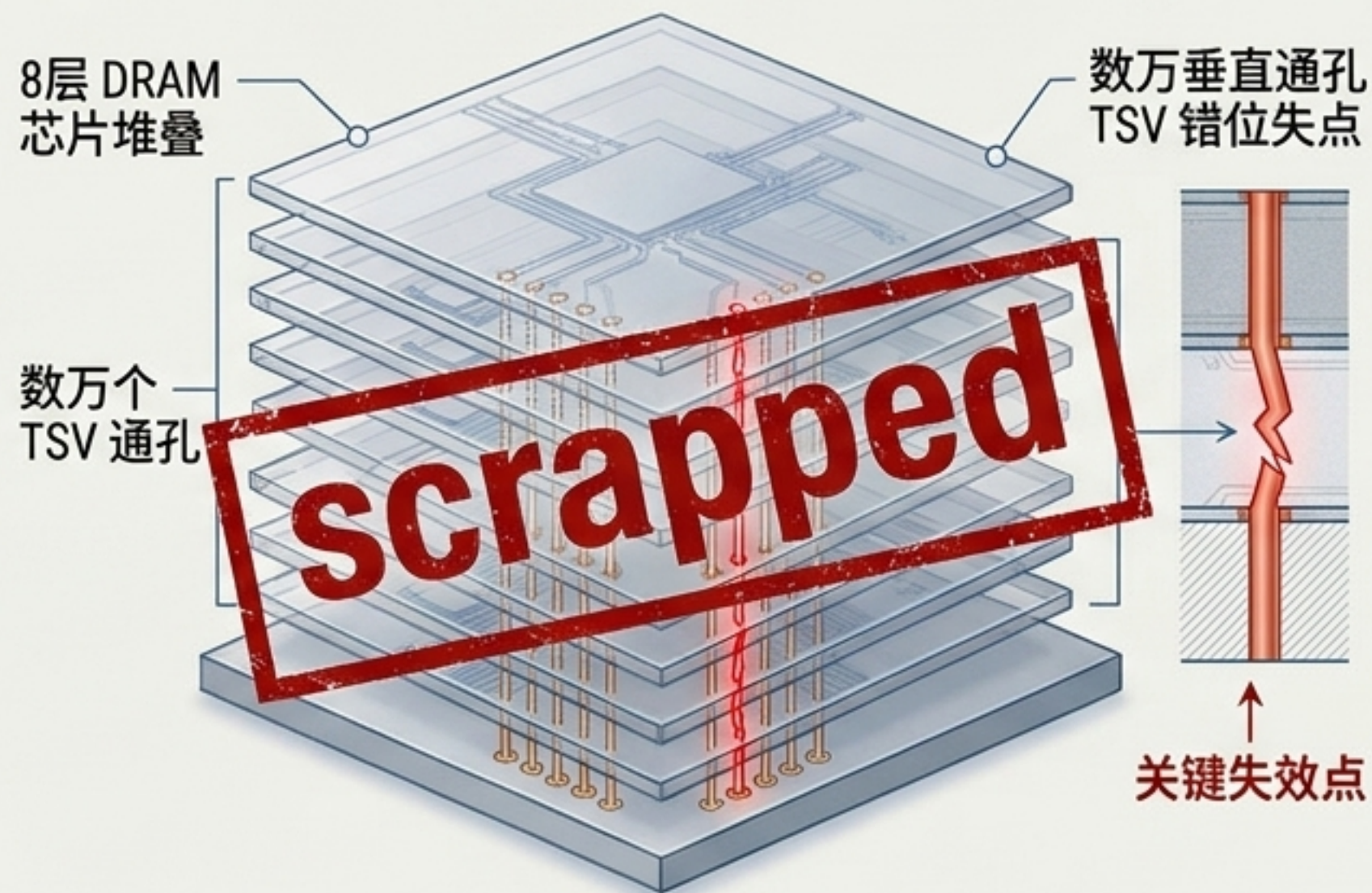


最大的成本黑洞：指数级暴跌的良率



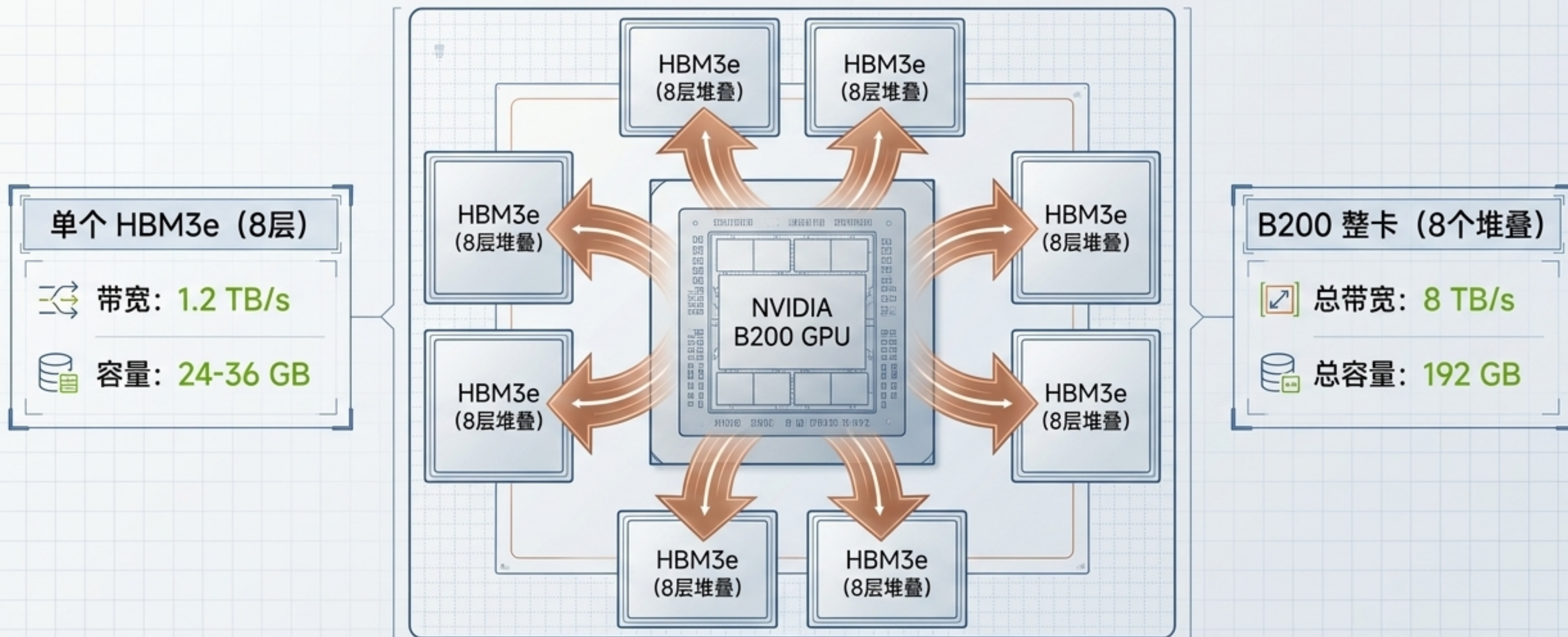
1 个坏点 = 100% 报废。

单点故障引发全盘崩溃



第一性原理分析：TSV 的良率是 HBM 最大的成本瓶颈。几千个垂直通孔穿透 8 层芯片，只要有一个通孔微米级错位或未导通，整颗价值连城的 HBM 芯片直接报废。HBM 良率远低于普通 DRAM——这是它极度昂贵的第一性原理。

当前最优解：HBM3e 与 B200 的极致重合



算力的怪兽，需要海量的粮草。H100 使用 HBM3，而 H200 与 B200 已全面转向 HBM3e 榨取极限性能。

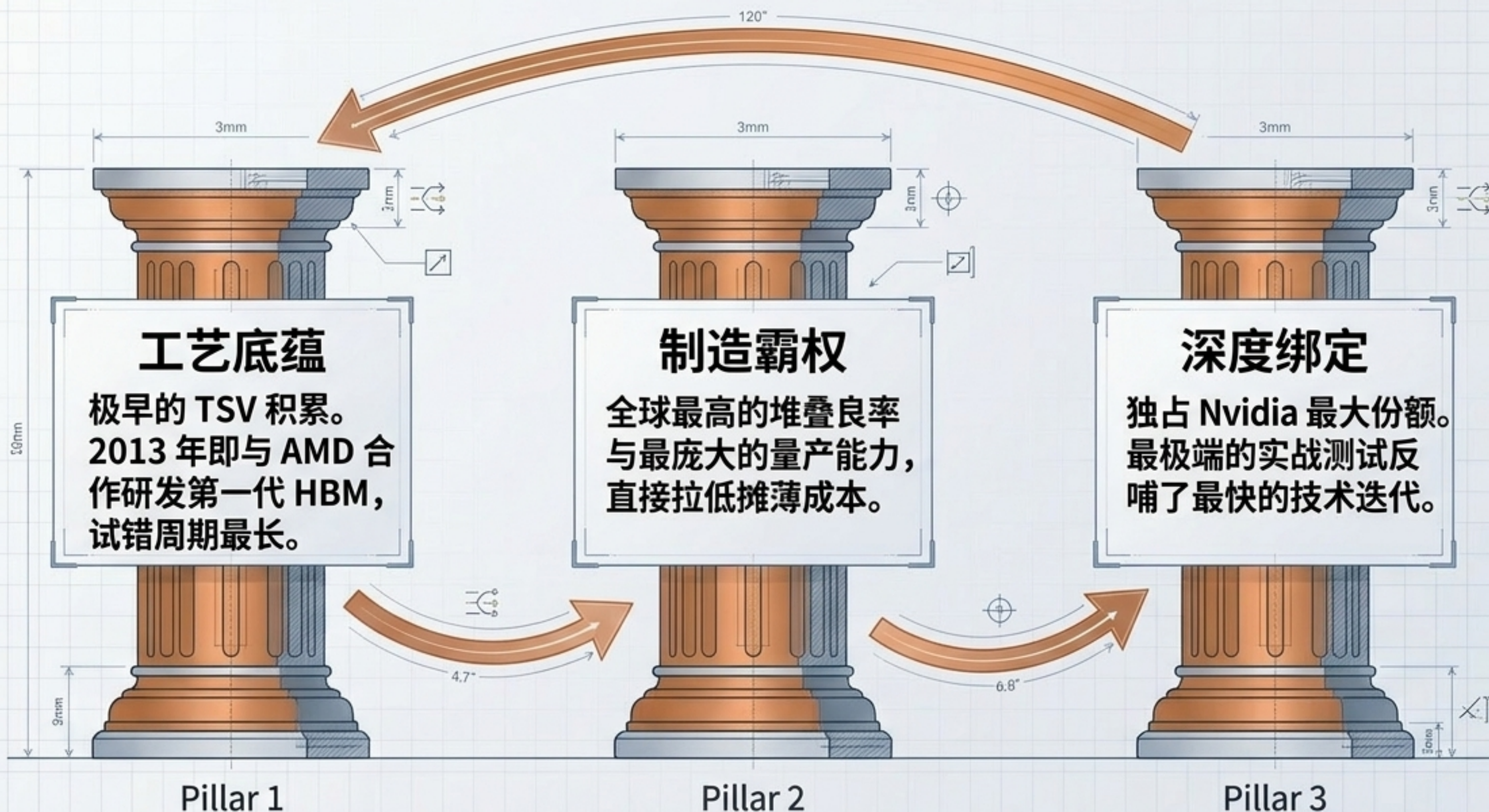
SK 海力士的绝对垄断地位



~90%
(HBM3 初期全球市场份额)

在 HBM 市场，SK 海力士拥有压倒性统治力。尽管三星和美光正在全力追赶，但核心技术差距与产能壁垒[®]至今未能实质性缩小。Nvidia 超过 70% 的 HBM 需求由海力士吞下。

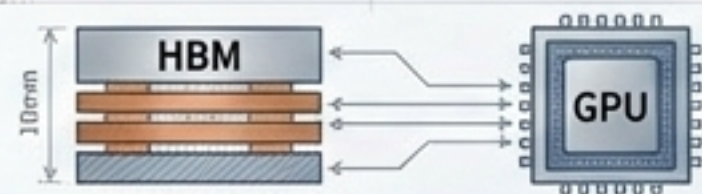
坚不可摧的护城河：从先发到绝对壁垒



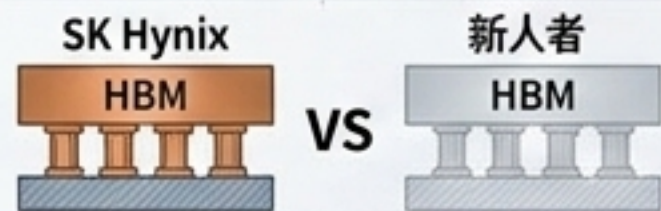
最隐蔽的壁垒：1-2 年的联合验证周期

The Lock-in Cycle

竞争者进展
(Competitor Progress)



HBM 无法即插即用。它必须与 GPU 一起进行极其复杂的联合验证与先进封装磨合。

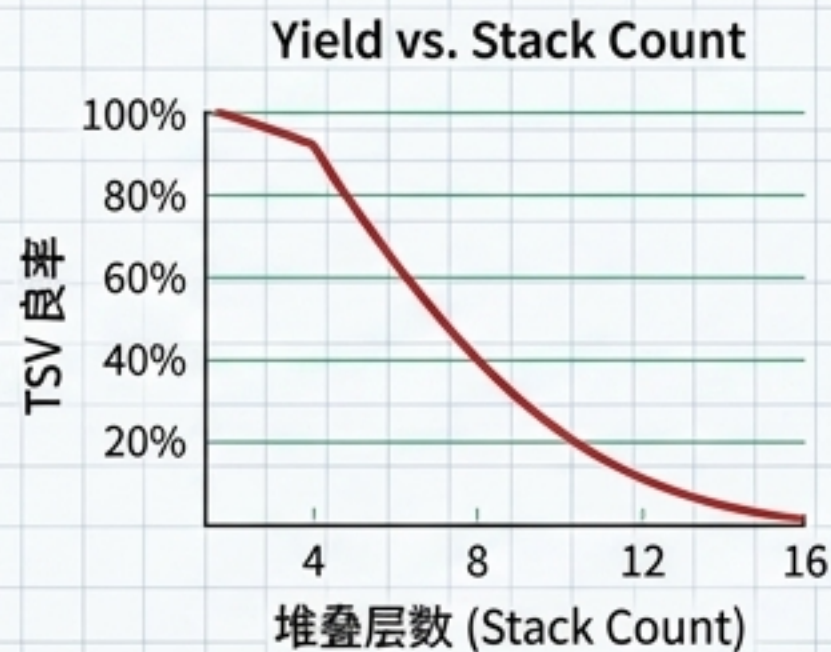
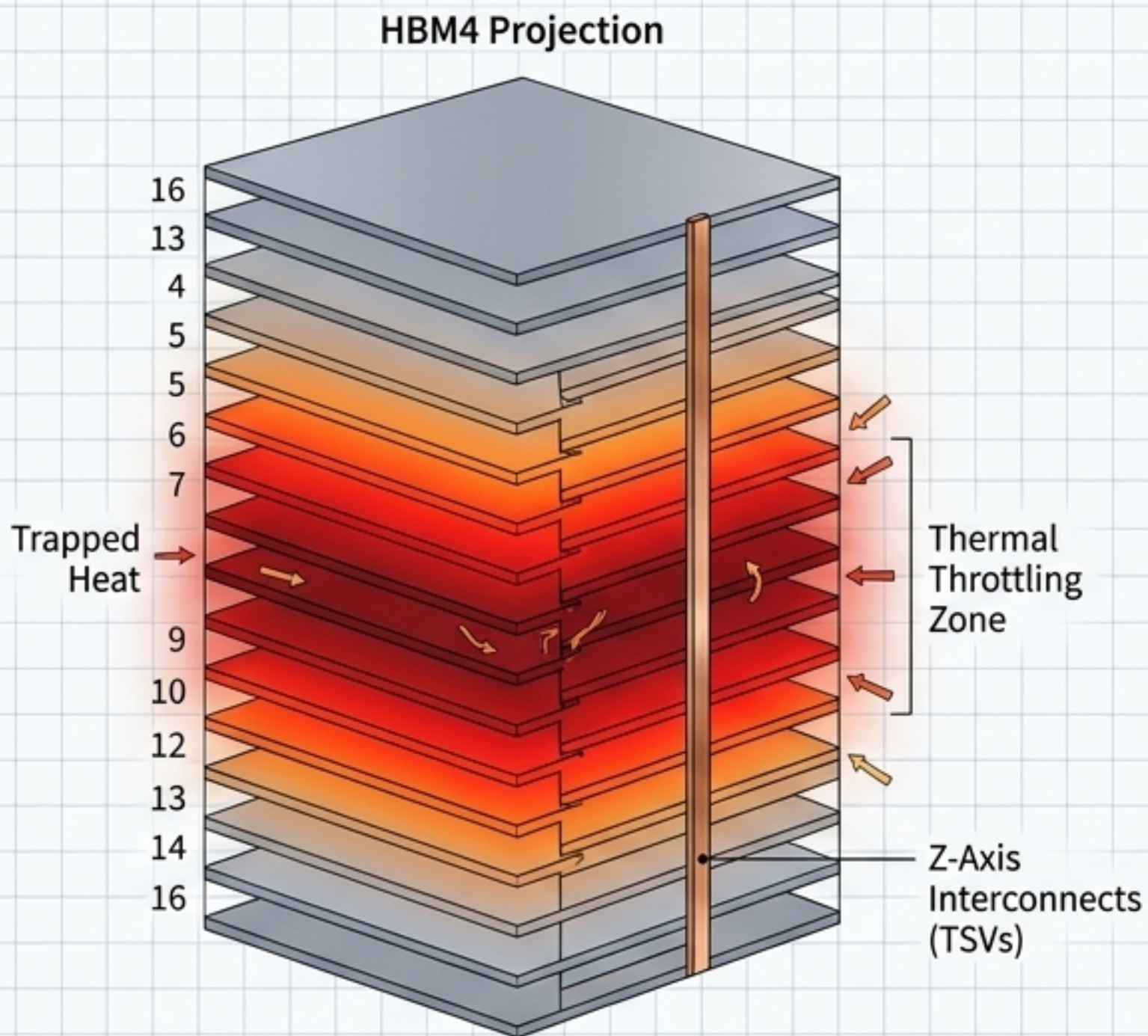


海力士已与 Nvidia 走完数代产品的磨合。后来者即使造出同等质量的 HBM，也必须重走 1-2 年的验证流程。



当你的验证走完时，市场早已被下一代 HBM 占满。一步慢，步步慢。

逼近物理极限：Z 轴无法无限向上



1 良率崩溃

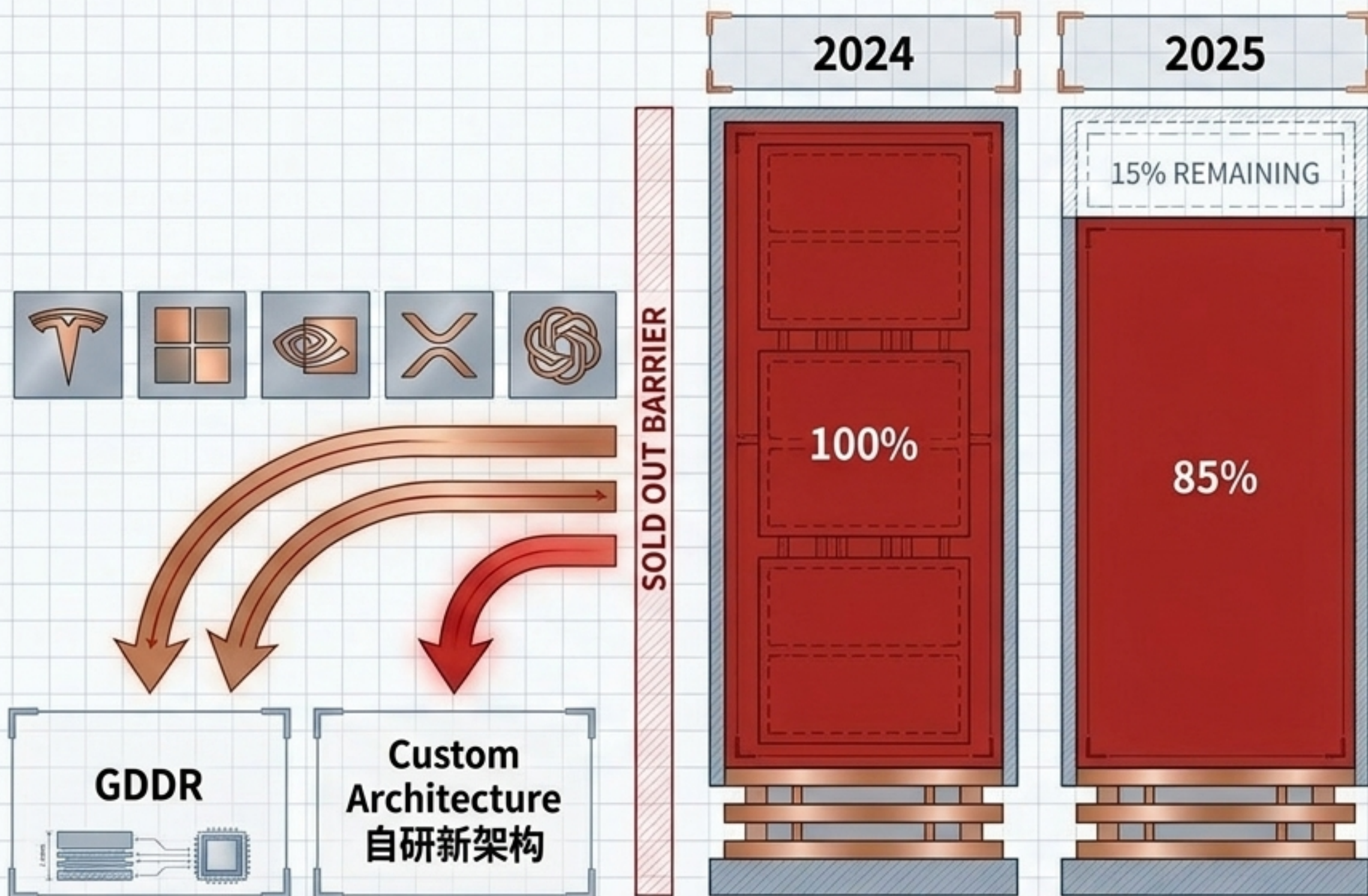
TSV 良率随堆叠层数呈指数级下降。

2 散热地狱

高发热的 DRAM 垂直紧贴发热量巨大的 GPU，热量互相烘烤，极难散发。

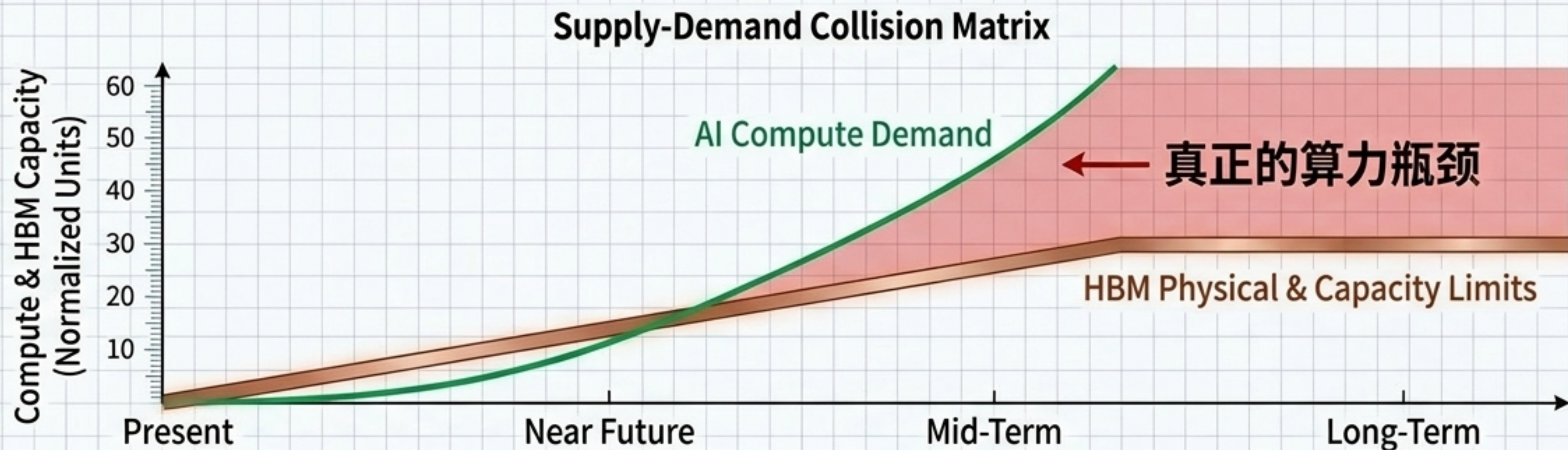
⚠ HBM4 计划挑战 16 层堆叠，但业界对其物理可行性与经济性已产生严重分歧。

产能黑洞：被买断的 2024 与 2025



- 2024 年全球 HBM 产能已被完全预定，2025 年绝大部分已被抢空。
- 产能完全跟不上 AI 需求。新进入者（如特斯拉、微软自研芯片）根本排不上供货。
- 这是迫使部分大厂转向传统 GDDR 或重金自研新架构的核心动因。

核心博弈：指数级需求撞上线性物理极限



Synthesis Box

AI 革命目前制约因素，并非 GPU 核心逻辑架构的算力上限，而是给硅片打微米级孔洞、以及垂直堆叠内存的物理极限。



Final Takeaway

HBM 很贵，且在可预见的未来会更贵。算力需求与内存物理屏障的碰撞，将是未来五年全球半导体行业最核心、最残酷的博弈。