



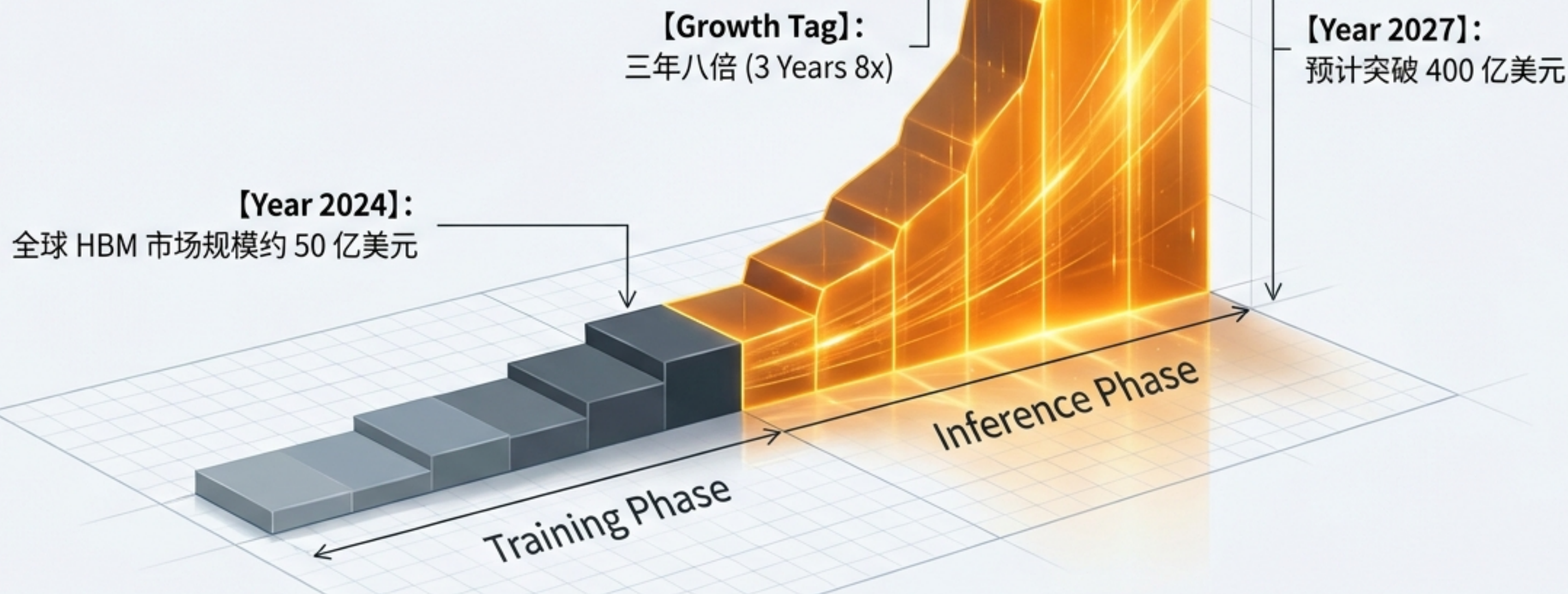
存储超周期： 算力飞轮的另一面

为什么推理阶段的物理法则正在创造 长达十年的半导体超级周期

The relentless ascent of AI inference is colliding with fundamental physical limits, driving an unprecedented demand for advanced memory solutions. This phenomenon, far from a temporary spike, is forging a decade-long "storage supercycle"-the essential counterweight to the compute flywheel.

```
[DATA_STREAM: INFERENCE_LOAD]
> DATA_THROUGHPUT: >10^21 OPS
> BANDWIDTH_REQ: 2.5 TB/S/CHIP
> LATENCY_TARGET: <50 NS
> SILICON_STACK_HEIGHT: 128 LAYERS
> ENERGY_EFFICIENCY: 0.1 PJ/BIT
> MEMORY_WALL_PROXIMITY: CRITICAL
```


惊人的增长曲线： 被忽略的真正核心



这个数字背后是一个被 **99% 的人忽视** 的事实：AI 的重心正在从**训练不可逆转地转向推理**。

范式转移：从“造模型”到“用模型”



造模型 - 训练

发生一次。耗资数亿至数十亿美元构建基础。



用模型 - 推理

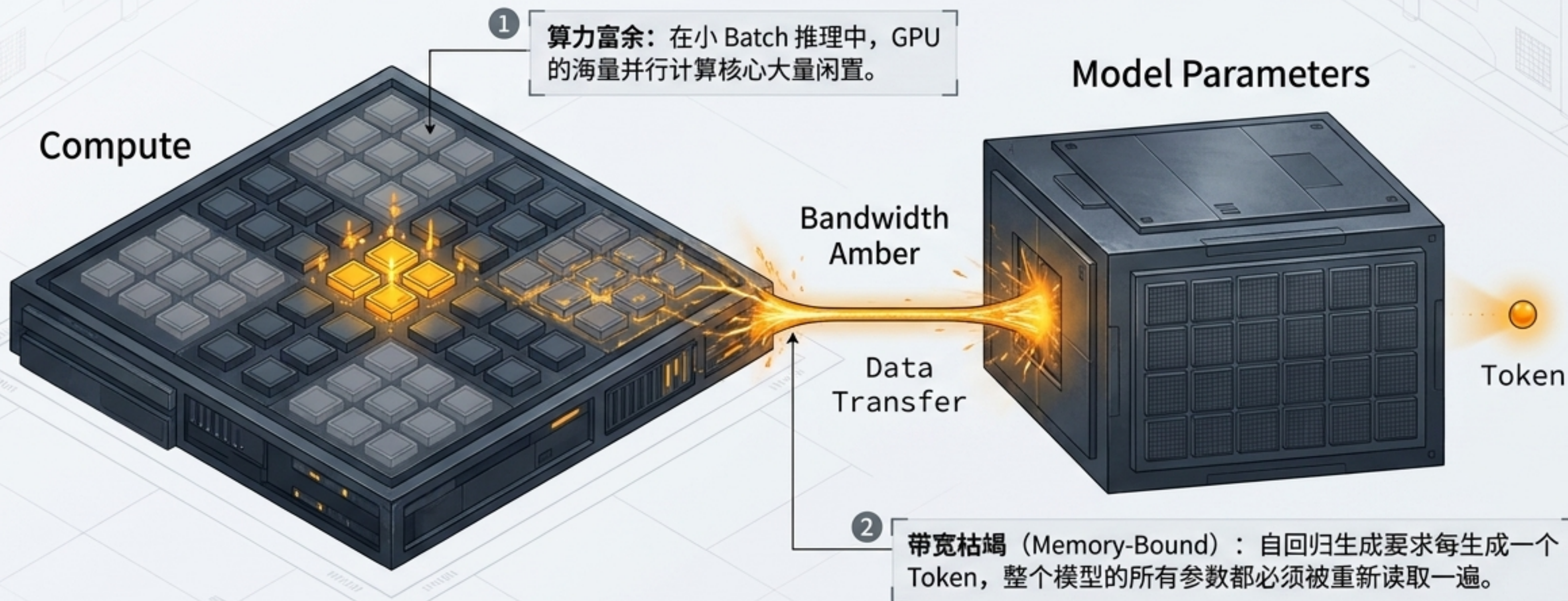
发生无数次。全世界每次调用 API 都在做推理。

Deloitte 预测：2026 年推理将占据总算力的 2/3。
推理芯片市场将从两百亿美元跃升至五百亿美元以上。

AI 工作负载矩阵：截然不同的底层逻辑

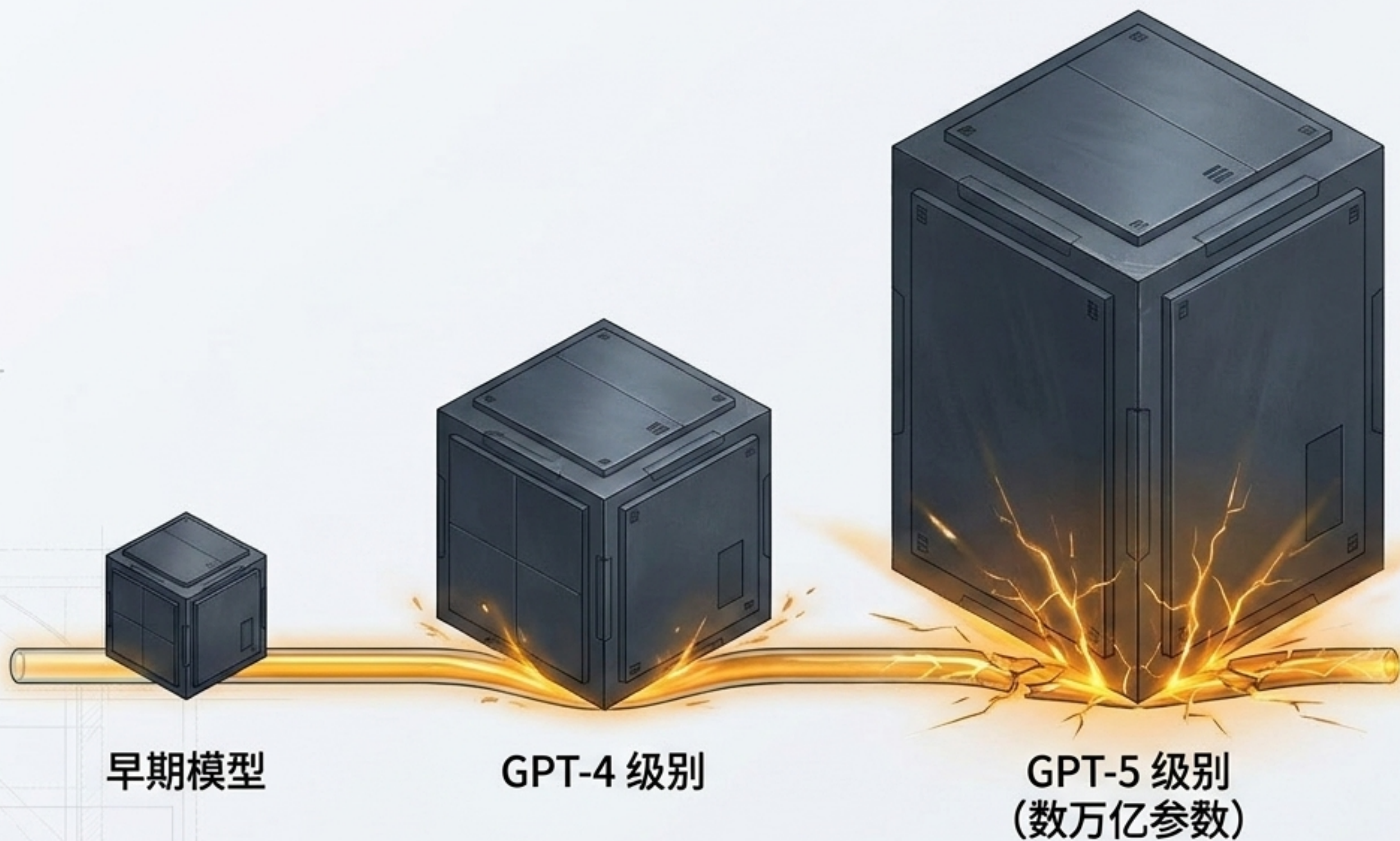
维度 (Dimension)	训练 (Training)	推理 (Inference)
频率 (Frequency)	单次极长周期 (Once)	无限次实时调用 (Infinite)
处理机制 (Processing)	并行大批量 (Large Batches)	自回归单 Token 生成 (Auto-regressive, Token-by-Token)
核心瓶颈 (Core Bottleneck)	算力/TFLOPS (Compute)	内存带宽 (Memory Bandwidth)
商业模式 (Business Model)	资本支出 (Capex)	运营成本 (Opex)

推理的物理法则：“饥饿”的 GPU



推理卡的关键指标不再是 TFLOPS（算力），而是 **HBM 带宽**。

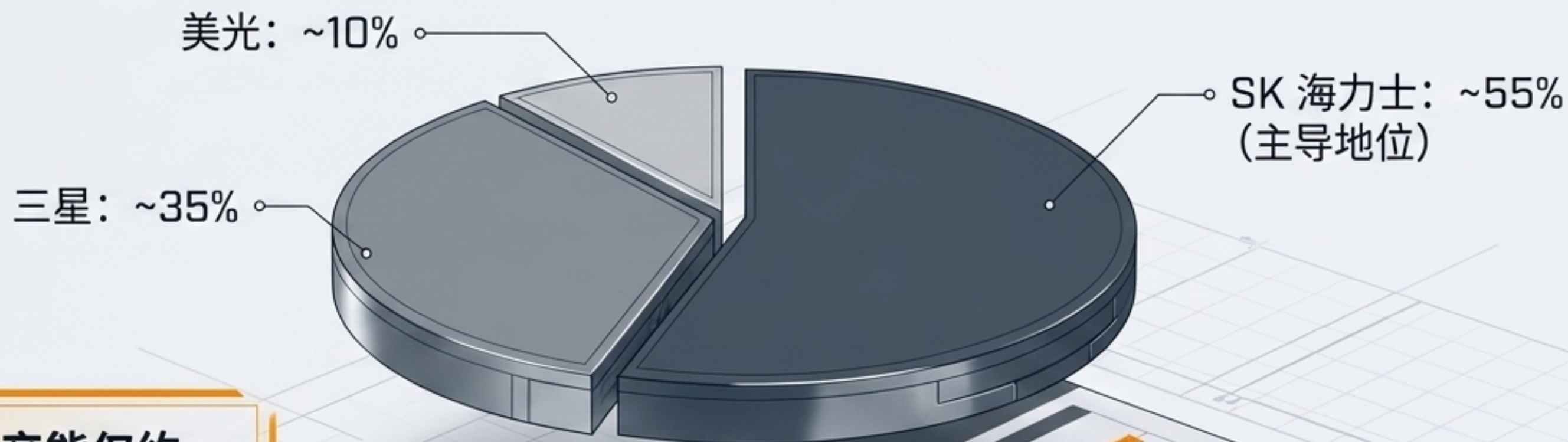
模型越大，带宽矛盾越极端



即使只推理单个 Token，
GPT-5 级别的
参数本身就要瞬间
占用数十 GB
的内存带宽。

模型参数越大，
对海量内存的绝对
依赖越深。

全球 HBM 产能博弈：极度稀缺的战略资源



全球 HBM 月产能仅约
15-20 万片 (12英寸等效)



需求远超供应——2025 年 HBM 产能已全部被预定。

真正的物理瓶颈不在 DRAM 颗粒

全球 DRAM 产能

供过于求

误区：存储颗粒短缺。
事实：全球标准 DRAM
产能供过于求。

真相：瓶颈在于 TSV
(硅通孔) 封装。



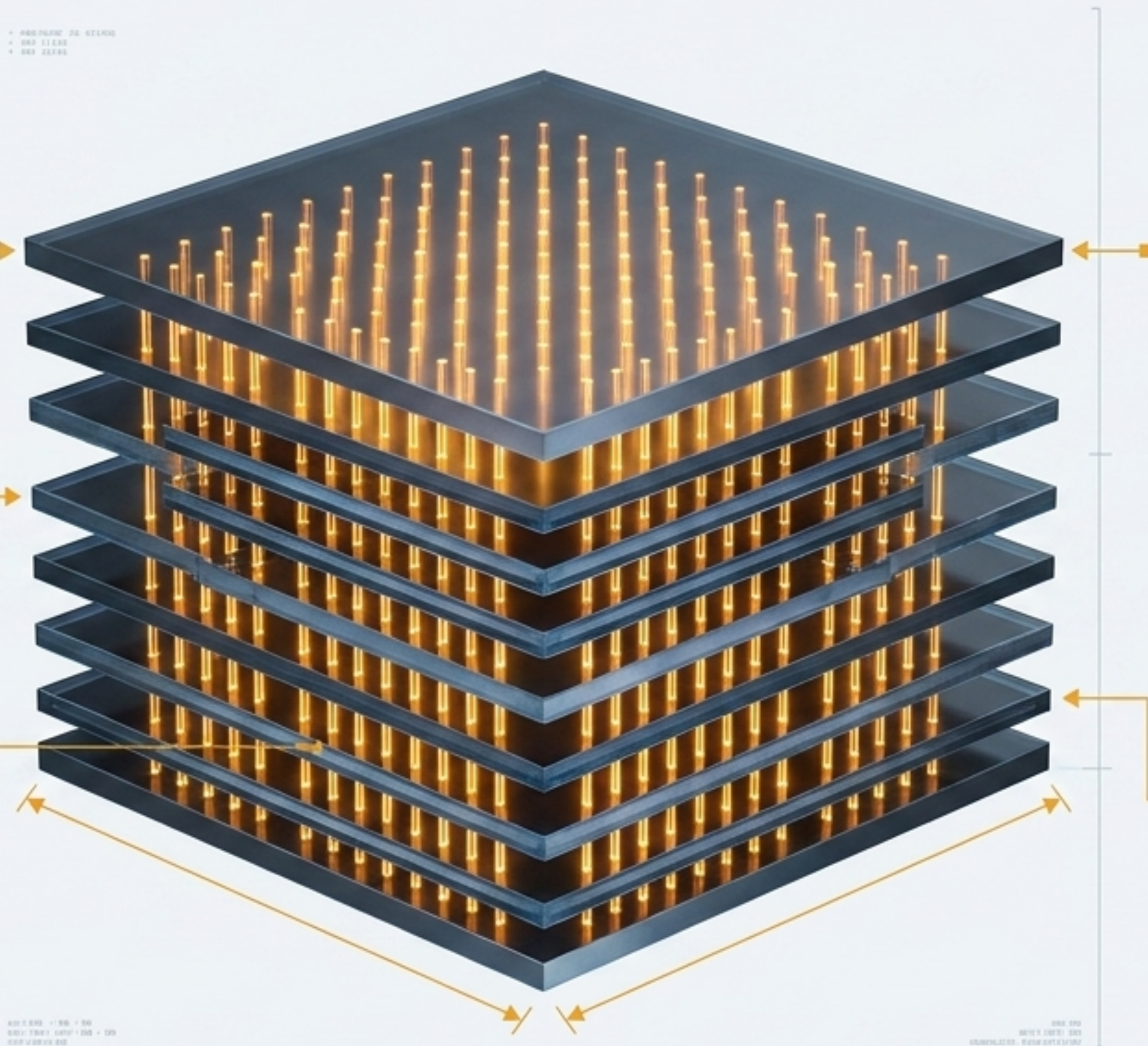
TSV 封装环节

产能扩张的非对称性：这是导致 HBM 无法像标准芯片那样快速扩产的核心原因。

解剖 HBM：TSV（硅通孔）的工程极限

多层垂直堆叠：不是平铺，而是立体堆叠。

微观“穿针引线”：每个 HBM 堆叠需要钻出并对齐几千个极其微小的硅通孔 (TSV)。

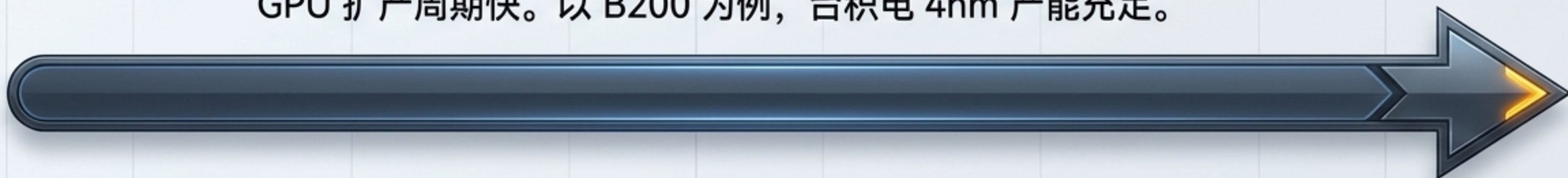


良率深渊：对对准精度和热力学稳定性的要求极高，极大地限制了良率和产出速度。

为什么 GPU 等不到它的内存？

GPU 扩产周期快。以 B200 为例，台积电 4nm 产能充足。

算力端



内存端



TSV 产能扩张周期长达 12-18 个月。

GPU 空有算力，内存却配不上。交货速度的错位，直接导致了 B200 必须等待 HBM3e 的产能释放。

地缘政治版图：制裁网中的“存储缺口”

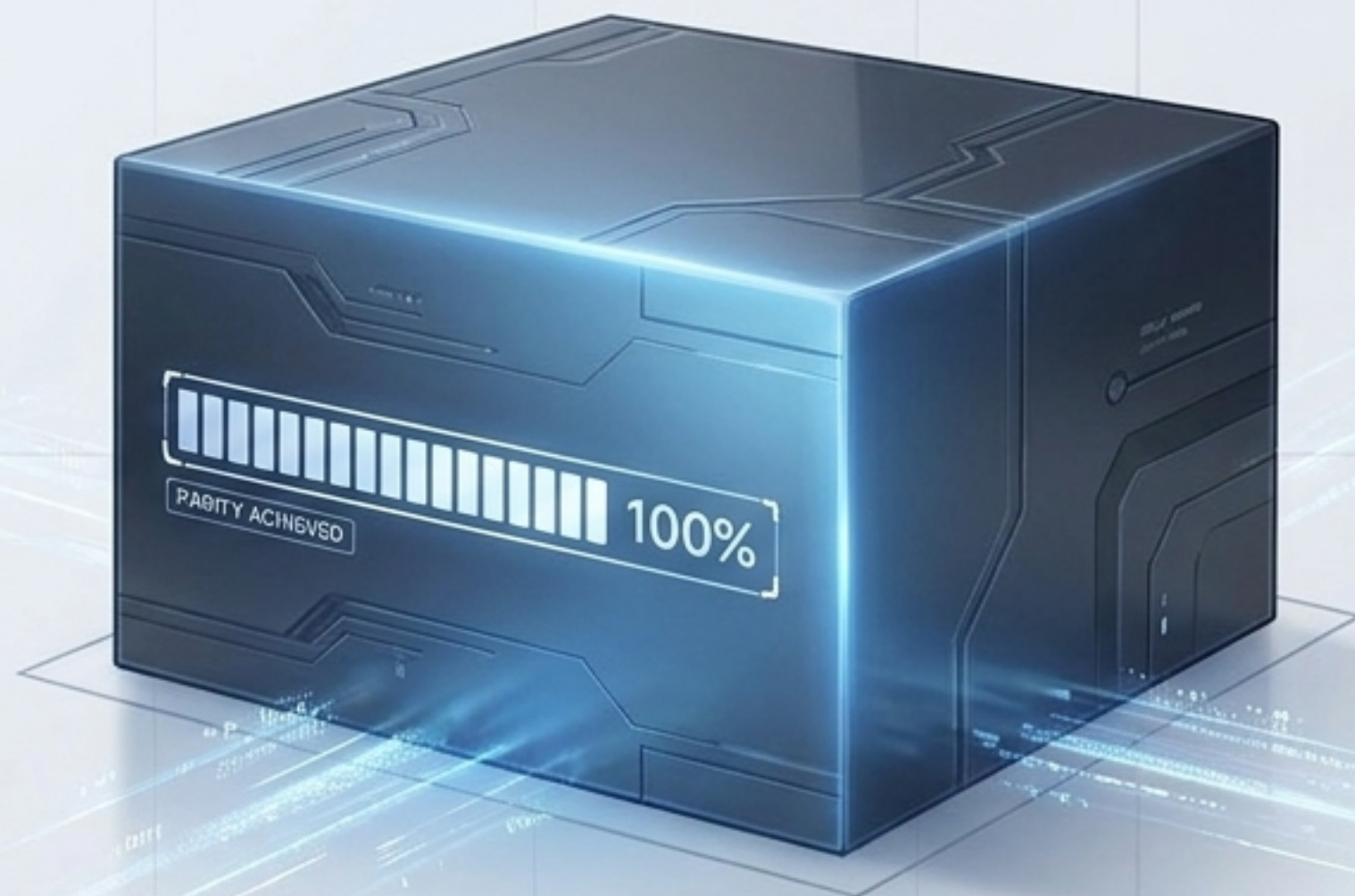


追平主流：中国在标准存储上的突围



DRAM - 长鑫存储 (CXMT)

进展超出预期。DDR5 已进入量产阶段，LPDDR5 开始批量供货智能手机厂商。



NAND - 长江存储 (YMTC)

成功跨越技术门槛。232 层 3D NAND 技术已实质性追平国际主流水平。

HBM 真空地带：最危险的短板

The Gap

国内目前还没有能力量产 HBM2e 及以上级别的产品。



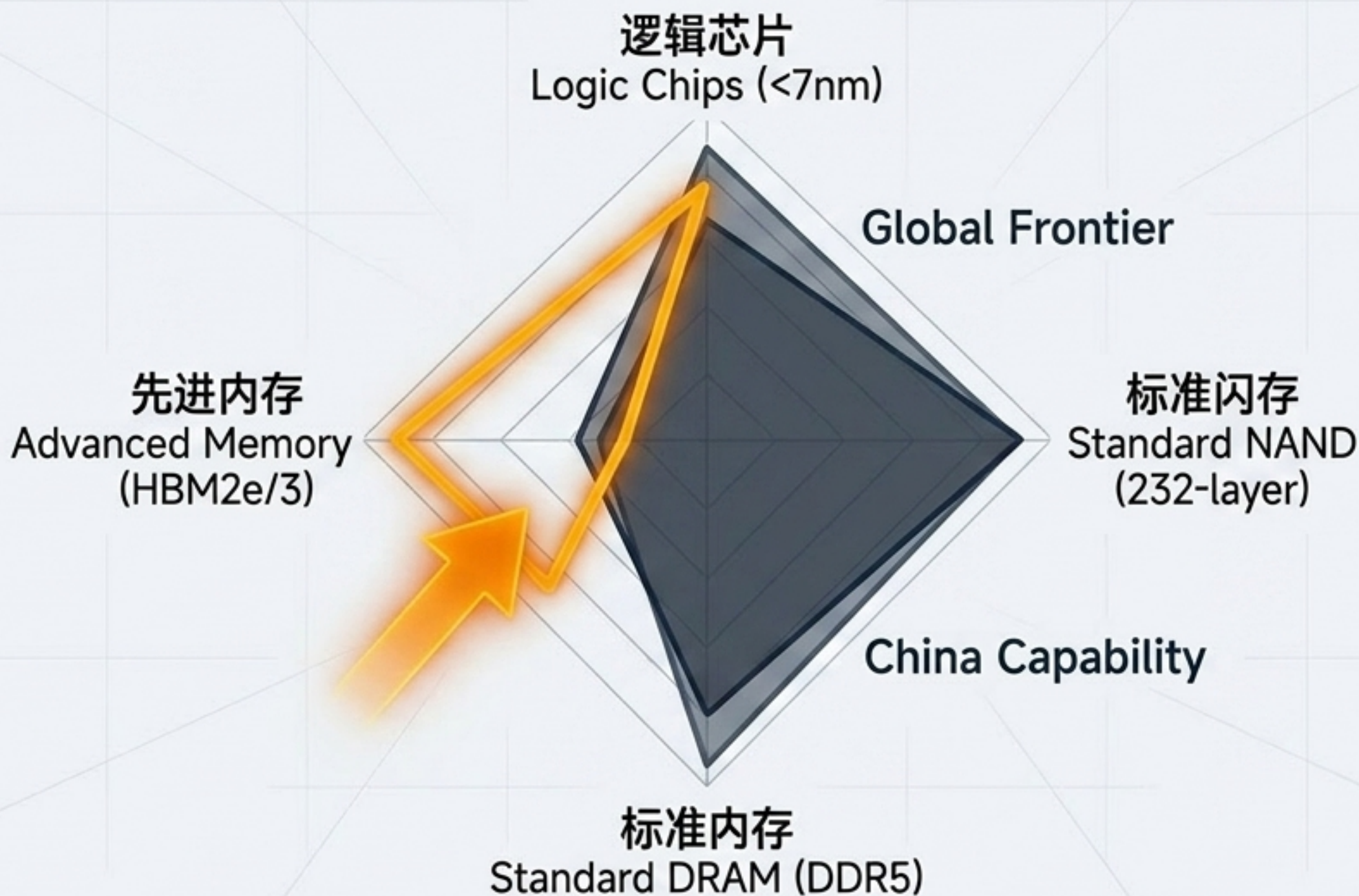
Missing Core Tech

HBM 高度依赖的三项核心技术在国内成熟度极低：

1. 高精度 TSV 封装
2. 大面积硅中介层 (Silicon Interposer)
3. 复杂的多层 3D 堆叠工艺

这是美国制裁下一步最可能收紧，也是最具杀伤力的领域。

半导体能力矩阵：制裁的下一个十字线



中国在 2D 存储阵列地取得胜利，但在 3D 堆叠的高带宽内存 (HBM) 上存在巨大断层。

综合演进：TCO（总所有成本）的终极重构



结论：内存成本已经历史性地成为 AI 算力基建 TCO 模型中占比最大的单一项。存储不再是边缘组件，而是决定 AI 时代天花板的绝对核心约束。

迎接十年存储超周期

存储超周期的结构性驱动绝非短期的“芯片短缺”，而是供需缺口在中长期内的持续性重塑。

尖端设备 (Equipment)

特别是 TSV 钻孔、
对准与检测设备。

先进封装 (Packaging)

掌握 2.5D/3D 堆
叠与中介层技术的
代工体系。

核心材料 (Materials)

高散热、高纯度封
装材料供应商。

在算力飞轮的另一面，掌控带宽的人，将掌控 AI 的下一个十年。