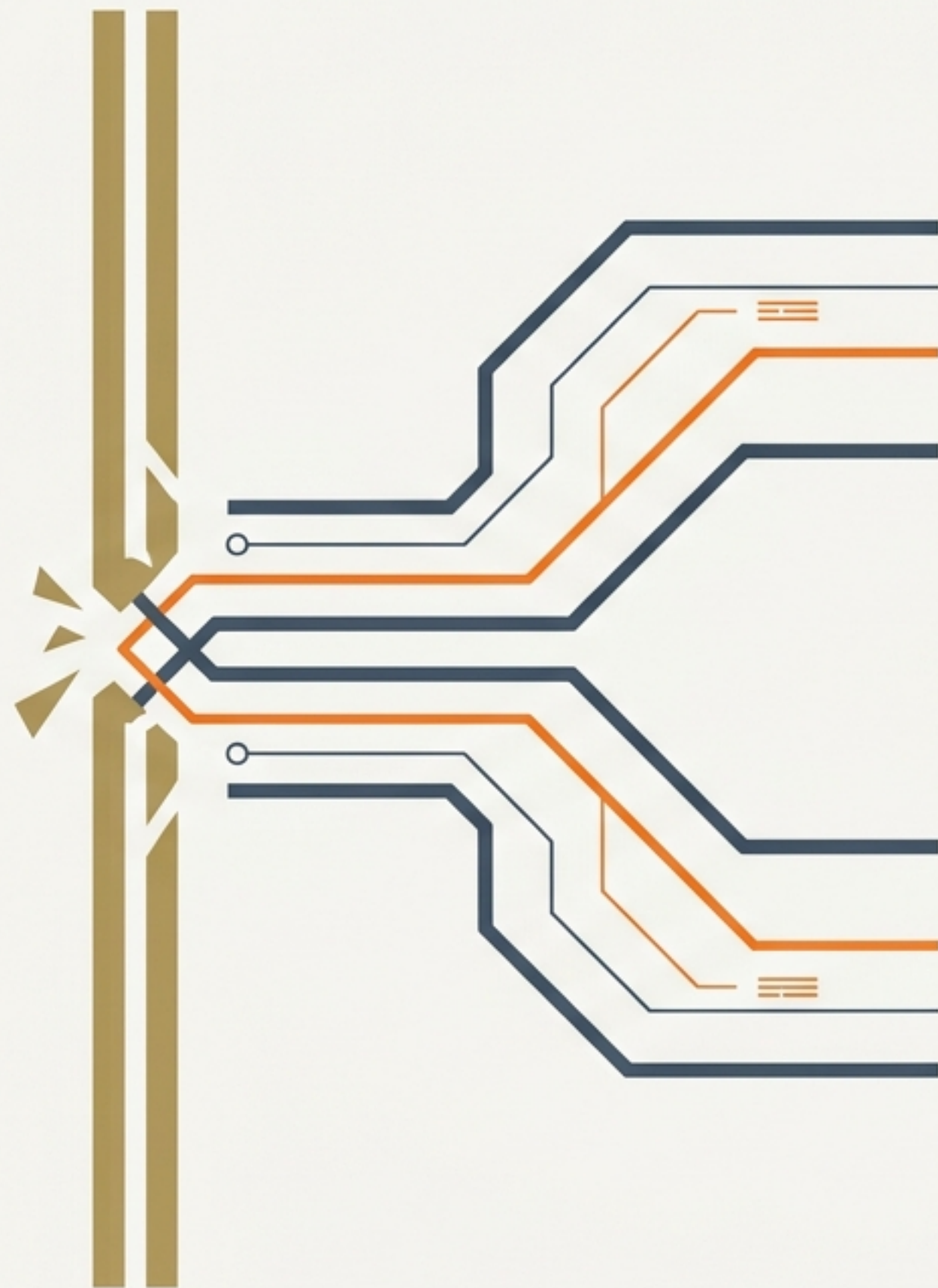
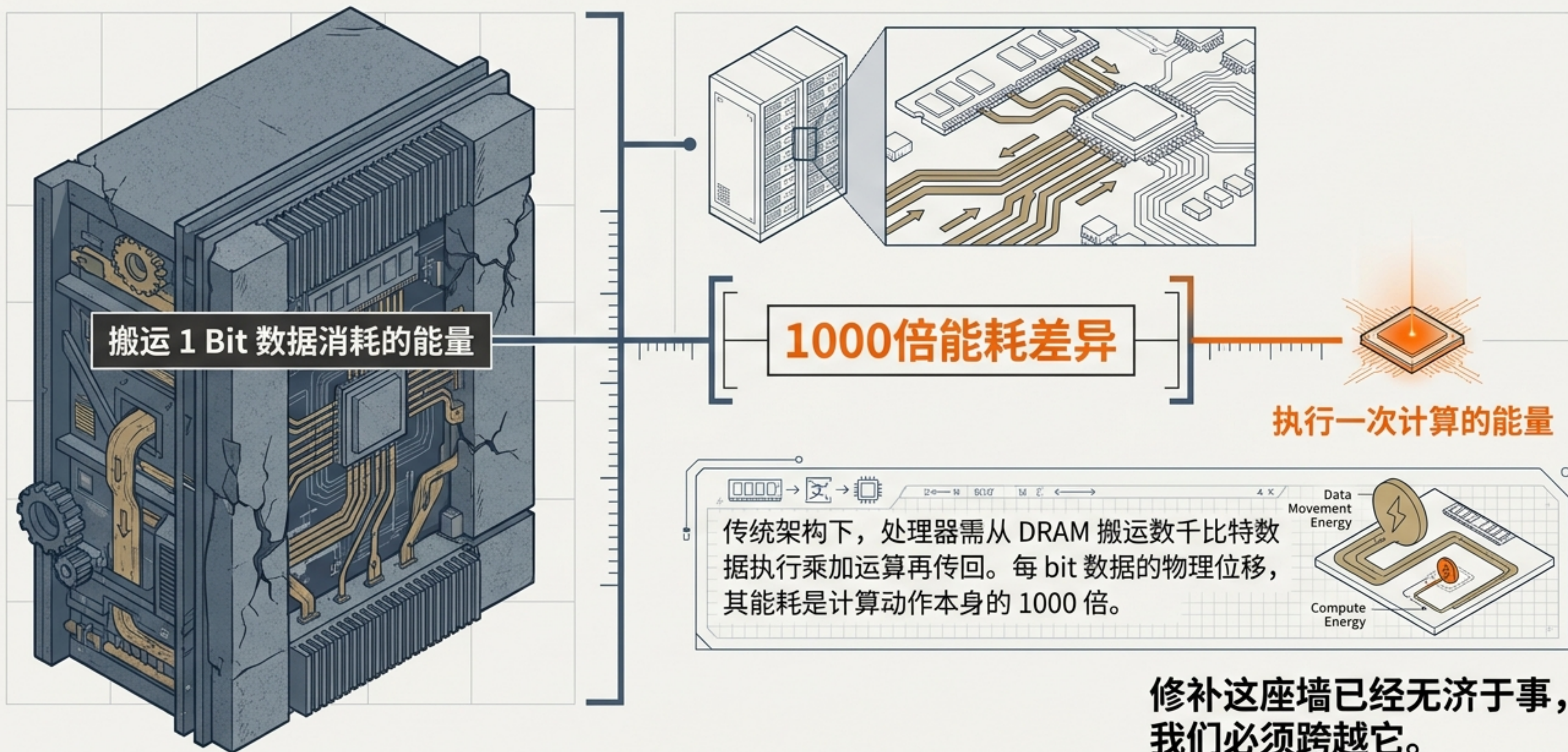


存储墙之外： 计算架构的下一站

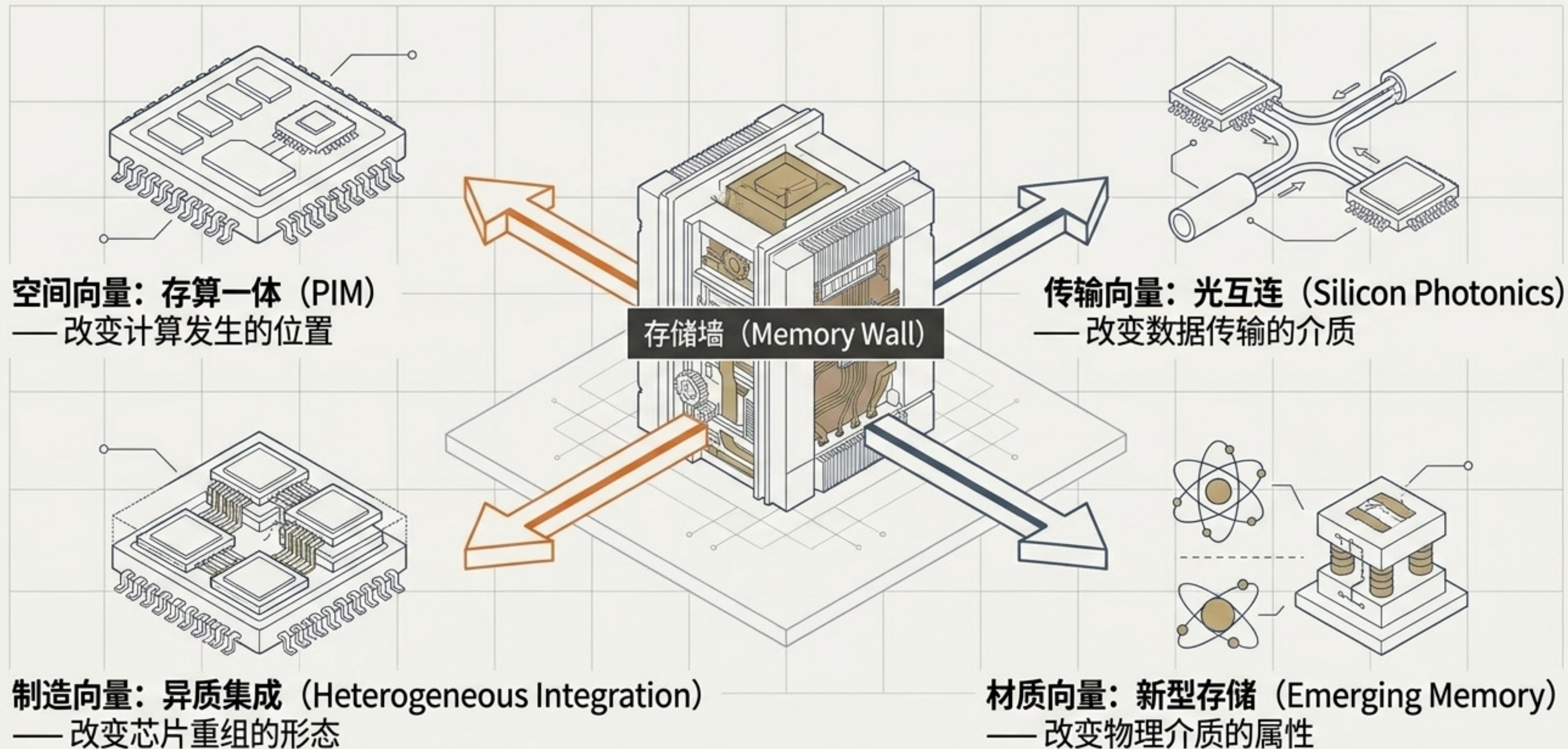
当物理堆叠达到极限，
AI 算力底座的演化全景图



数据的物理搬运正在吞噬计算的能量

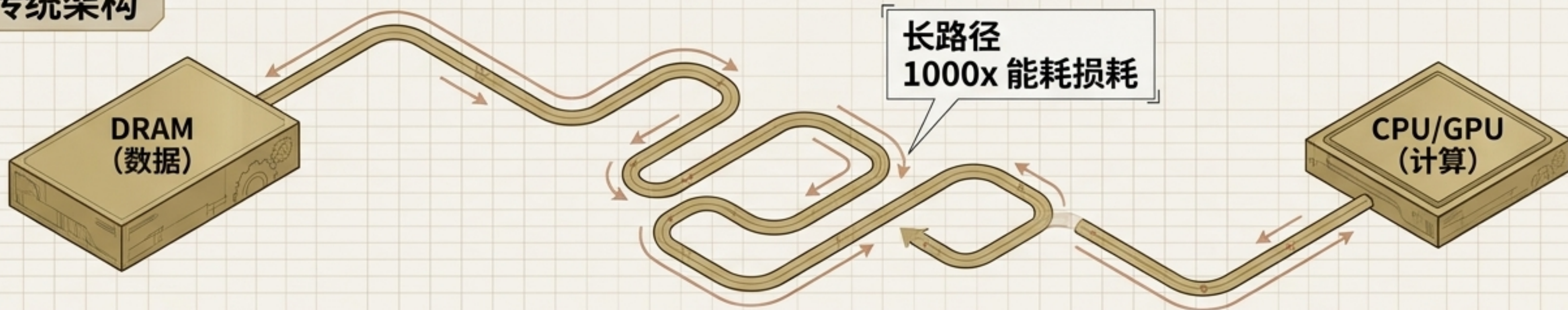


突破物理极限的四个演化向量

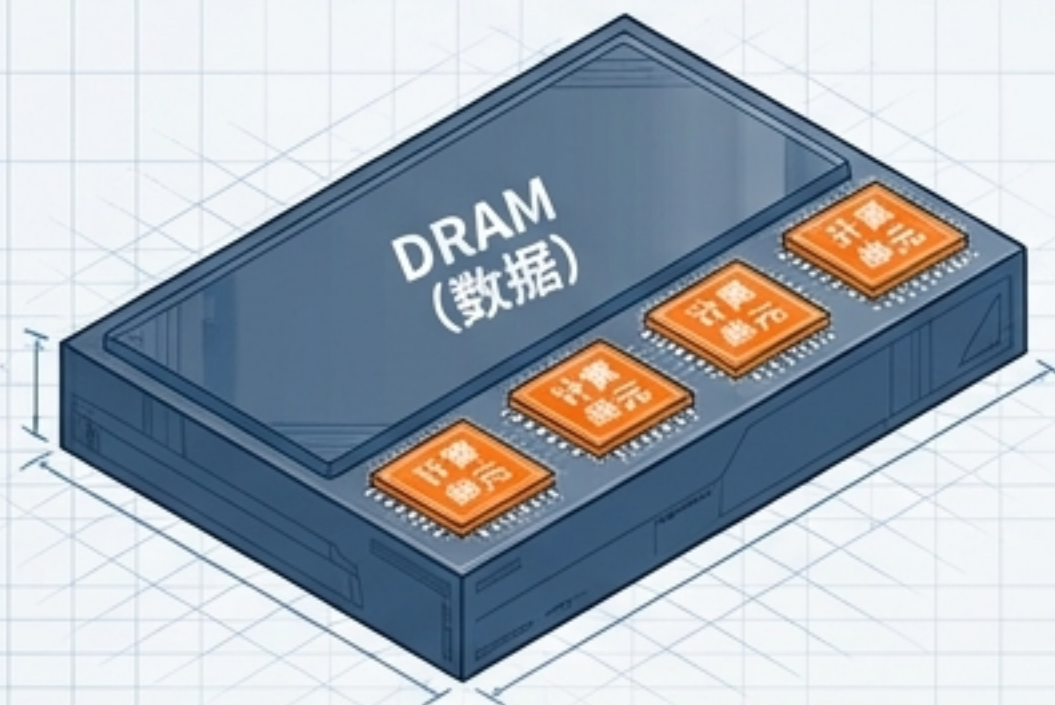


空间向量：让计算单元直接进驻数据腹地

传统架构



存算一体架构



数据零搬运（就地计算）

与其把数据搬到计算单元，不如把计算搬到数据旁边。

存算一体在 AI 推理中的现实投影与边界

2.5x

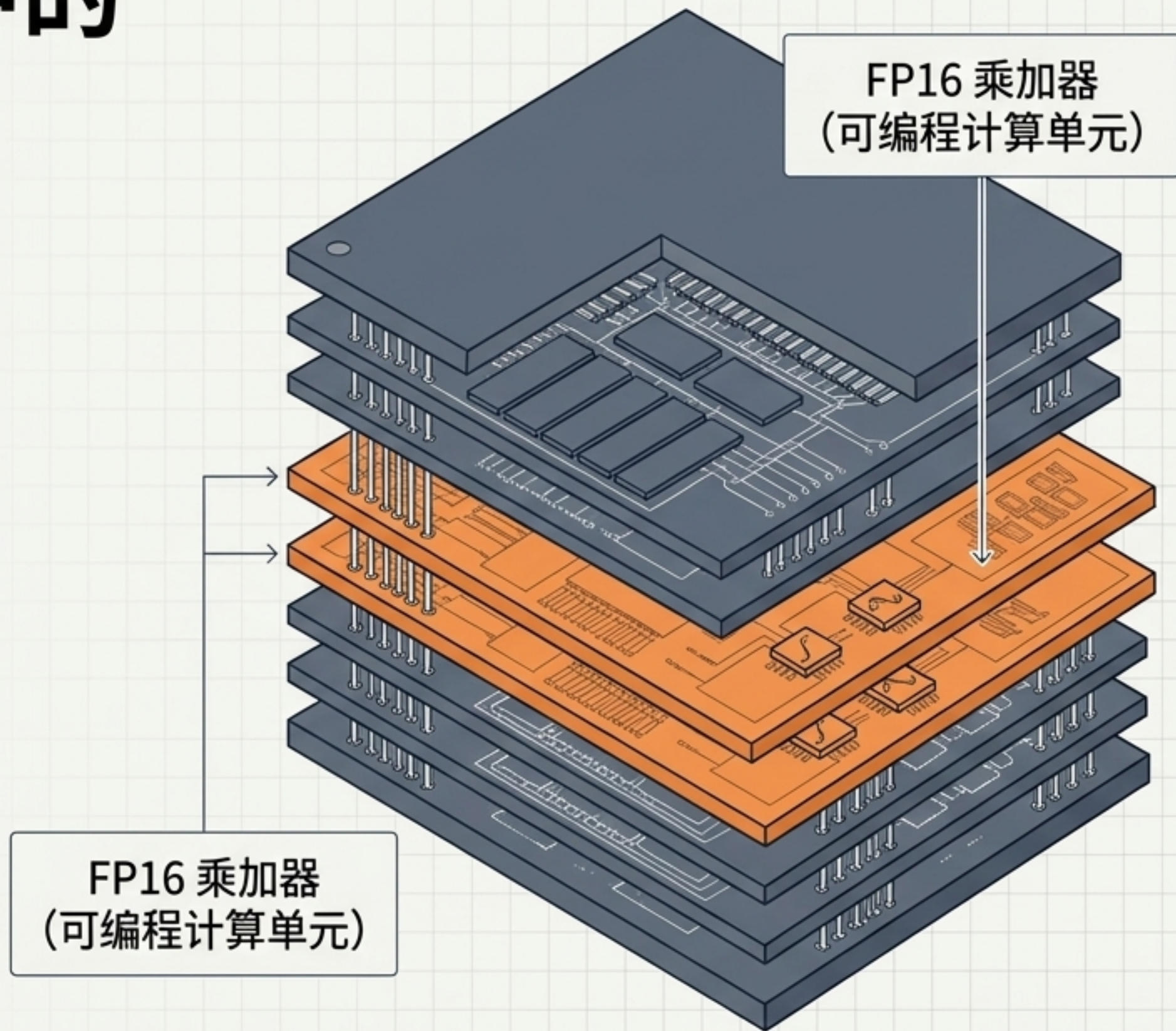
实测能效提升
(以三星 HBM-PIM 为例)

核心挑战

DRAM 工艺精度远低于逻辑工艺，片上计算的精度和速度面临妥协。

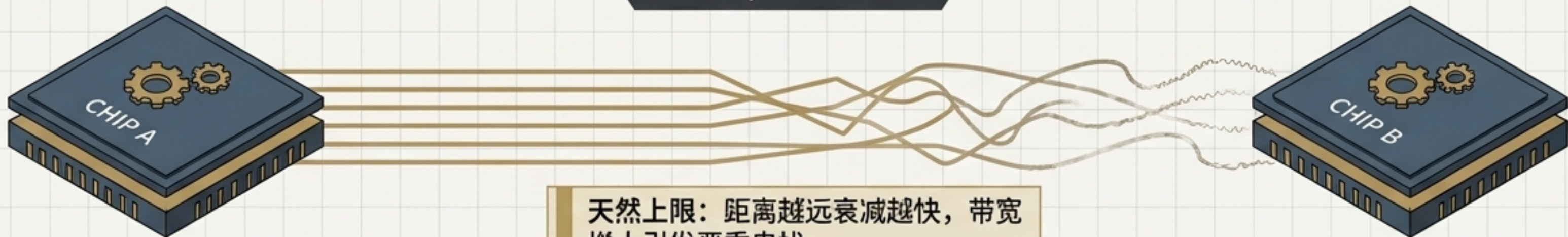
最佳落地场景

存算一体并非替代 GPU 训练，而是完美契合大数据量大数据量、简单计算的 AI 推理场景。



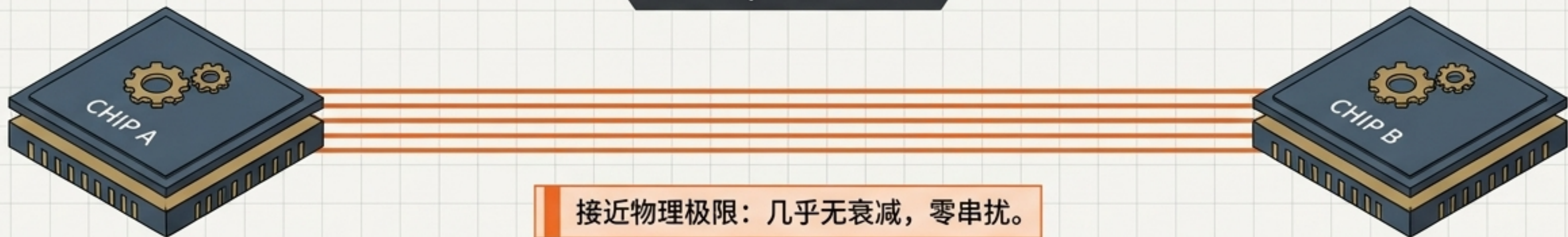
传输向量：用光子替代电子跨越物理衰减

电子 / 铜线互连



天然上限：距离越远衰减越快，带宽增大引发严重串扰。

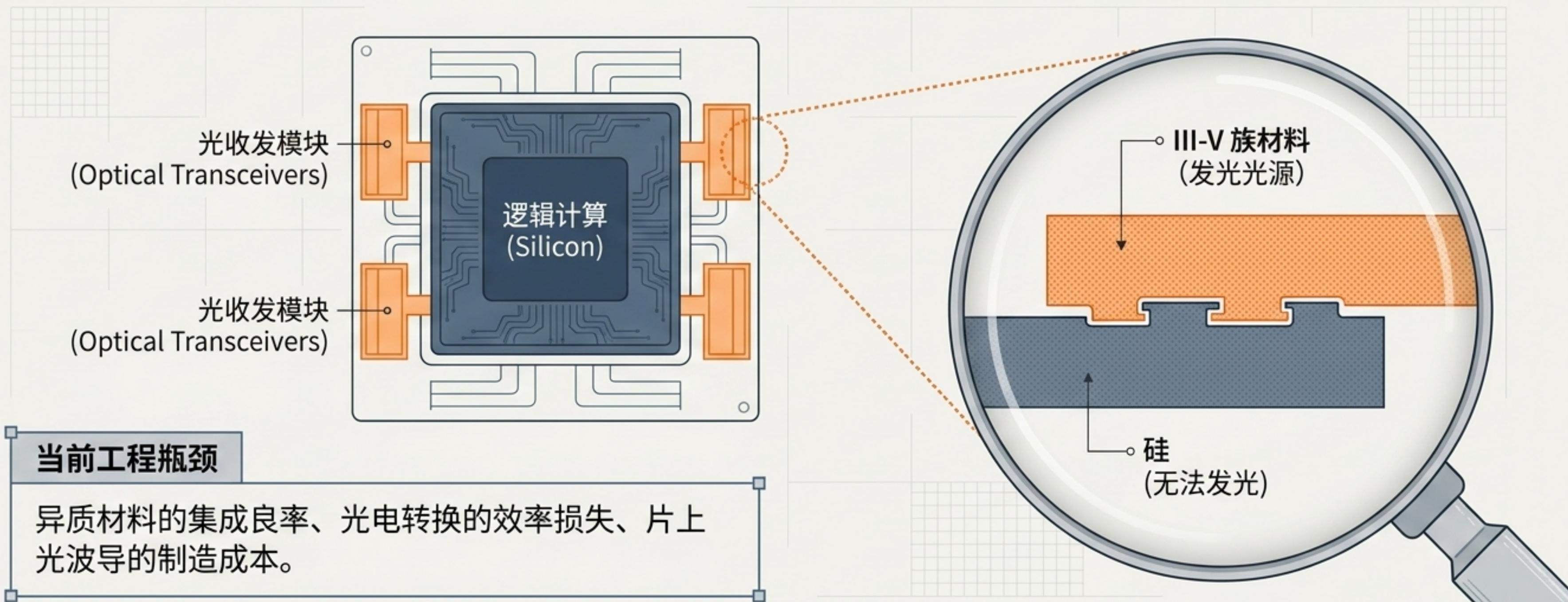
光子 / 波导互连



接近物理极限：几乎无衰减，零串扰。

功耗可降低 5-10 倍 | 带宽提升 10倍以上

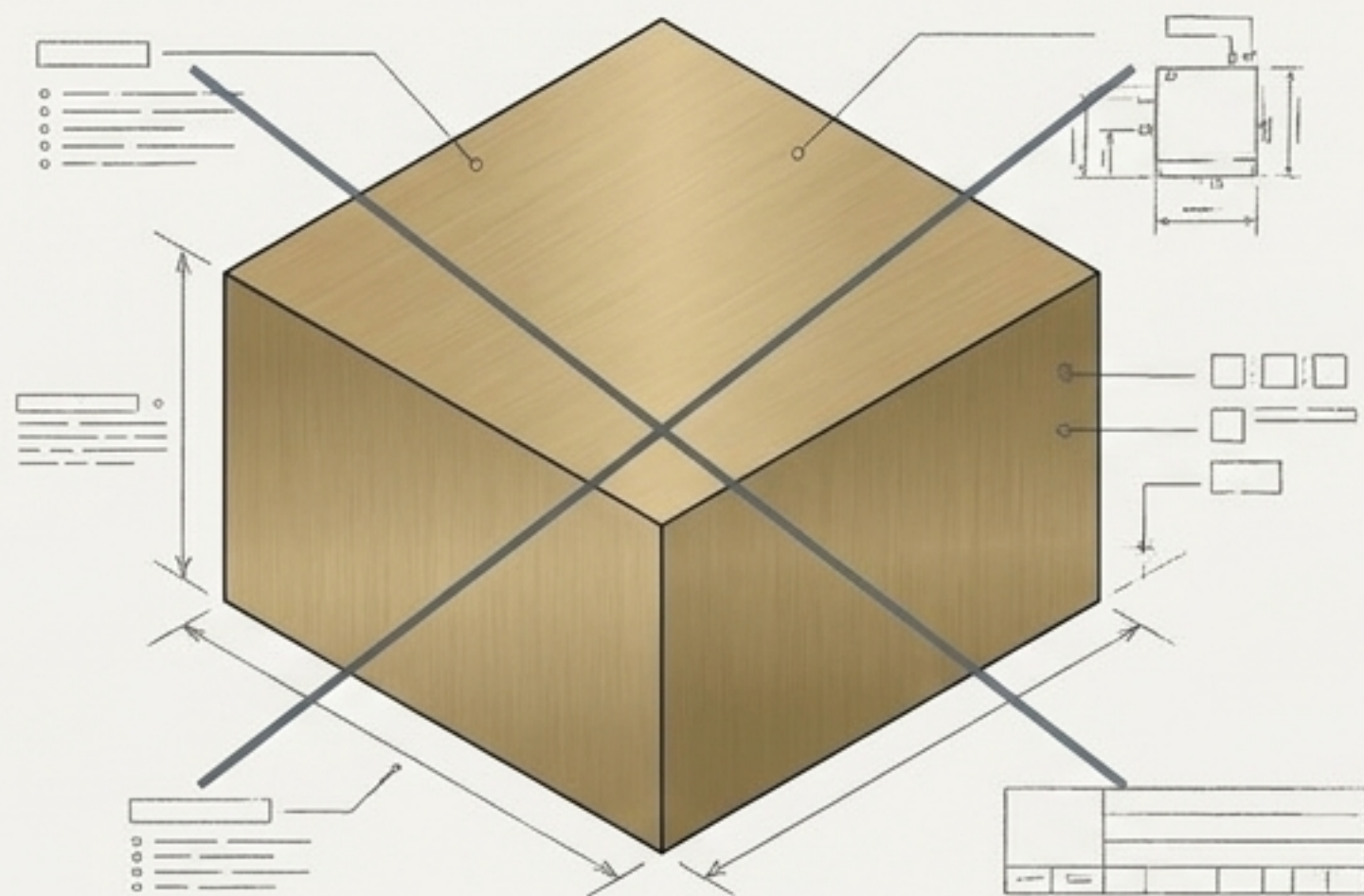
光电分工重塑芯片间的大规模通信逻辑



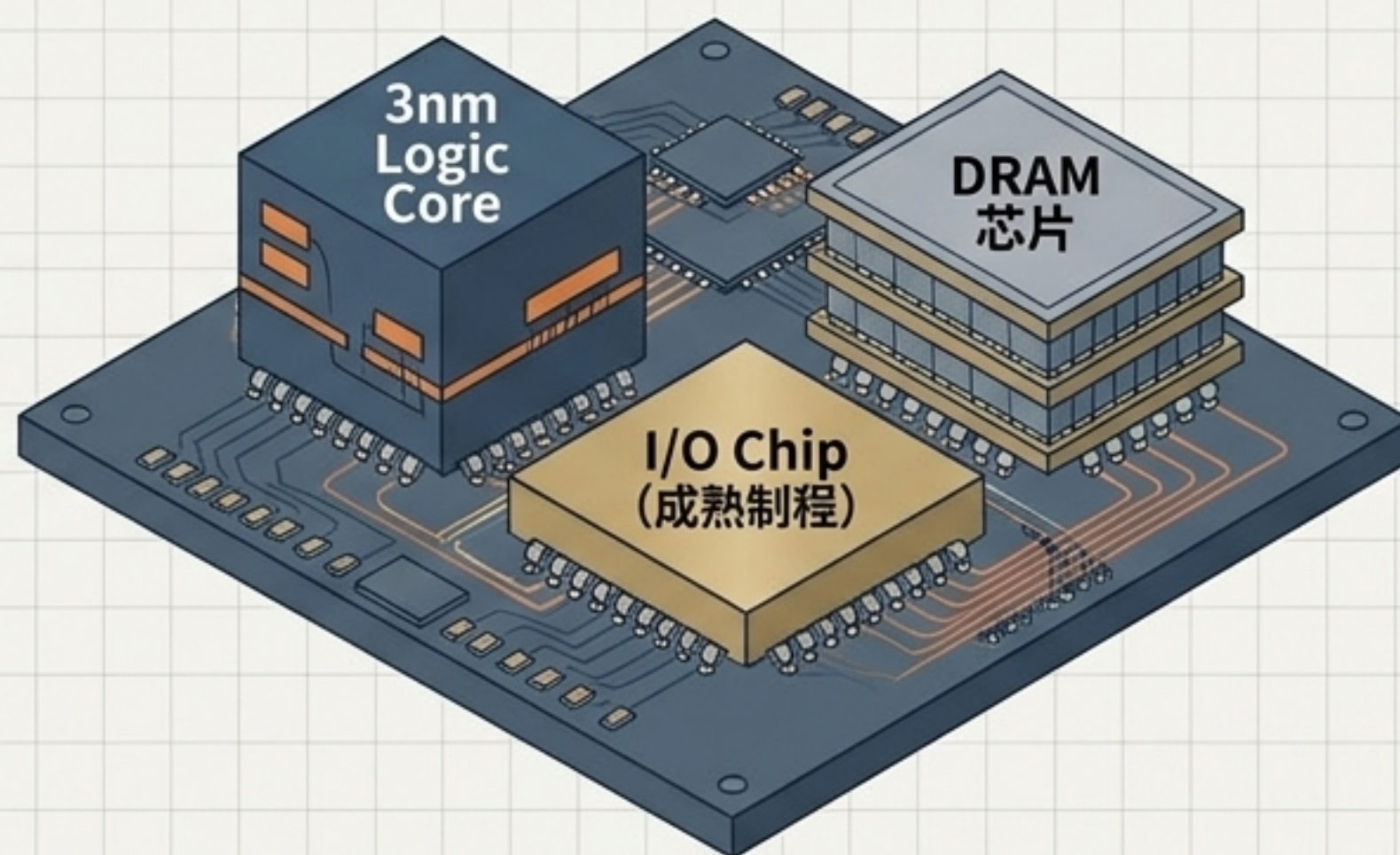
计算的归电子，传输的归光子。各自在最擅长的物理定律下工作。

制造向量：将庞大的单体芯片拆解重组

传统的单片 SoC (妥协的制程)



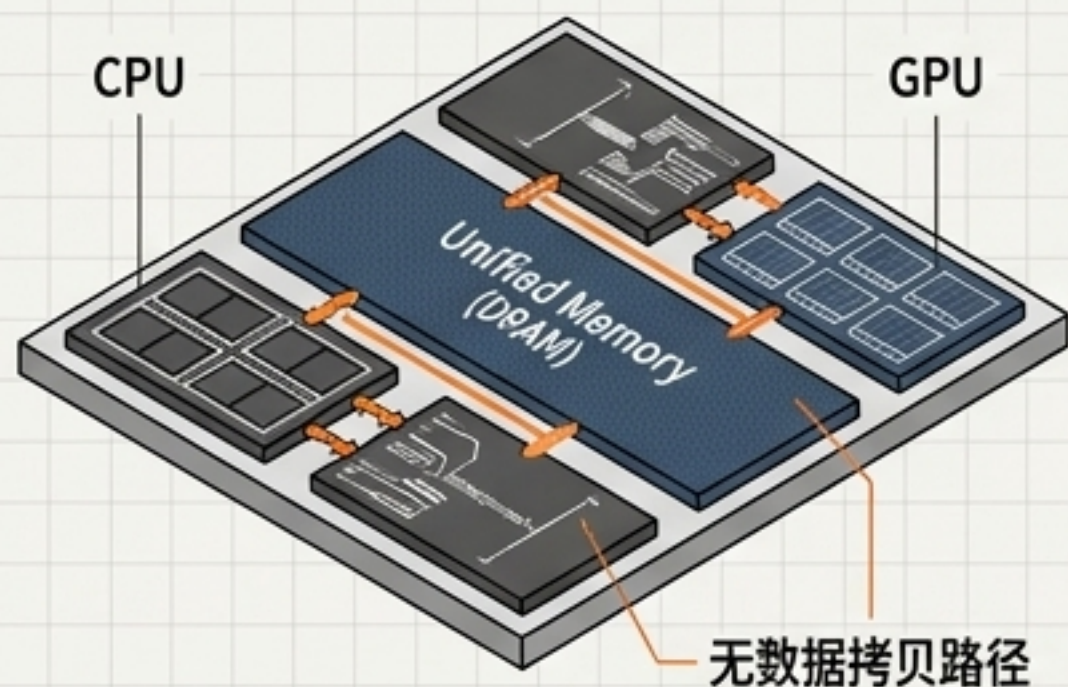
硅中介层 / 高级封装



放弃让所有组件在同一工艺下妥协。让逻辑、I/O、存储各自采用最优制程，再通过 2.5D/3D 封装在基板上重新连结。

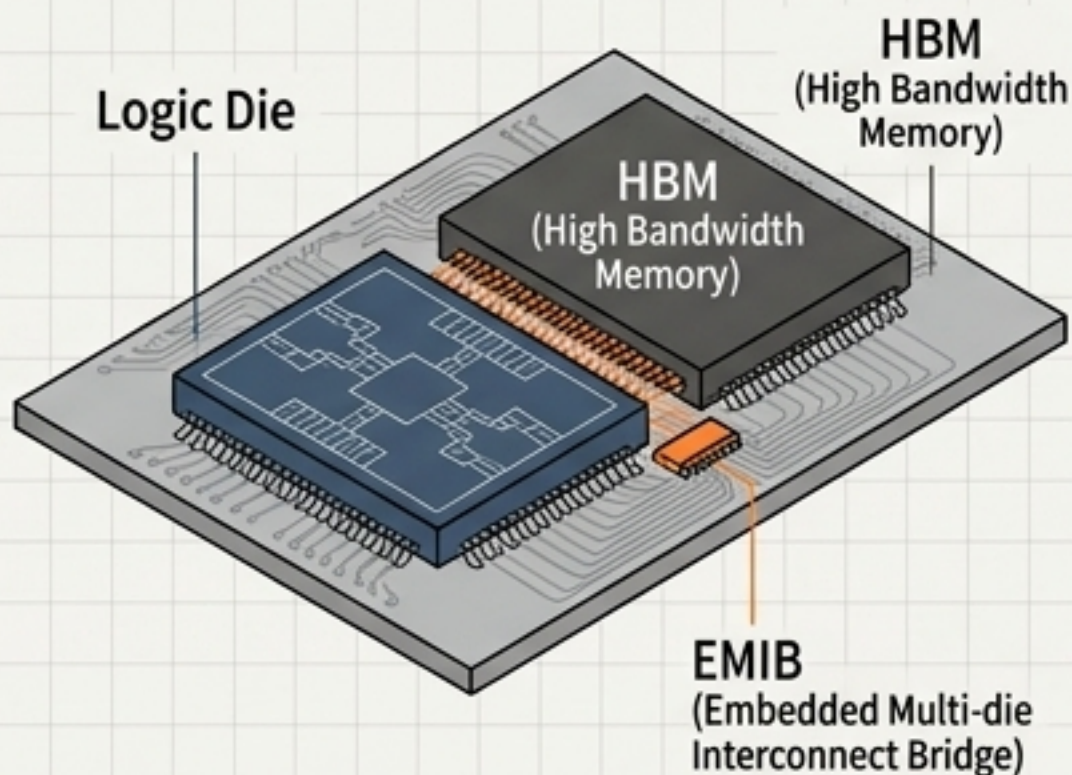
消除数据搬运的封装级统一架构

Apple M系列



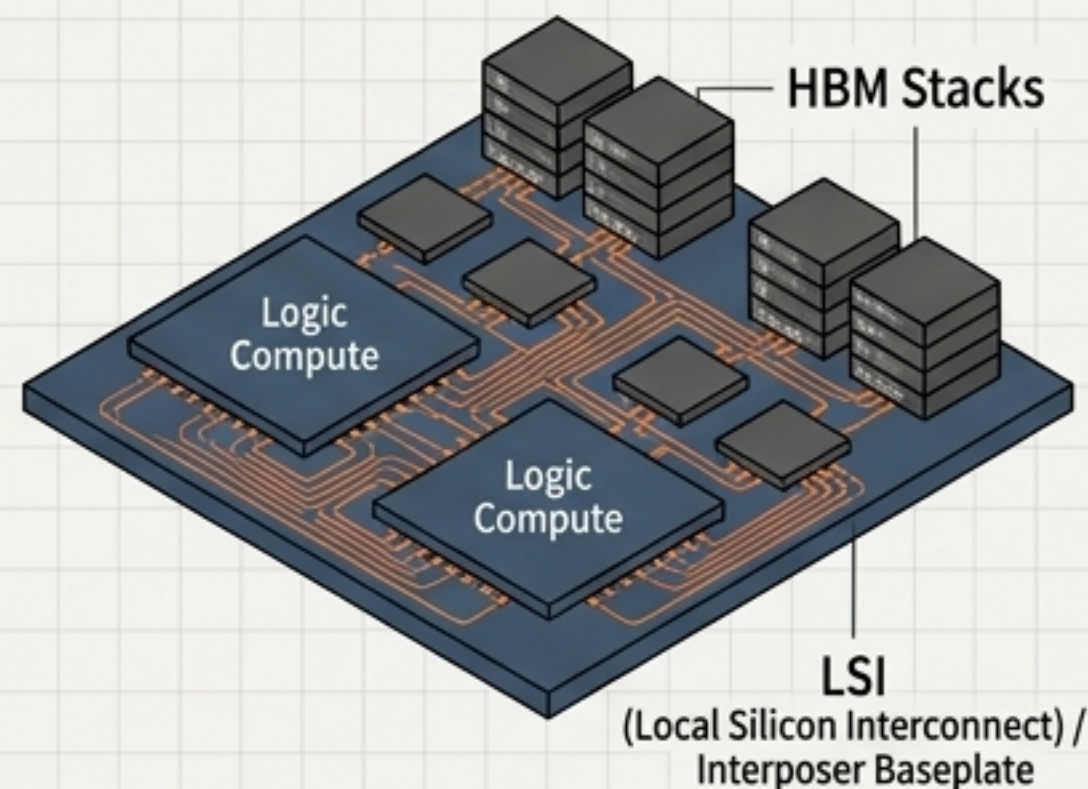
统一内存架构，消除内部数据拷贝

Intel EMIB



超高密度硅桥连结异构小芯片

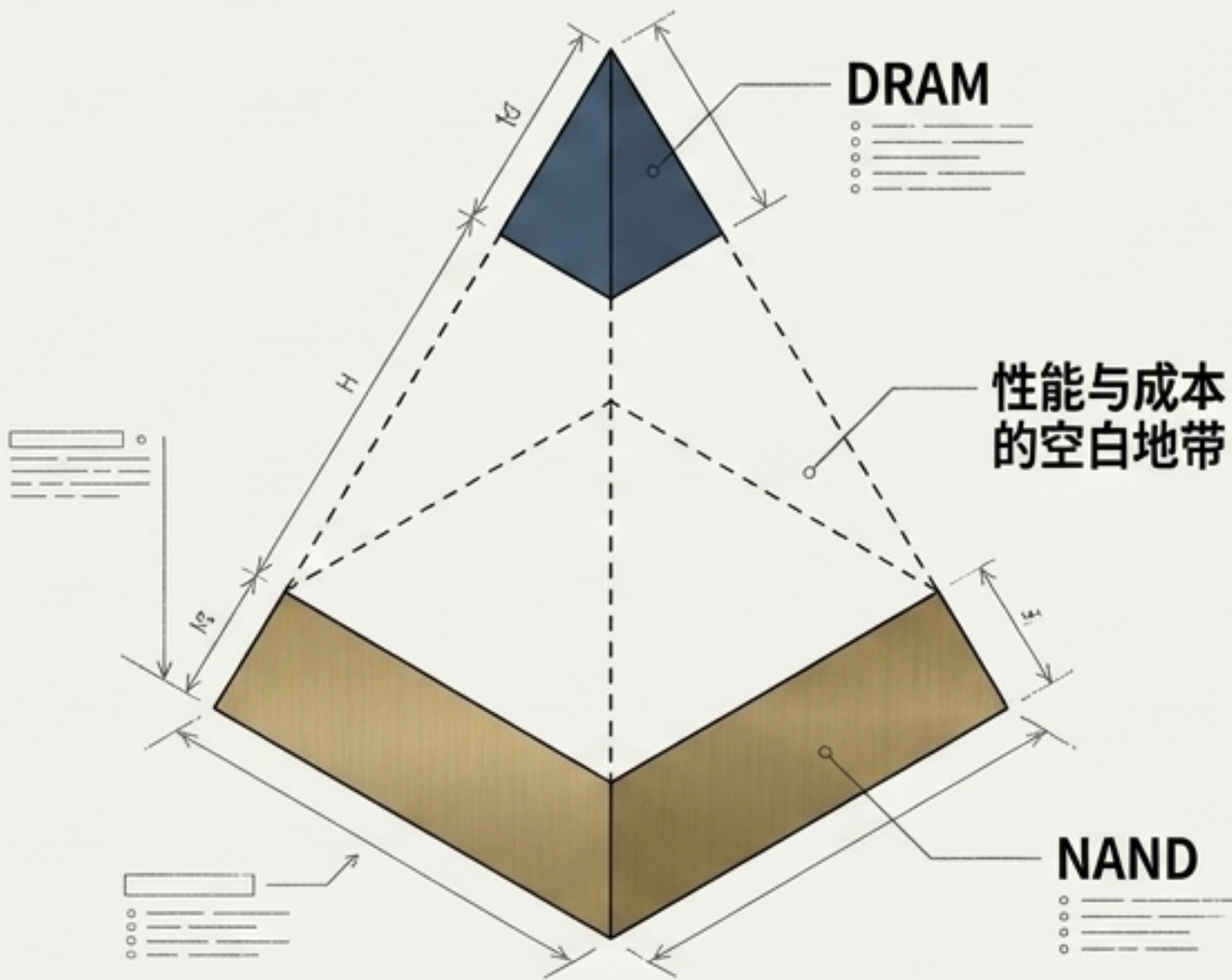
台积电 CoWoS-L



超大面积中介层实现多芯片扩展

核心诉求完全一致：让异质集成变得更便宜、面积更大、能塞下更多芯片。

材质向量：寻找性能与成本之间的理想介质



Diagnostic Matrix			
介质	速度	成本	易失性
DRAM	极快	高	断电丢失
新型存储 (MRAM/ReRAM/ FeRAM)	填补断层	填补断层	非易失性
NAND	极慢	低	永久保存

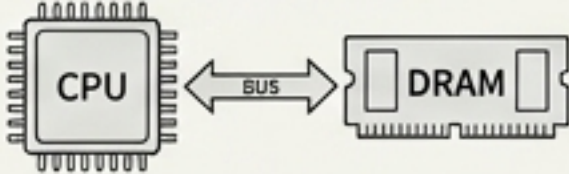
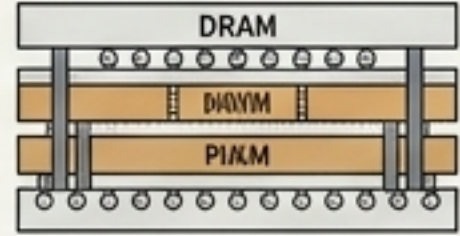
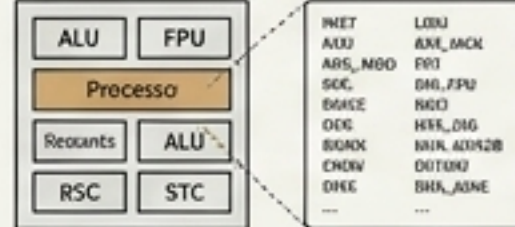
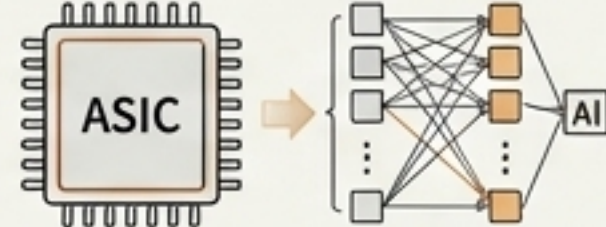
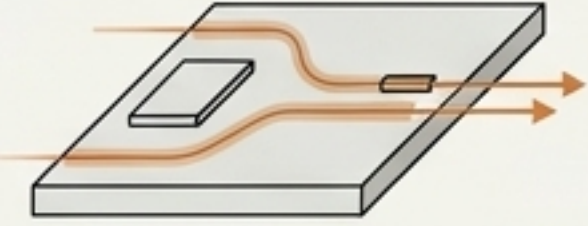
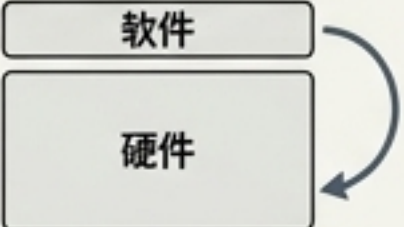
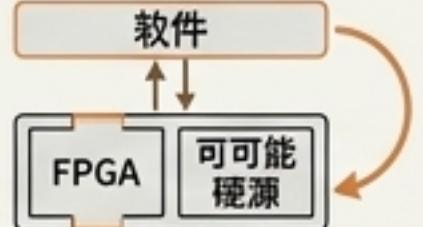
AI 时代急需一种“持久化的大容量工作内存”，让模型权重可以直接驻留，彻底免去从 SSD 缓慢加载到 DRAM 的过程。

填补空白的候选者与三十年的终极圣杯



行业现实：寻找比 DRAM 快、比 NAND 便宜、非易失、无限耐久的“Universal Memory”已经持续了 30 年。至今，没有任何一种新型介质在成本、可靠性和存储密度上同时跨过量产门槛。

架构演化的大趋势：一场底层的范式转移

	传统范式 (Past Paradigm)		未来范式 (Future Paradigm)
空间距离	通用 CPU 与内存物理分割 	➡	数据相邻计算 (PIM) 与 3D 垂直微米级堆叠 
专用化程度	x86 全套通用指令集 	➡	特定算子固化 ASIC (如极度优化的矩阵乘法) 
互连介质	受制于铜线的电信号衰减 	➡	无拘无束的硅光子互连 
软硬关系	硬件定义软件, 迭代周期数年 	➡	软件定义硬件, 迭代周期数月 

存算一体、光互连、异质集成与新型存储并非孤立创新，它们共同指向一个离数据更近、更三维、高度专用的未来架构。

曾经最枯燥的组件，成为了 AI 时代的决战之地

二十年前，内存仅仅是一个插上能用、容量够大就行的无聊配件。

而今天，跨越存储墙的每一步索，都在定义下一个十年人工智能的计算计算边界。而在于大转入之地。

