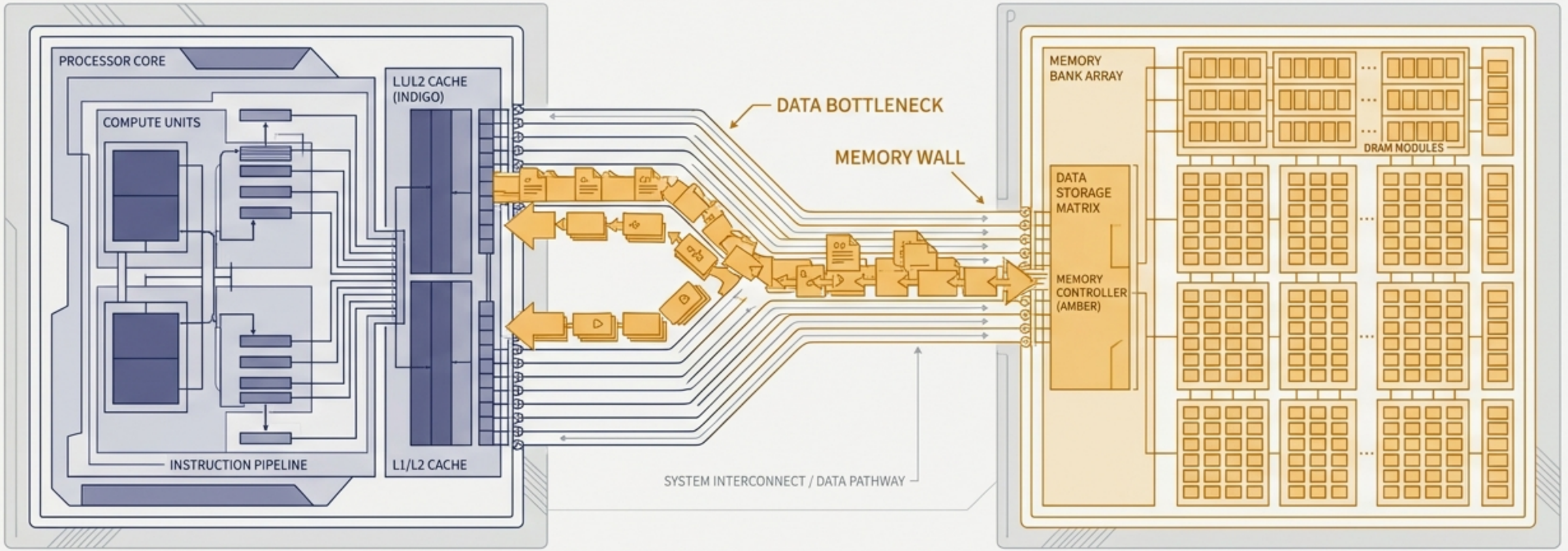
解释性诊断 // 存储墙

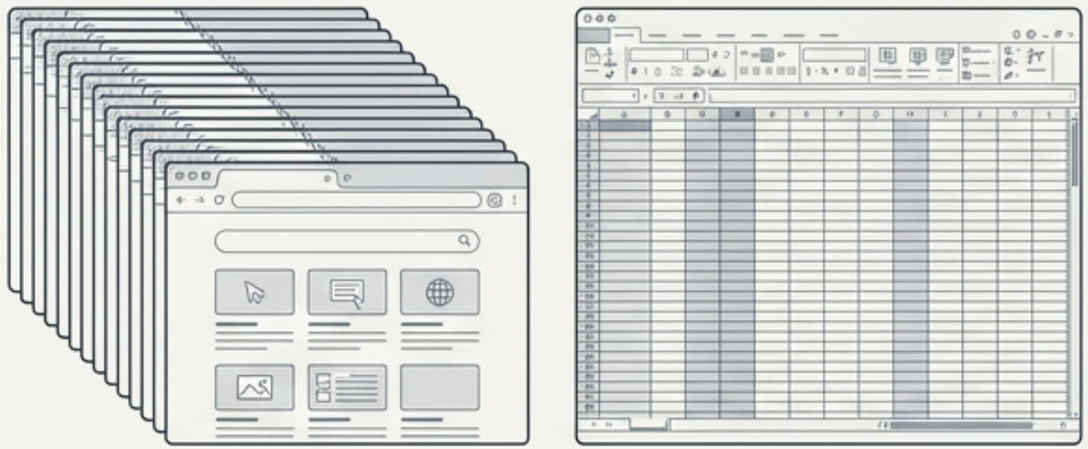
电脑卡，十有八九是处理器在等内存

算力呈指数级爆发，但数据传输的物理法则正锁死计算架构的未来。

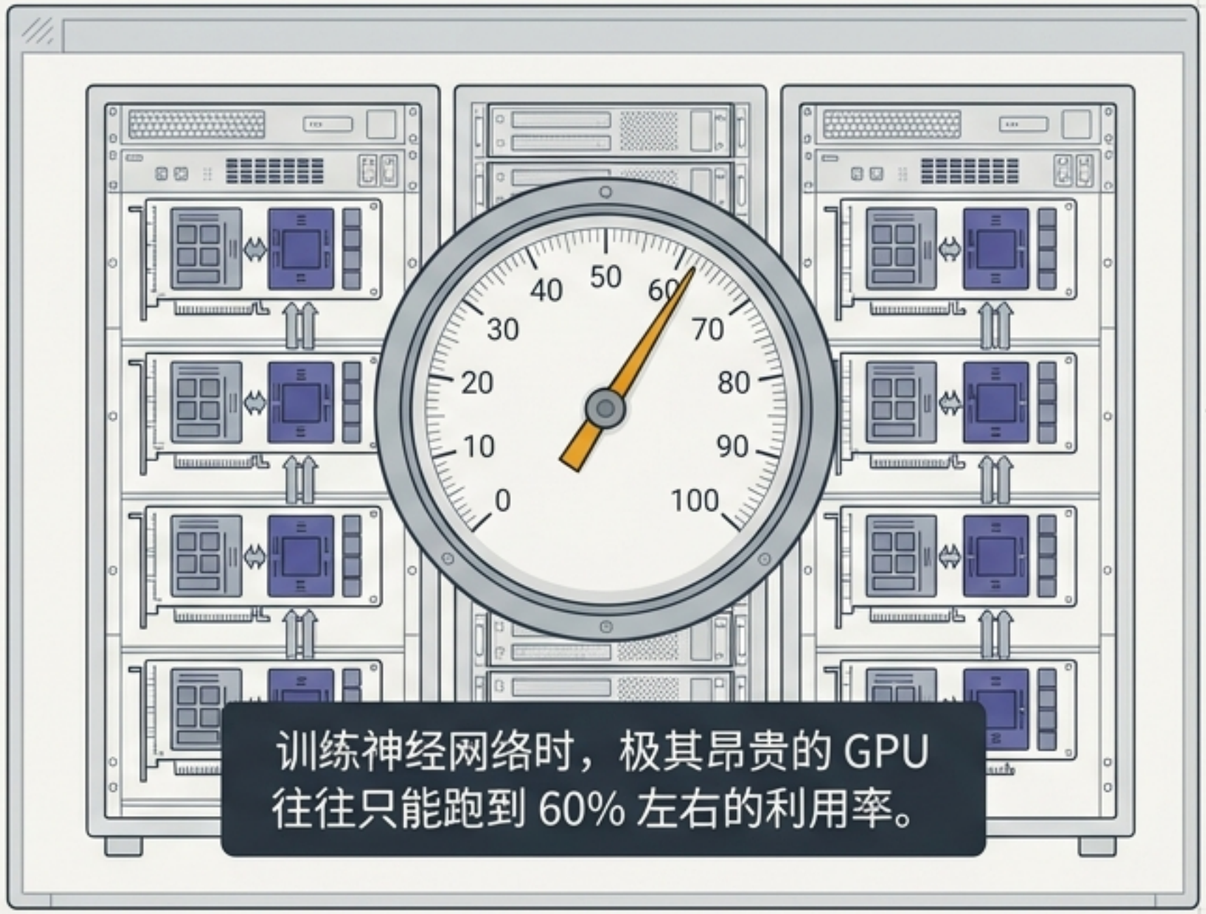


GPU 利用率停滞在 60% 的物理谜团

白屏半秒



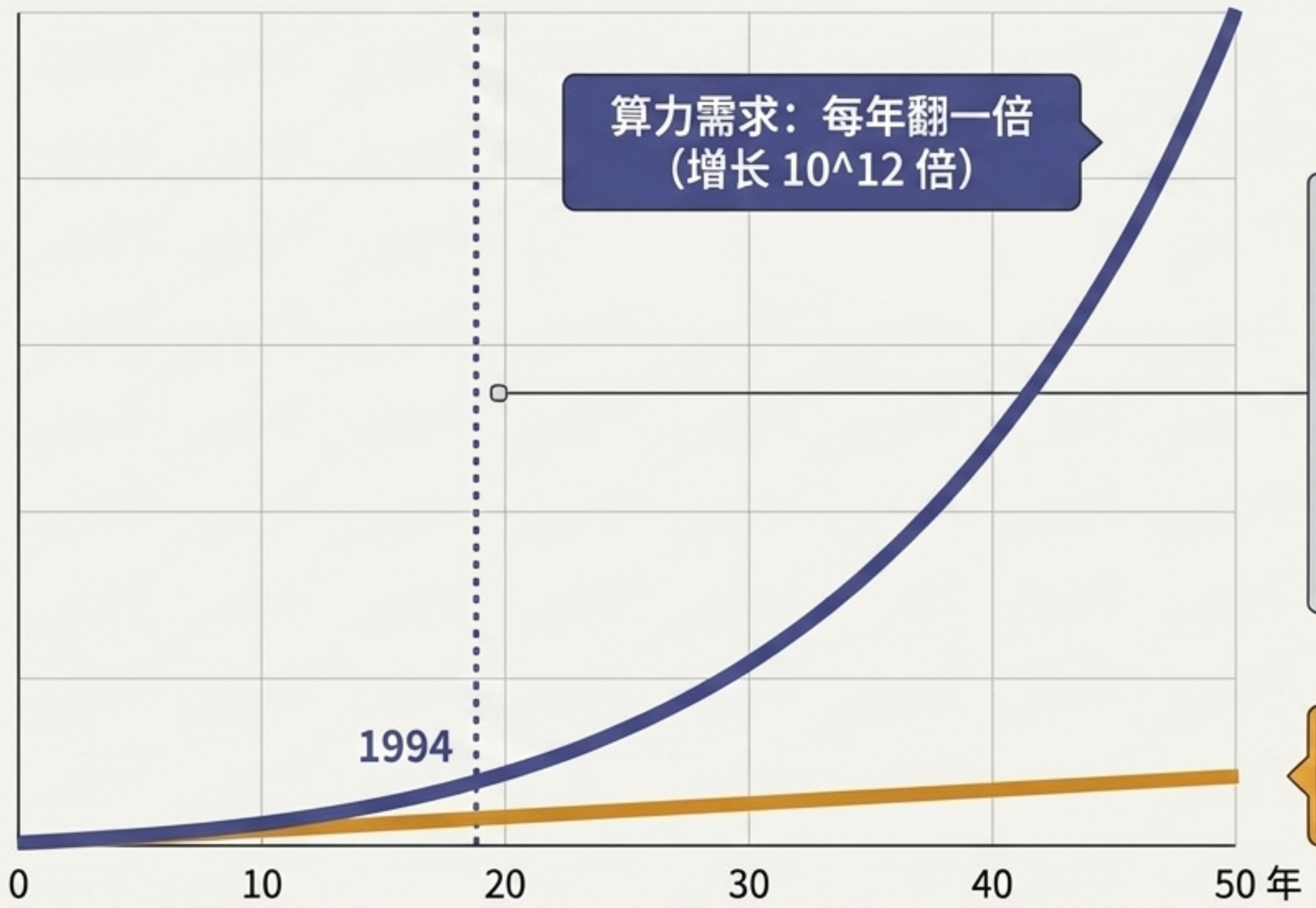
无论是 Chrome 开三十个标签页时的卡顿，还是加载庞大 Excel 时的僵死，并非处理器算力不足。



训练神经网络时，极其昂贵的 GPU 往往只能跑到 60% 左右的利用率。

钱花了，卡是好卡。但在剩下的 40% 时间里，处理器都在干等——等待数据从内存长途跋涉而来。

万亿倍与 4 倍的五十年鸿沟



1994年的致命预言

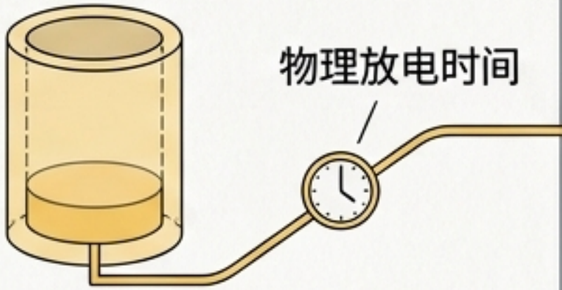
William Wulf 和 Sally McKee 发表论文《Hitting the Memory Wall》。他们预言：按此趋势，处理器速度将完全被内存带宽锁死，加再多核心也只是在“等饭吃”。三十年过去，预言已成现实。

DRAM 的微观病理：被物理定律定死的放电

1. 半个世纪未变的结构

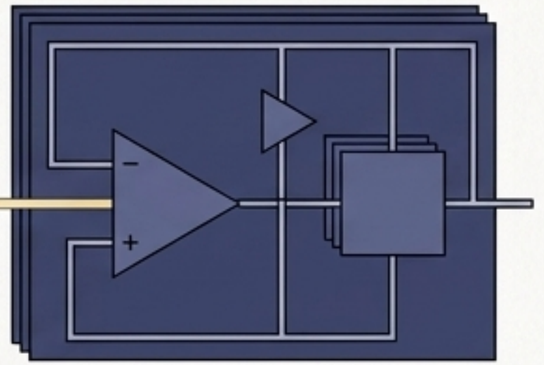
自 1970年代起，DRAM 的核心就是电容。存 1 或 0，必须不断刷新以防漏电。

2. 等待物理放电



读取数据时，必须等待电容放电电。这是一个纯物理过程，无法通过软件优化略过。

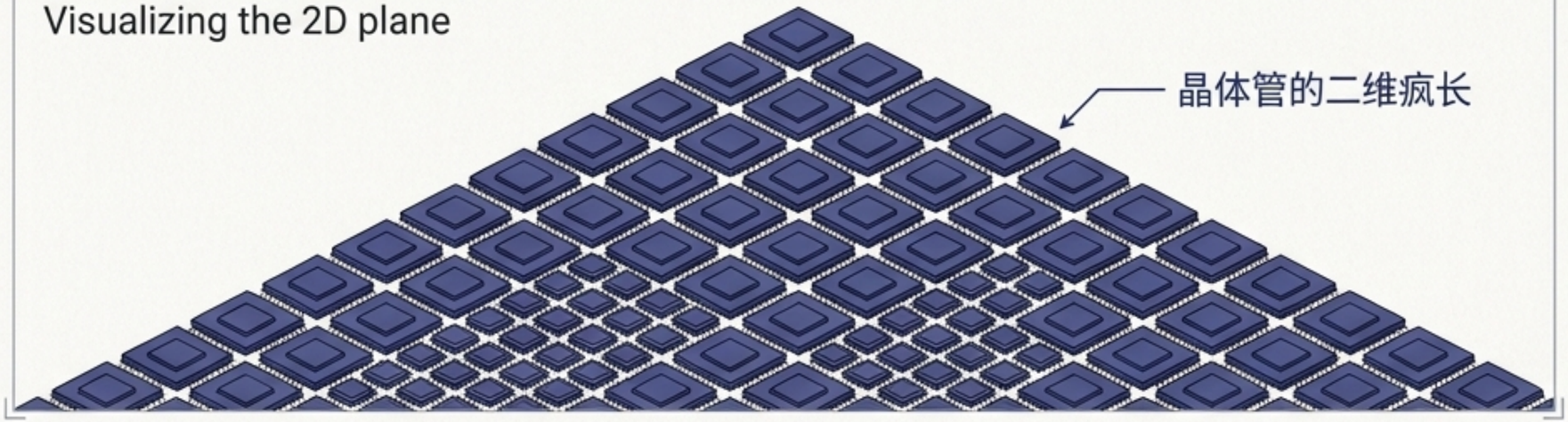
3. 优化空间的尽头



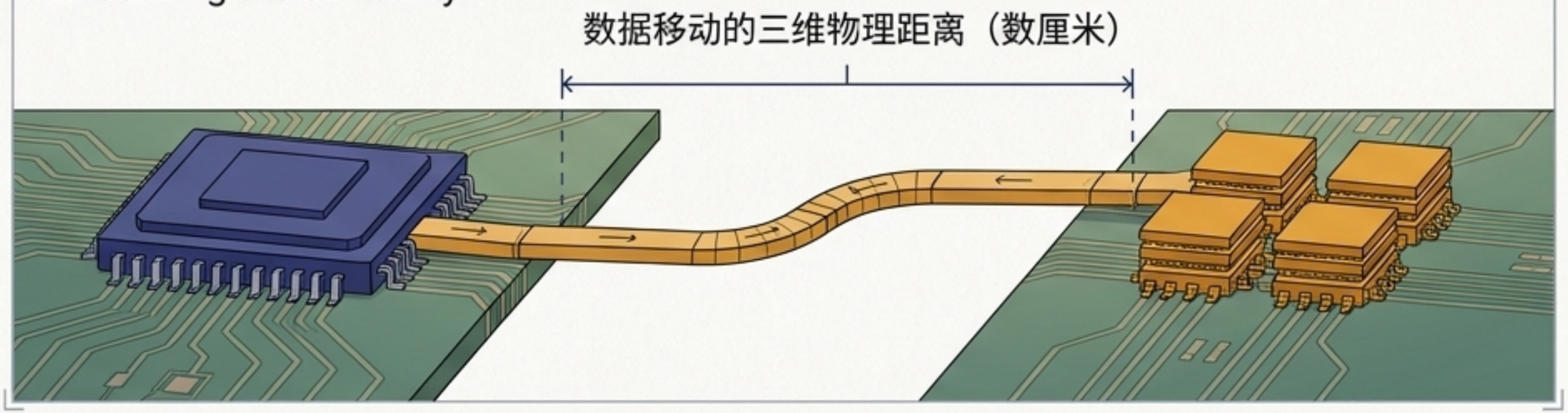
这个放电速度是物理定律定死的。你可以把马路修宽（增加带宽），但每辆车的车速（延迟）无法再提。换任何架构都快不起来。

空间悖论：二维的计算，三维的搬运

Visualizing the 2D plane



Visualizing the 3D reality



计算本身花不了多少能量。真正的能耗和时间大头，全砸在了把数据从内存搬到处理器的这一路上。

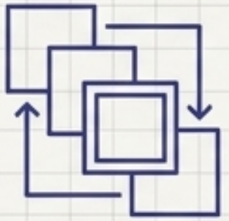
- ☑ 芯片上的晶体管在二维平面内指数级翻倍。
- ☑ 数据交换却要在不同芯片之间、三维物理空间中长途跋涉。

治标不治本：为什么传统的绕墙策略失效了？

CPU 策略：预测与缓存（L1/L2/L3）

将数据提前从慢速 DRAM 搬到快速 SRAM。
前提：程序访问数据有规律，能被预测。

GPU 策略：隐藏与切换（海量并发）



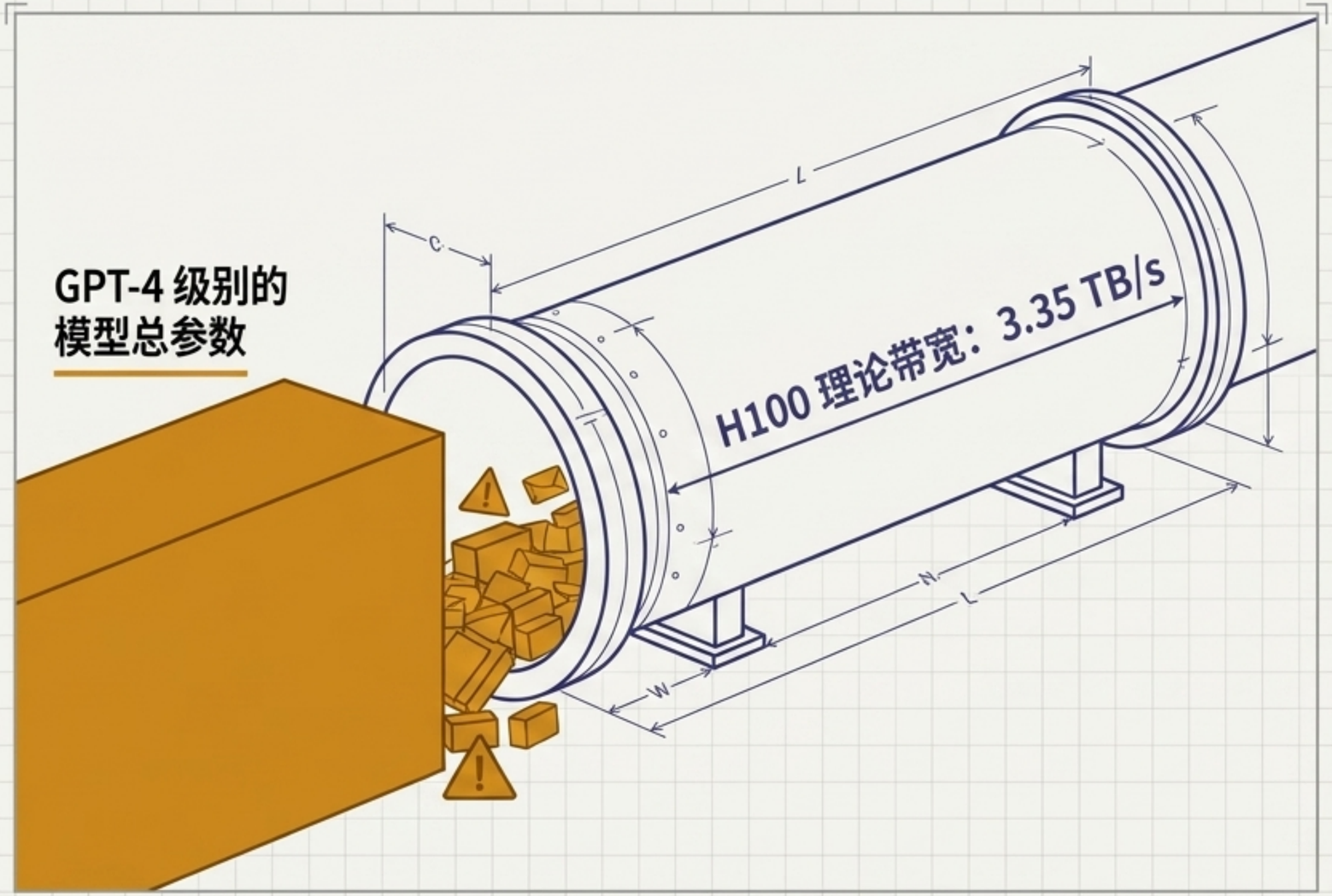
不猜了。一组线程等数据时，瞬瞬间切换到另一组运算以掩盖延迟。前提：手里有足够多可切换的线程。

AI 时代的降维打击（失效原因）

击穿 CPU：深度学习的注意力机制访问稀疏且随机，权重分散在内存各处，缓存预取算法完全失效。

击穿 GPU：实时推理场景下（Batch Size 小至 1），根本没有足够的线程可供切换来掩盖延迟。

墙的解剖学 I：带宽（路不够宽）



表面数据

H100 显存带宽高达 3.35 TB/s。听起来像极宽的高速公路。

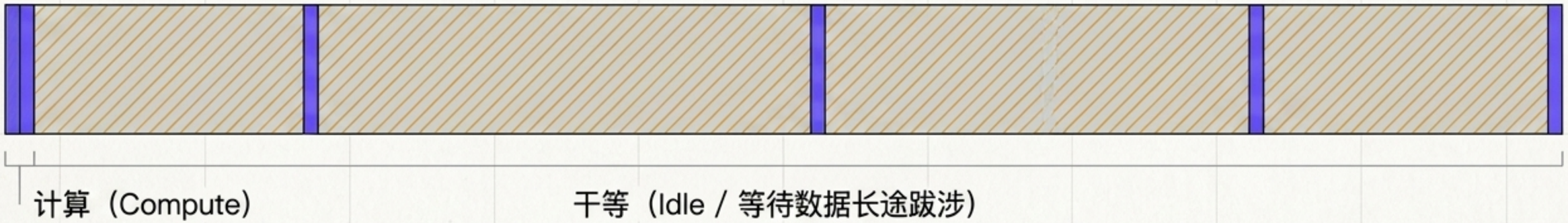
LLM 吞吐困境

对于 GPT-4 级别的推理，生成每一个 Token，都必须将模型的所有参数全部访问一遍。

残酷的除法

带宽总量 ÷ 模型参数量 = 每秒极限 Token 数。模型越大，每秒能处理。模型越大，每秒能处理的 Token 反而越少，实际吞吐完全由带宽说了算。

墙的解剖学 II：延迟（不可逾越的限速）



隐形的杀手

即使拥有无限的带宽（路足够宽），从发出数据请求到数据真正抵达，中间的物理等待时间依然无法省略。

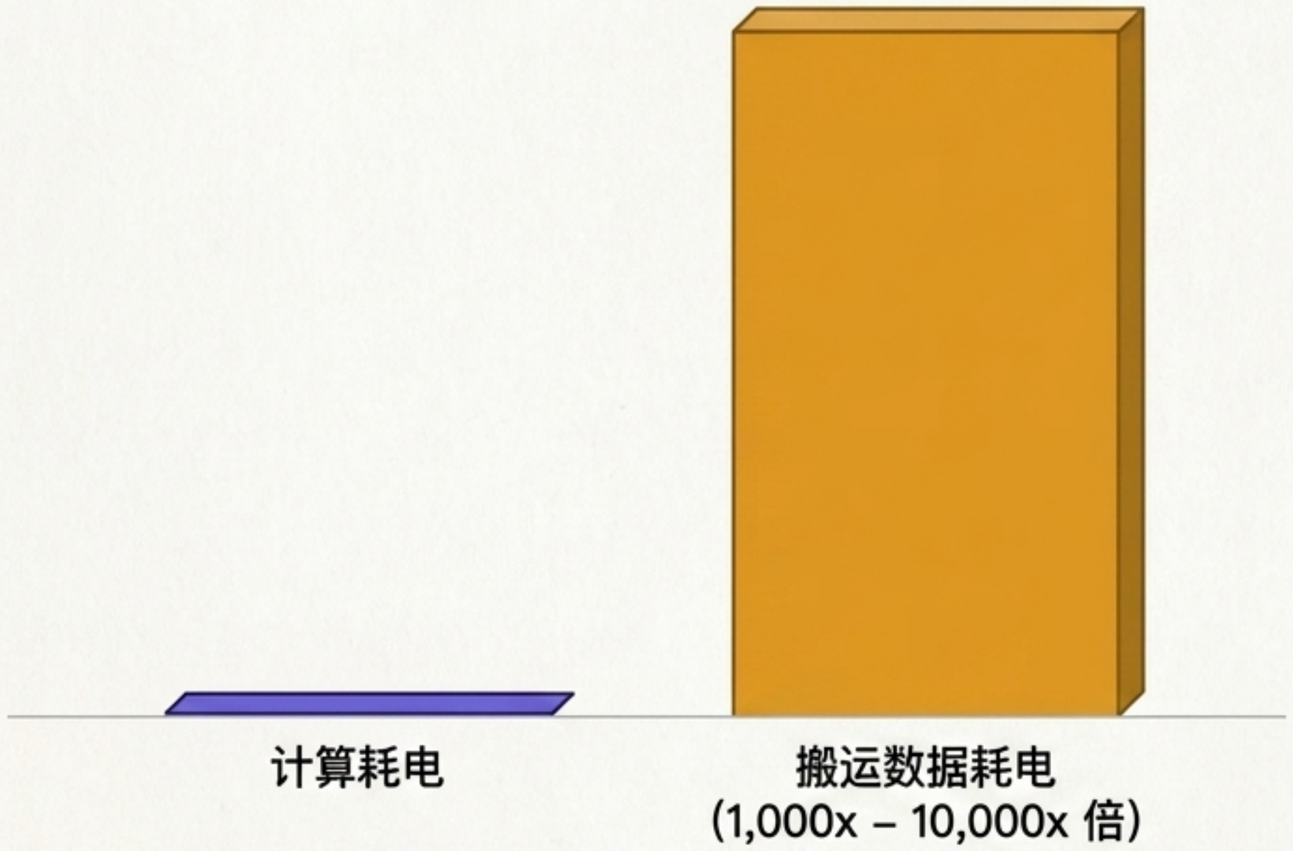
处理器的日常

在等待数据返回的这段时间里，拥有万亿次计算能力的处理器只能处于闲置状态，白白消耗资源。

被忽略的盲区

平时业界紧盯带宽数字，而这段纯粹粹受制于物理法则的延迟时间，却往往被忽略。

墙的解剖学 III：功耗（万倍的数据搬运税）



最致命的瓶颈：

同一颗芯片上的计算能耗微乎其微。但跨芯片、跨板卡搬运一次数据，耗电量是计算本身的 1,000 到 10,000 倍。

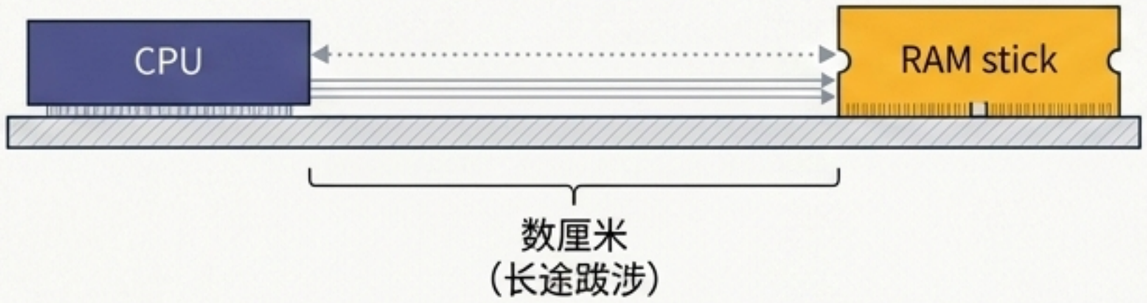
烧毁数据中心的大头：

现代 AI 数据中心的总功耗中，有 40% 到 60% 全都烧在了「把数据搬来搬去」这件事上，真正用于计算的只是零头。

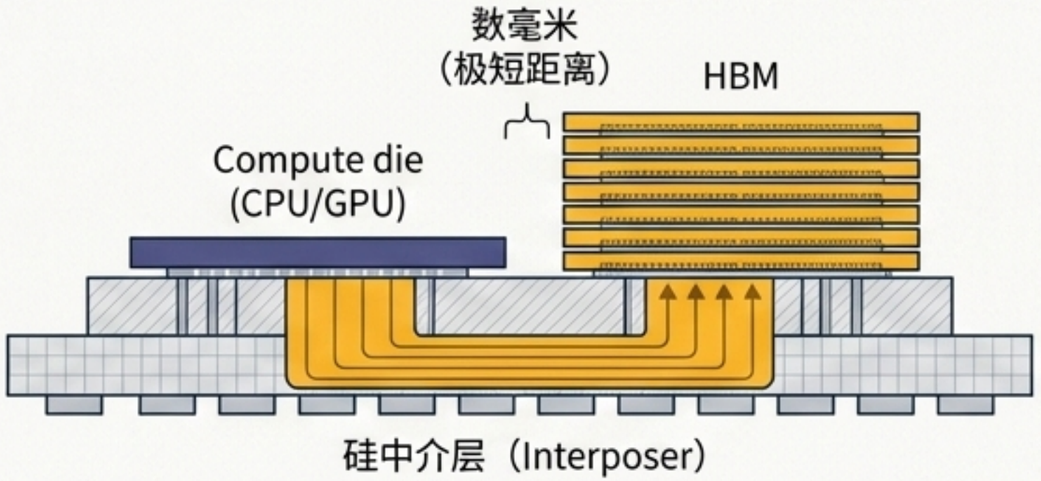
结论：

这是物理极限，工艺再进步也无法绕开。

昂贵的空间魔法：HBM（高带宽内存）



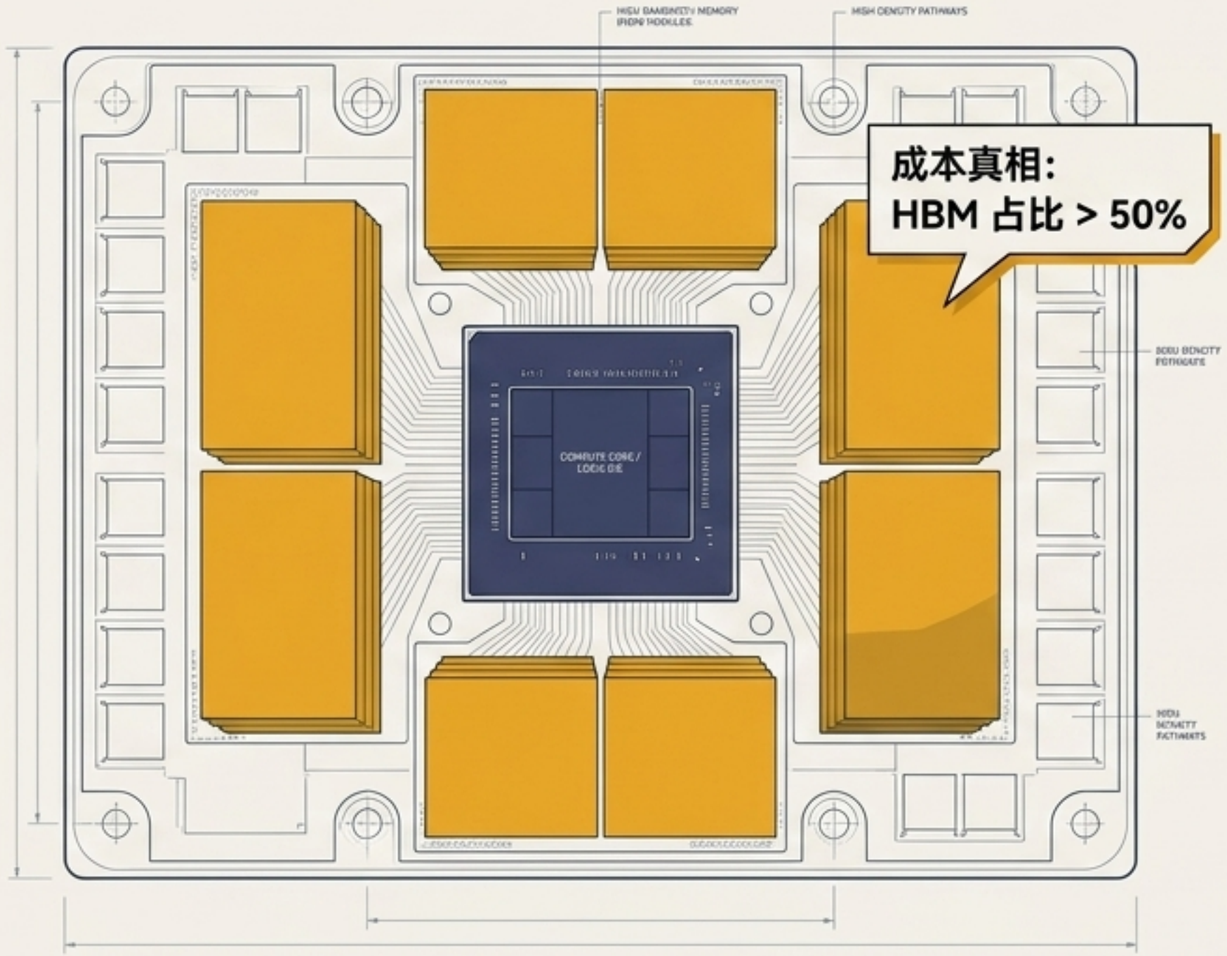
传统 DRAM：数据走线长，总线宽度受限。



HBM：放弃长距离 PCB，利用硅中介层将 DRAM 堆叠在芯片边缘。

不是更快，而是更近：HBM 的本质缩短了物理距离。路径极度缩短后，总线可以做得极其宽广从而实现指数级提升的带宽。

距离的代价：为什么 HBM 占据了主导权



并非算力太贵，而是存储太贵

在一张昂贵的 H100 显卡成本中，HBM 占比超过了 50%。AI 行业关心的功耗、成本、延迟，追根溯源，最终全部指向了内存。

存储墙是过去二十年最被低估的技术瓶颈。

下一章预告：
HBM 的良率为何如此之低？
为什么全球能稳定量产的仅有极少数厂商？
存储产业链的物理法则博弈，才刚刚开始...