大模型时代的智能芯片

李萌

人工智能研究院 & 集成电路学院

北京大学

提纲

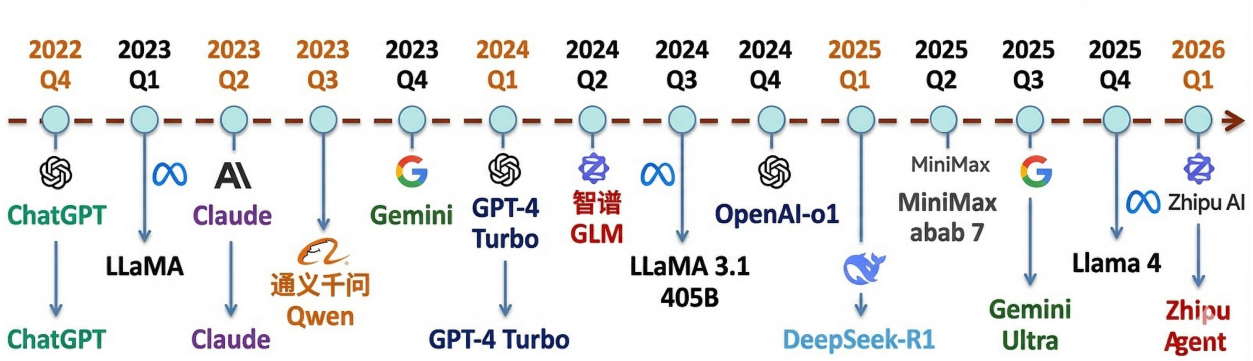
- 人工智能计算需求
- 智能芯片基础架构与指标
- 智能芯片架构演进
- 总结

提纲

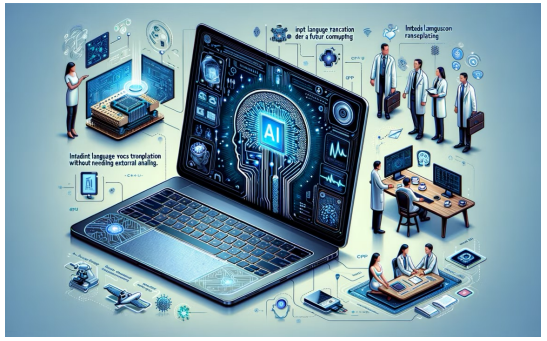
- **人工智能计算需求**
- 智能芯片基础架构与指标
- 智能芯片架构演进
- 总结

时代背景

- 自22年ChatGPT发布以来，大模型快速发展，展现出前所未有的模型能力
- 大模型的**端侧部署**具有低延迟、高隐私以及高个性化等优势，具有重要应用前景
  - 具身智能、自动驾驶、个人助手等



- 通用知识（MMLU）准确率：70% → 94%
- 数学逻辑（AIME）准确率：10% → 93%
- Agent能力（Tau-bench）准确率：0% → 89%



时代背景

大模型爆发 → 算力需求暴增 → 芯片成为战略资源

2017	2020	2022	2024
Transformer 架构提出	**GPT-3 发布**	**ChatGPT 引爆全球**	**DeepSeek R1 发布**
注意力机制革命	1750 亿参数	月活 1 亿用户	国产大模型突破

万亿级
大模型参数量级

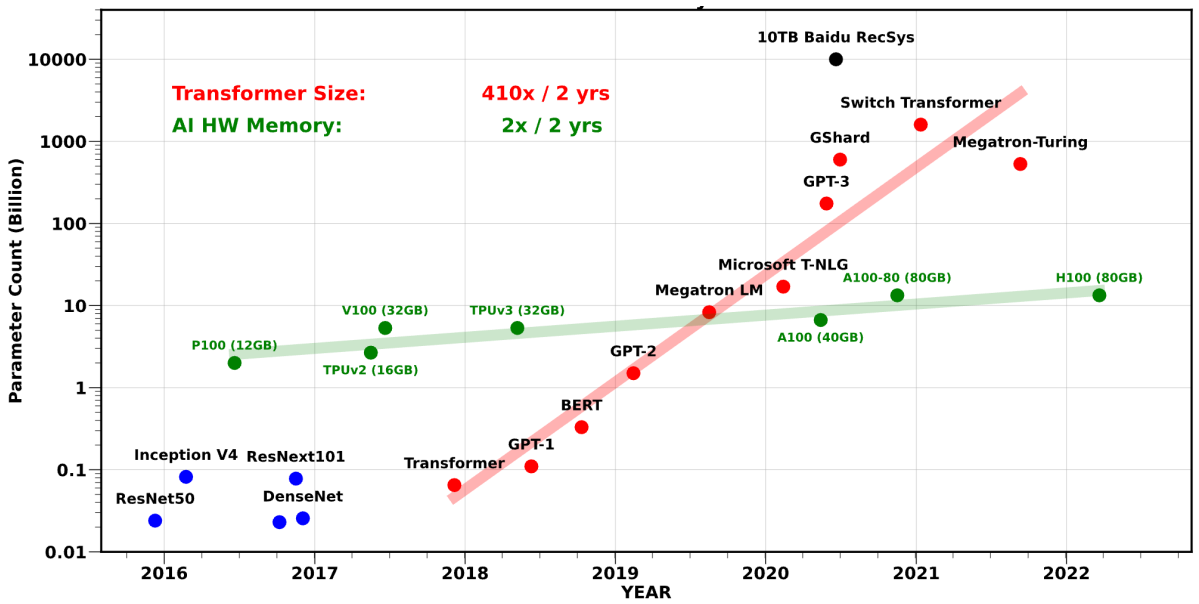
~$100M
GPT-4 训练成本

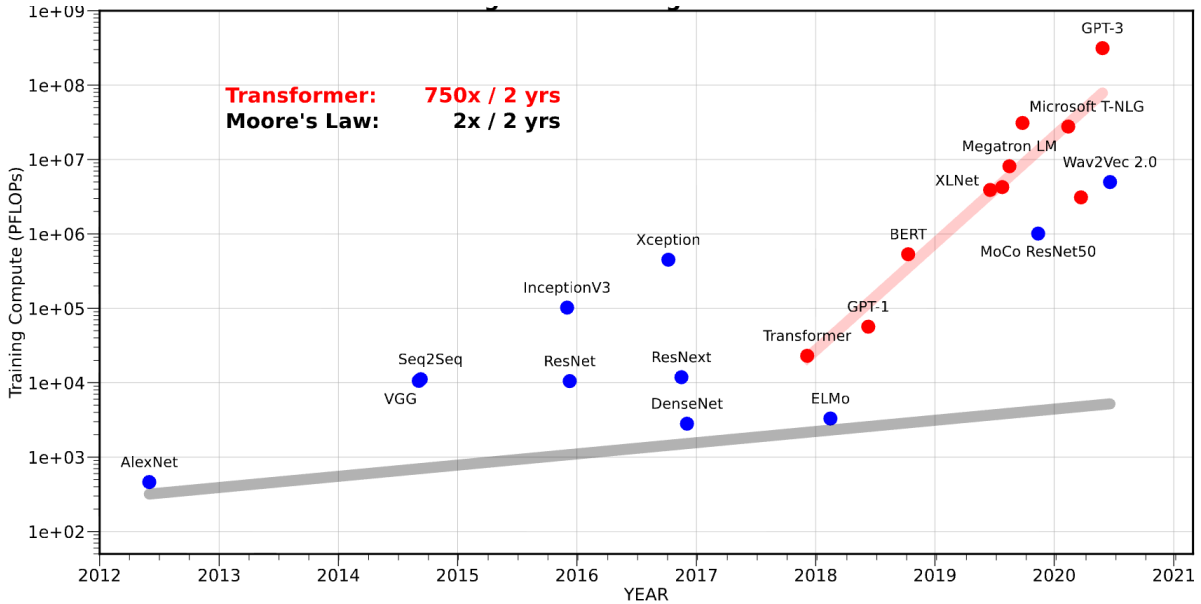
50,000+
训练用 H100 数量

卡脖子
芯片成战略制高点

时代背景

大模型爆发 → 算力需求暴增 → 芯片成为战略资源



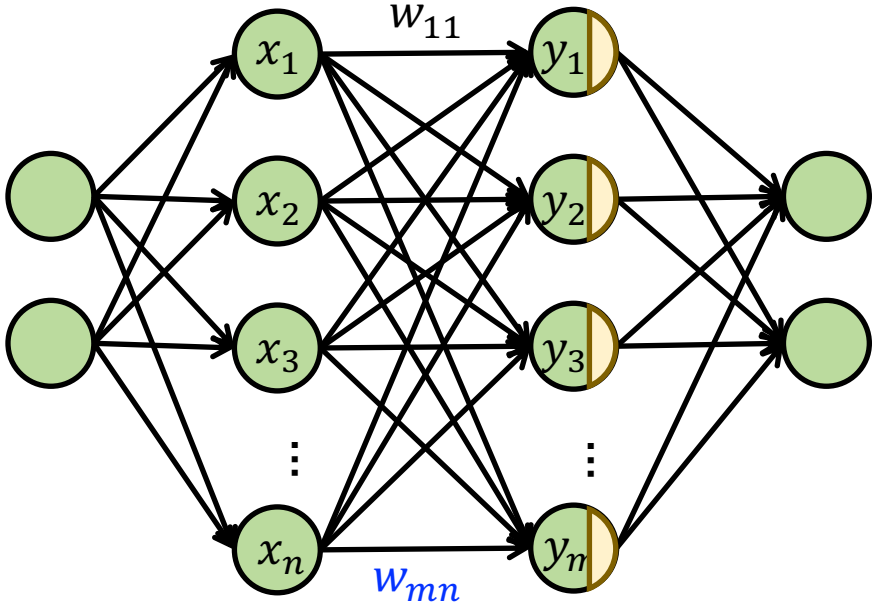


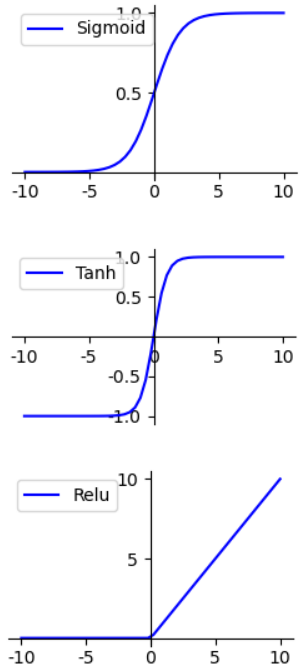
遵循Scaling Law，大模型的参数量和计算量指数级增长，进一步加剧算力以及智能芯片需求

神经网络

什么是神经网络？—— 生物神经元的数学抽象

1. **输入：** 接收来自上一层的数值信号（特征）
2. **权重：** 对每个输入赋予重要性系数 w
3. **激活：** 经过非线性函数后输出到下一层
4. **层叠：** 多层叠加 = 深度学习，逐层抽象特征

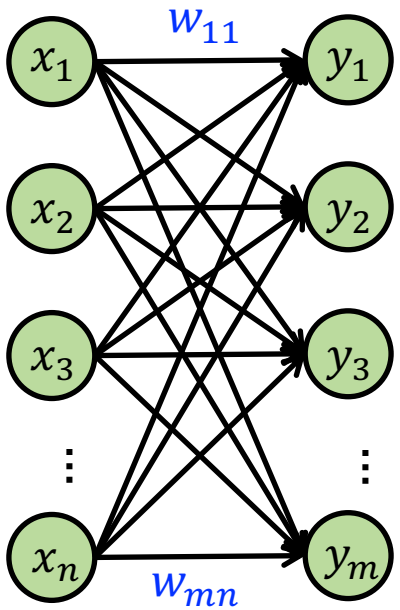




类比：神经网络就像多级投票过滤器——每层提取更抽象的特征，最终映射到目标输出

神经网络

什么是神经网络？—— 生物神经元的数学抽象



$$y_1 = x_1 w_{11} + x_2 w_{12} + \cdots + x_n w_{1n}$$

$$y_2 = x_1 w_{21} + x_2 w_{22} + \cdots + x_n w_{2n}$$

$$\vdots$$

$$y_m = x_1 w_{m1} + x_2 w_{m2} + \cdots + x_n w_{mn}$$

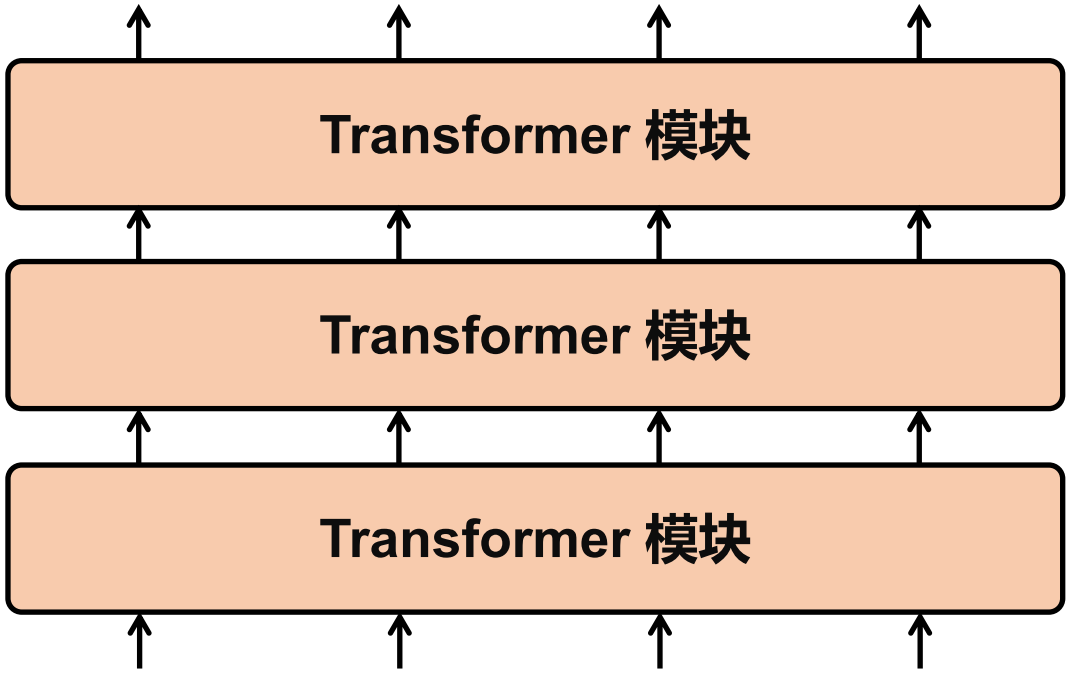
$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

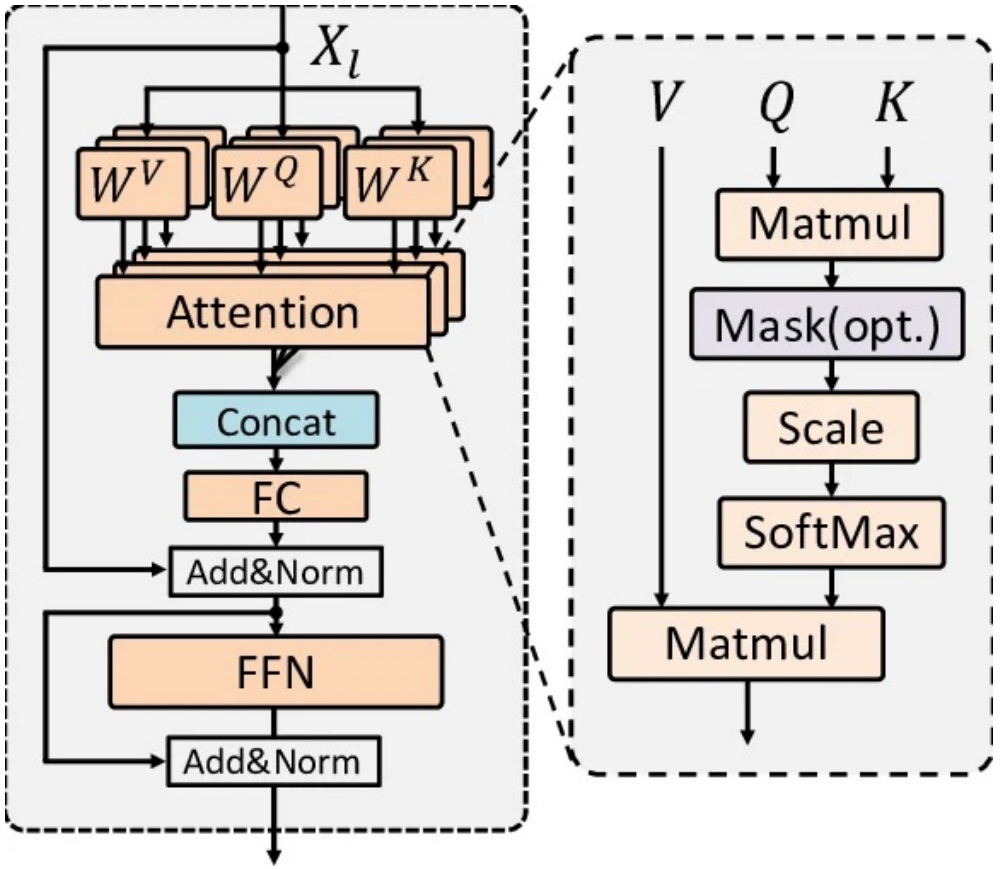
输出 **模型参数** **输入**

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1b} \\ y_{21} & y_{22} & \cdots & y_{2b} \\ \vdots & \vdots & \ddots & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mb} \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{bmatrix} \times \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1b} \\ x_{21} & y_{22} & \cdots & x_{2b} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nb} \end{bmatrix}$$

线性层的核心计算为矩阵乘法，主要是乘累加计算，神经网络几乎所有参数来自线性层

大模型的基础模块：Transformer模块





大模型由重复的Transformer模块组成，每个Transformer模块包含一系列的线性和非线性算子

神经网络计算需求

计算量的度量单位：FLOPs

FLOP 是什么?

Floating Point Operation（浮点运算次数）

衡量模型计算量的基本单位

一次乘法或加法 ≈ 1 FLOP

经验公式（大模型推理）

所需 FLOPs ≈ 参数量 × 2

GPT-3 示例：1750亿 × 2 ≈ 3.5×10^{11} FLOPs

单位换算

10^9	GFLOPS	早期 AI 模型，一个卷积层
10^{12}	TFLOPS	现代 GPU 单卡峰值算力量级
10^{15}	PFLOPS	
10^{18}	EFLOPS	

神经网络计算需求

神经网络计算的关键特性

基础计算简单

神经网络计算主要以乘累加为主，计算简单，且可以量化到低比特进行计算

计算高度并行

线性层涉及大量乘累加计算，且同一线性层内计算可以完全独立并行进行

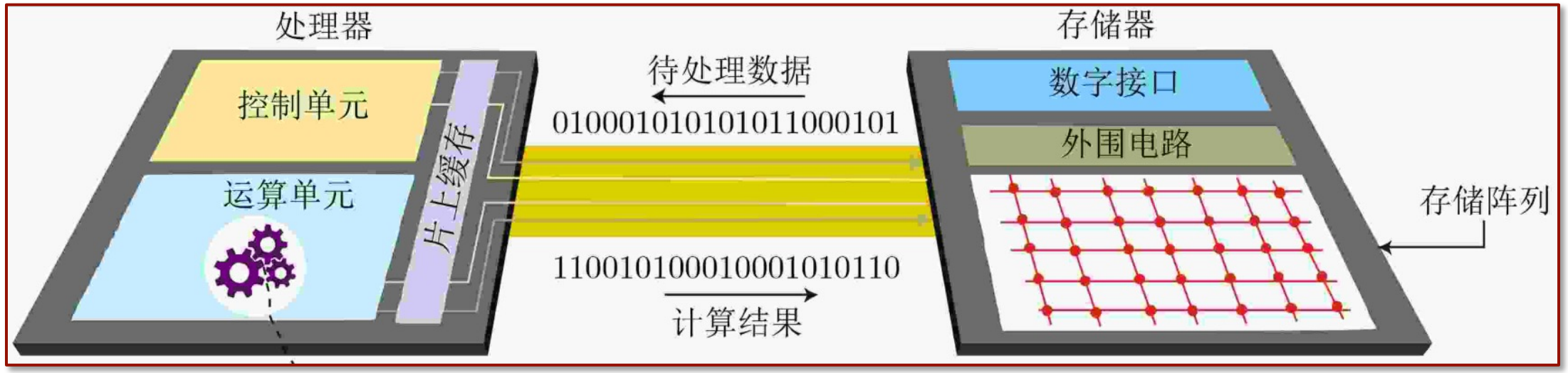
计算确定性强

神经网络计算确定性强，无需进行过多条件判断

提纲

- 人工智能计算需求
- **智能芯片基础架构与指标**
- 智能芯片架构演进
- 总结

冯诺依曼架构：现代计算机的基石



- **智能芯片**是实现边缘智能的**基础平台**，传统芯片往往采用**冯•诺依曼架构**
  - 核心组成部分包括存储器、中央处理器、数据传输总线等
- 模型参数存储在存储器中，计算需要在处理器进行，因此需要通过数据总线把模型参数载入到处理器中

智能芯片关键指标

关键指标 ①：算力（FLOPS / TOPS）

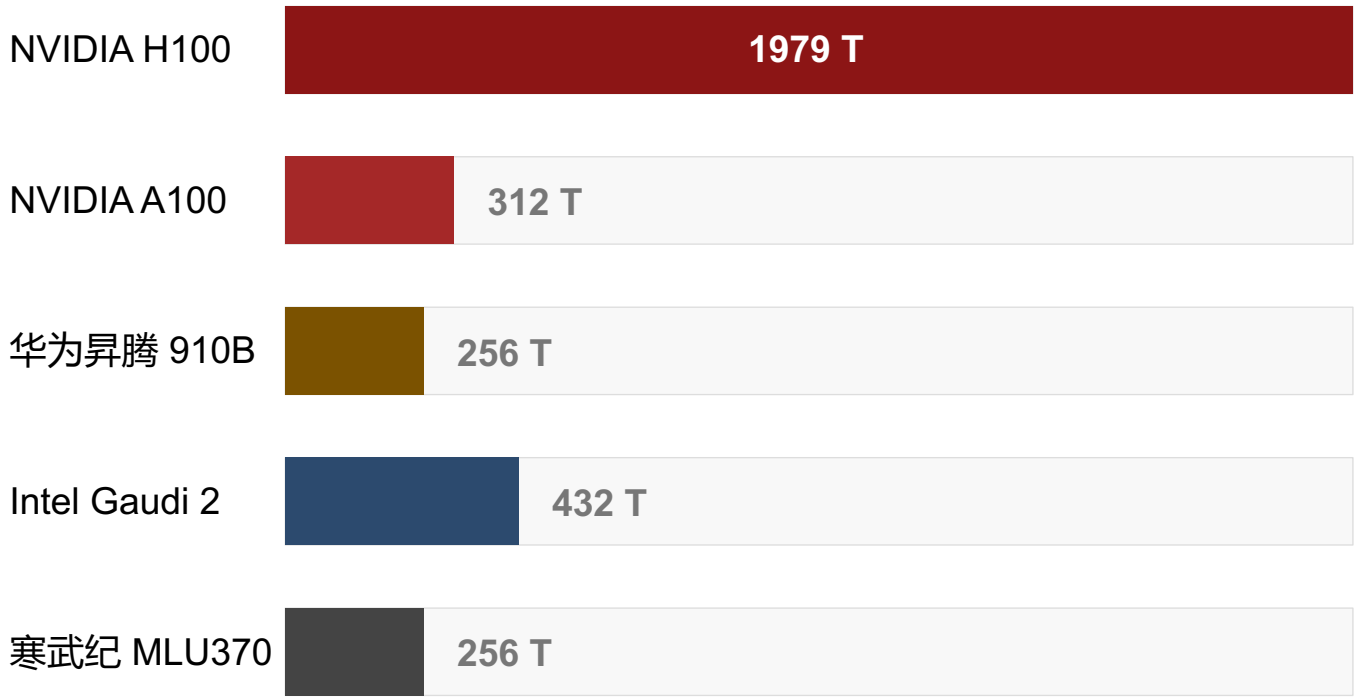
定义：每秒能完成多少次浮点运算

FLOPS = FLoating-point Operations Per Second

精度对算力的影响

1x	FP64（双精度）	科学计算
2x	FP32（单精度）	通用 AI 训练
8x	FP16（半精度）	混合精度训练
16x	INT8（整数）	推理加速

主流 AI 芯片算力对比（FP16 TFLOPS）



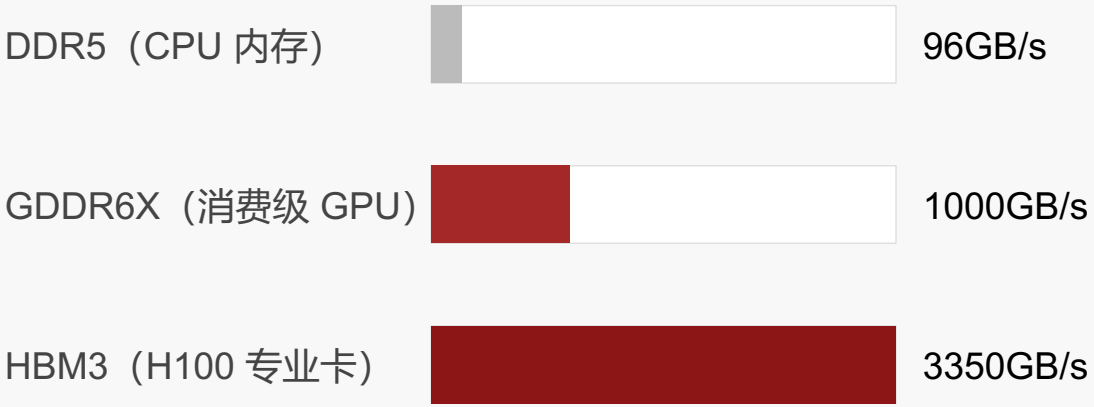
智能芯片关键指标

关键指标 ②③：内存带宽 与 显存容量

② 内存带宽（Memory Bandwidth）

定义：单位时间内 CPU/GPU 能从内存读取的数据量，单位 GB/s

类比：高速公路的车道数



③ 显存容量与 HBM

为什么大模型需要大显存?

- 模型权重必须全部加载到显存（70B 模型 ≈ 140GB FP16）
- 训练时还需存梯度、优化器状态（约 3 倍权重大小）
- 推理时 KV Cache 随序列长度快速膨胀

HBM（高带宽内存）

多层 DRAM 垂直堆叠，宽总线连接 GPU，带宽是 GDDR 的 3-5 倍，但成本高约 5 倍，是数据中心 AI 卡的标配。

智能芯片关键指标

关键指标 ④⑤：能效比 与 多卡互联带宽

④ 能效比（TOPS/W 或 FLOPS/W）

为什么能效比很重要?

~30MW	单个大型 AI 数据中心峰值功耗
$0.1/kWh	算力成本很大一部分来自电费
GPT-4	训练一次碳排放约等于 500 辆汽车一年

能效比越高 = 相同功耗下算力越强，对数据中心 TCO（总拥有成本）有决定性影响

⑤ 互联带宽（NVLink / InfiniBand）

大模型训练为何需要多卡?

单卡显存不够装下整个模型，必须将模型切分到多块 GPU 上并行训练（模型并行/数据并行）

卡间互联对比

PCIe 5.0	128 GB/s	普通服务器
NVLink 4.0	900 GB/s	NVIDIA H100
NVSwitch	3.6 TB/s	8 卡互联全带宽

智能芯片关键指标

关键指标综合对比：主流 AI 芯片横向评测

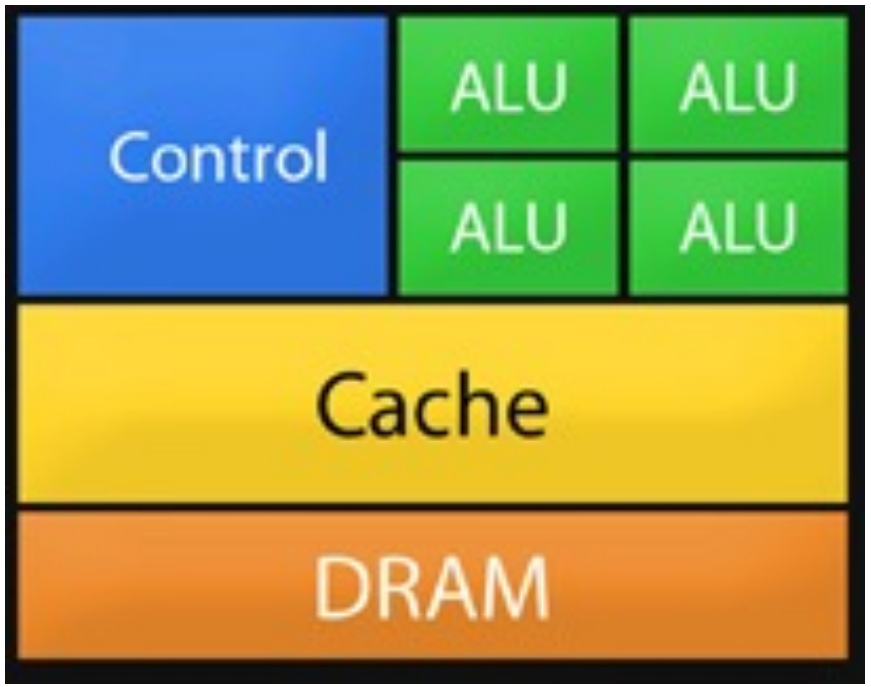
芯片型号	算力 FP16 (TFLOPS)	显存容量	显存带宽 (GB/s)	TDP 功耗 (W)	互联带宽
NVIDIA H100 SXM	1,979	80 GB HBM3	3,350	700 W	NVLink 900 GB/s
NVIDIA A100 SXM	312	80 GB HBM2e	2,000	400 W	NVLink 600 GB/s
华为昇腾 910B	256	64 GB HBM2e	2,000	400 W	HCCS 392 GB/s
寒武纪 MLU370	256	96 GB LPDDR5	614	250 W	MLU-Link
Google TPU v4	275	32 GB HBM2	1,200	170 W	ICI 1.2 TB/s

数据来源：各厂商官方白皮书（截至 2024 年）。H100 因 NVLink 与 HBM3 的组合，在 AI 训练中总体领先 1-2 代。

提纲

- 人工智能计算需求
- 智能芯片基础架构与指标
- **智能芯片架构演进**
- 总结

智能芯片架构演进：CPU



计算核心数量少，但每个核计算能力极强

CPU 擅长的任务

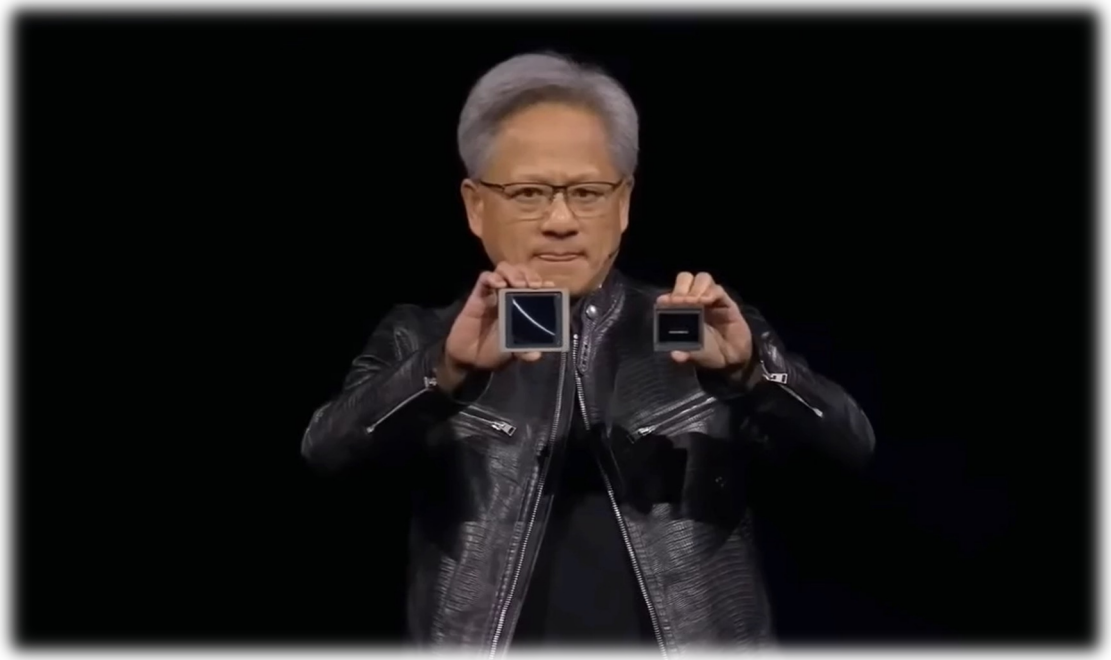
- **复杂控制流：** if/else、递归、异常处理，逻辑分支丰富
- **低延迟响应：** 单线程延迟极低，适合实时交互任务
- **通用操作系统：** 运行 OS 和各类通用软件，生态最广

CPU 做 AI 的瓶颈

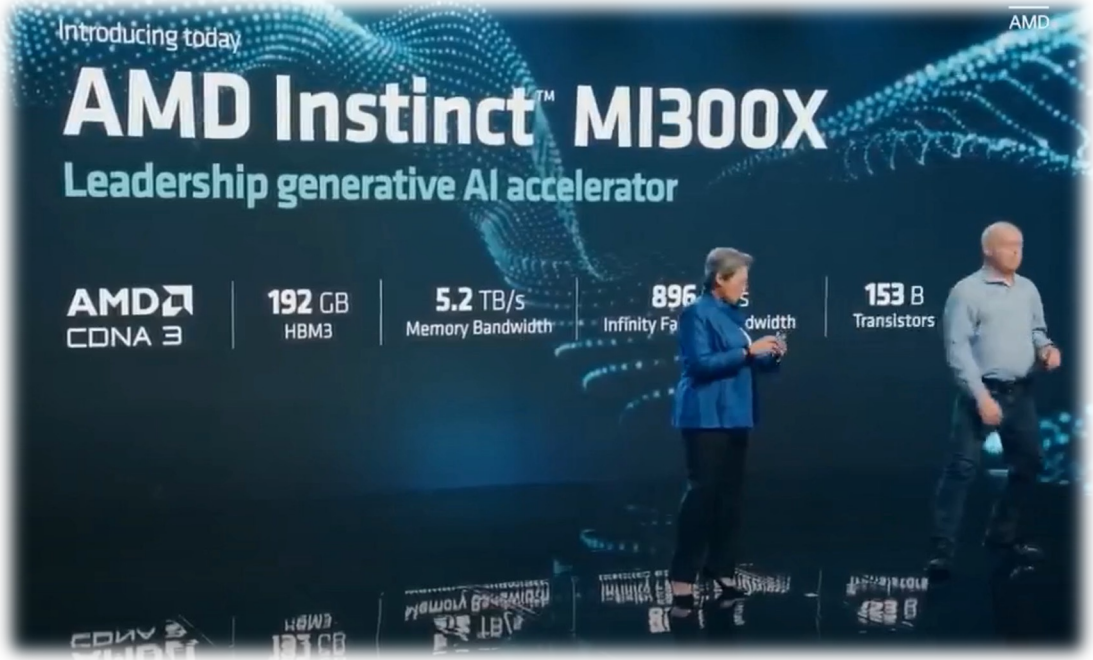
- **核心少：** 数十核，并行度远不足以处理亿级矩阵运算
- **主频≠并行：** 提升频率只加快串行任务，矩阵乘法受益有限
- **带宽不足：** DDR5 带宽 ~96GB/s，难以喂饱大模型权重搬运
- **控制逻辑复杂：** 神经网络计算逻辑简单，浪费了控制逻辑

智能芯片架构演进：GPU

GPU：Graphics Processing Unit 图形处理单元

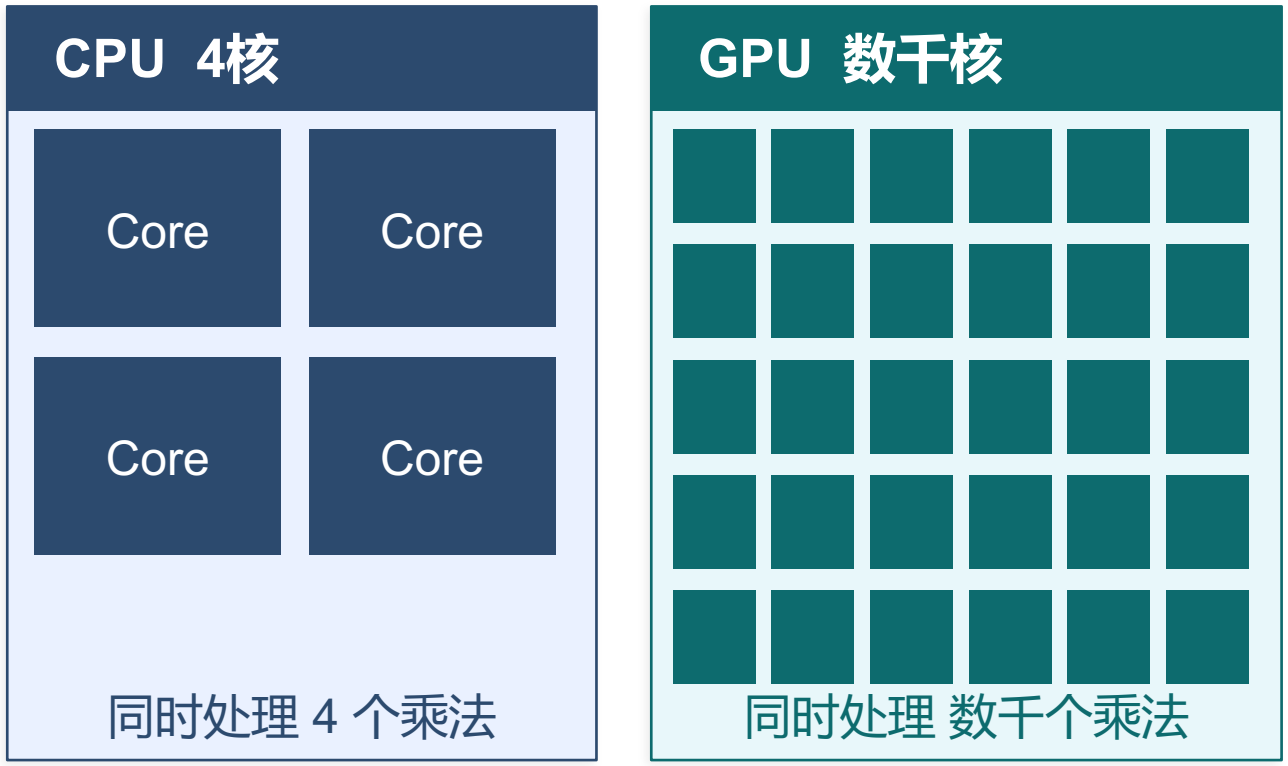


黄仁勋 (英伟达CEO) 展示B200 GPU
2024年



苏姿丰 (AMD CEO) 展示MI300X GPU
2023年

智能芯片架构演进：GPU



1999 — **NVIDIA GeForce 256**

首款商用 GPU，用于图形渲染

2006 — **CUDA 发布**

允许用 C 语言编程 GPU，GPGPU 时代开启

2012 — **AlexNet 引起轰动**

GPU 训练深度网络速度是 CPU 的 50 倍

2017 — **Volta / Tensor Core**

专为矩阵乘法的混合精度计算单元

2022+ — **H100 / B200**

AI 专用 HBM3+NVLink，算力千级 TFLOPS

智能芯片架构演进：GPU

GPU 内部结构：SM 阵列

数千个简单 SM 并行，天然匹配矩阵乘

Tensor Core：AI 专属加速单元

什么是 Tensor Core

Volta 架构（2017）引入，专门执行 D=A×B+C 的混合精度矩阵乘加，一条指令完成 4×4 矩阵运算

精度支持

FP64/FP32/TF32/FP16/BF16/INT8/INT4，每降一档精度算力翻倍

性能跃升

H100 含 528 个 Tensor Core，FP16 算力达 1979 TFLOPS，是仅靠 CUDA Core 的 8× 以上

CUDA 护城河：百万开发者 + cuDNN/cuBLAS 算子库，是最难复制的竞争壁垒

智能芯片架构演进：GPU

A100

7nm 540亿晶体管

620TFLOPS@BF16

面积：826mm²

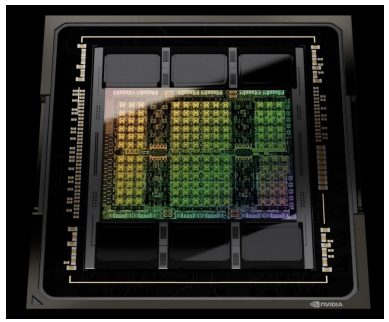
2021

H100

4nm 800亿晶体管

4000TFLOPS@FP8

面积：826mm²



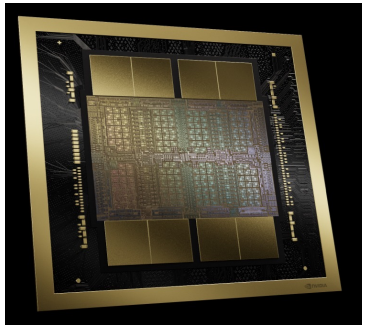
2023

B200

4nm 2080亿晶体管

20000TFLOPS@FP4

面积：2×826mm²



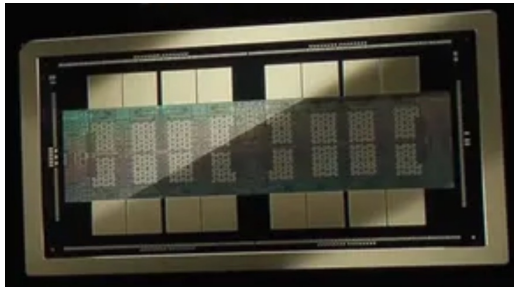
2024

Rubin GPU

3nm 3360亿晶体管

100000TFLOPS@FP4

面积：4×826mm²

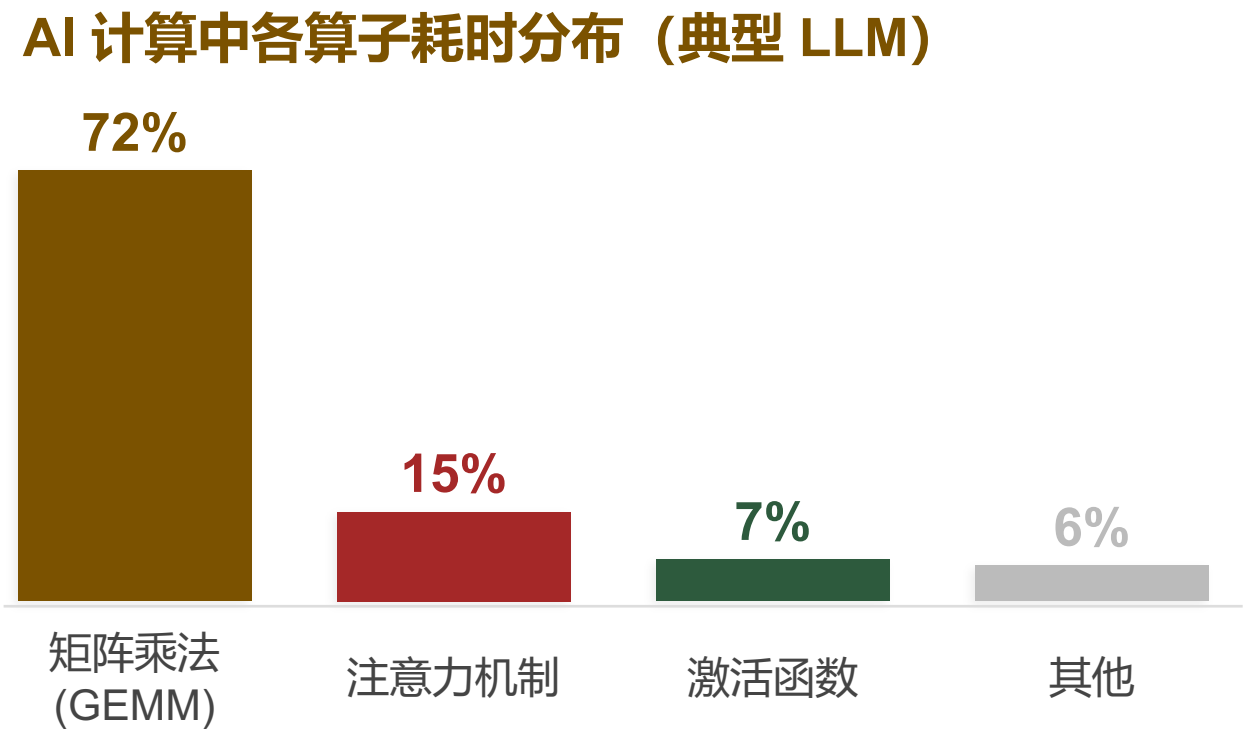


2026

智能芯片架构演进：DSA

DSA：领域专用架构的崛起

当 90% 的算力都花在矩阵乘法上，通用设计就会引起浪费

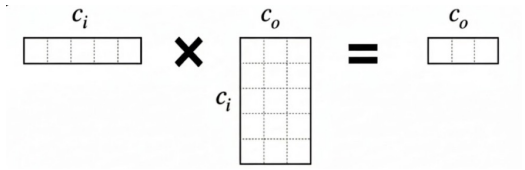


结论：针对矩阵乘法做极致优化，可获得最大收益

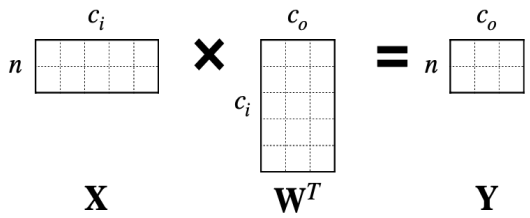
智能芯片架构演进：DSA

Google TPU：脉动阵列与专用架构的优化

```
1: procedure FullyConnectedLayer(X, W)
2:     for b ← 1 to B do
3:         for c_o ← 1 to C_o do
4:             Y[n][c_o] ← 0
5:             for c_i ← 1 to C_i do
6:                 Y[b][c_o] ← Y[b][c_o] + X[n][c_i] × W[c_i][c_o]
```

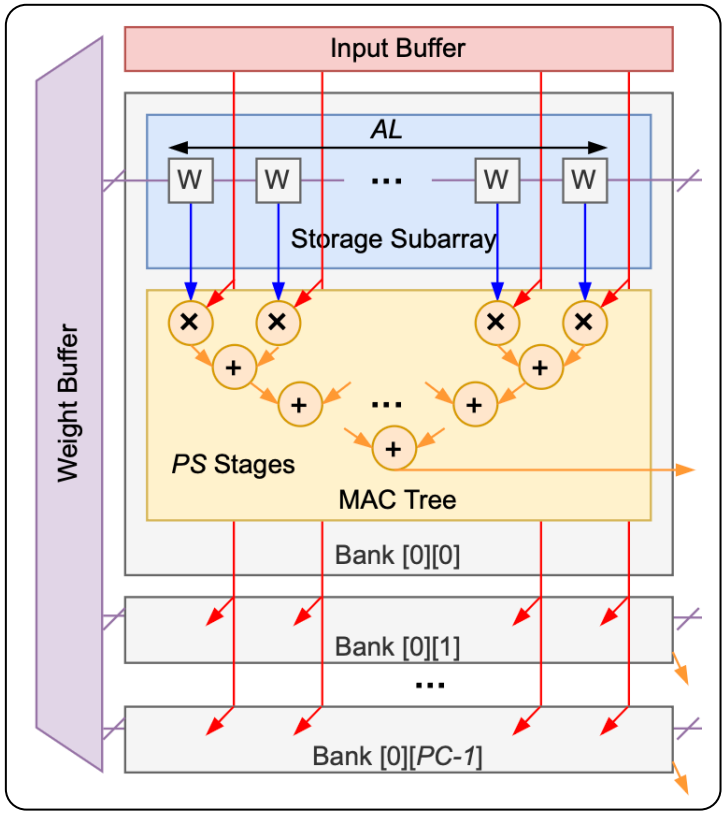


矩阵-向量乘 (GEMV)

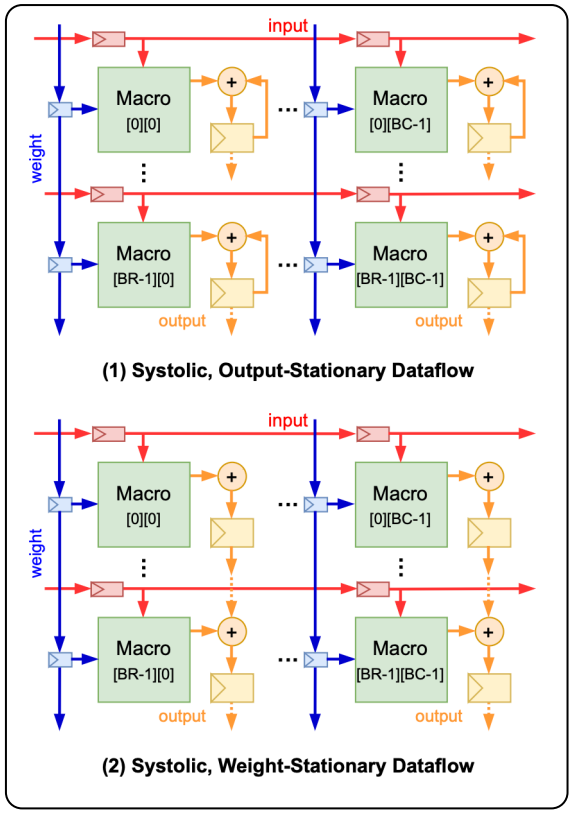


矩阵-矩阵乘 (GEMM)

(a) Tensor计算



(b) 数据广播、加法树



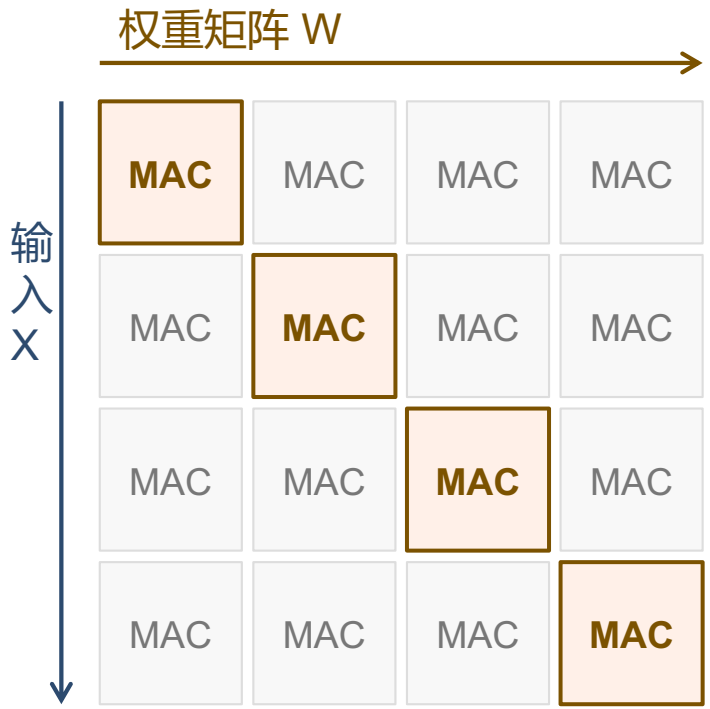
(c) 脉动阵列

智能芯片架构演进：DSA

Google TPU：脉动阵列与专用架构的优化

脉动阵列（Systolic Array）

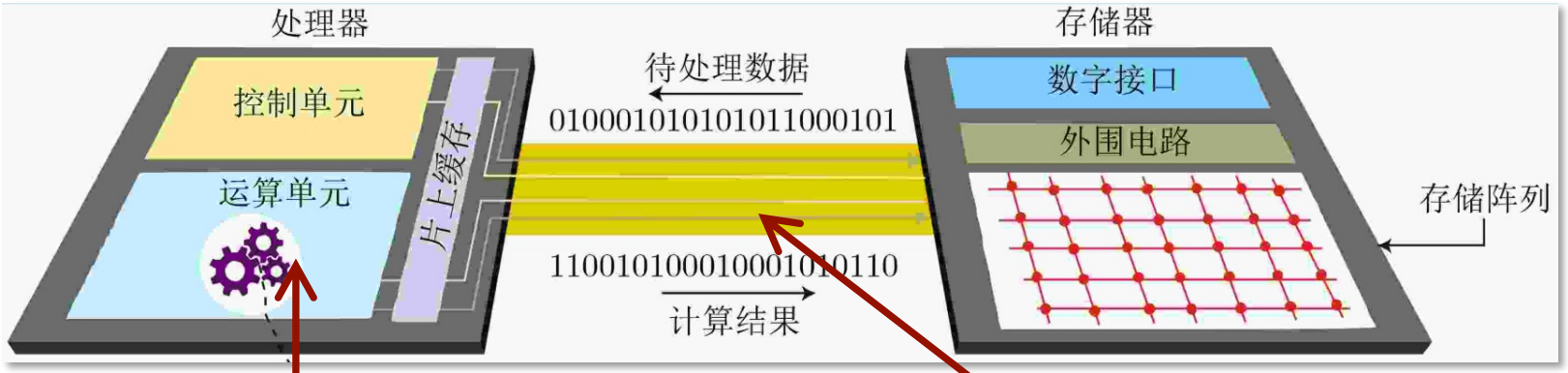
数据像脉搏一样从左→右、上→下流过计算阵列，每个节点执行一次乘累加，结果自然流出，全程无需读写内存

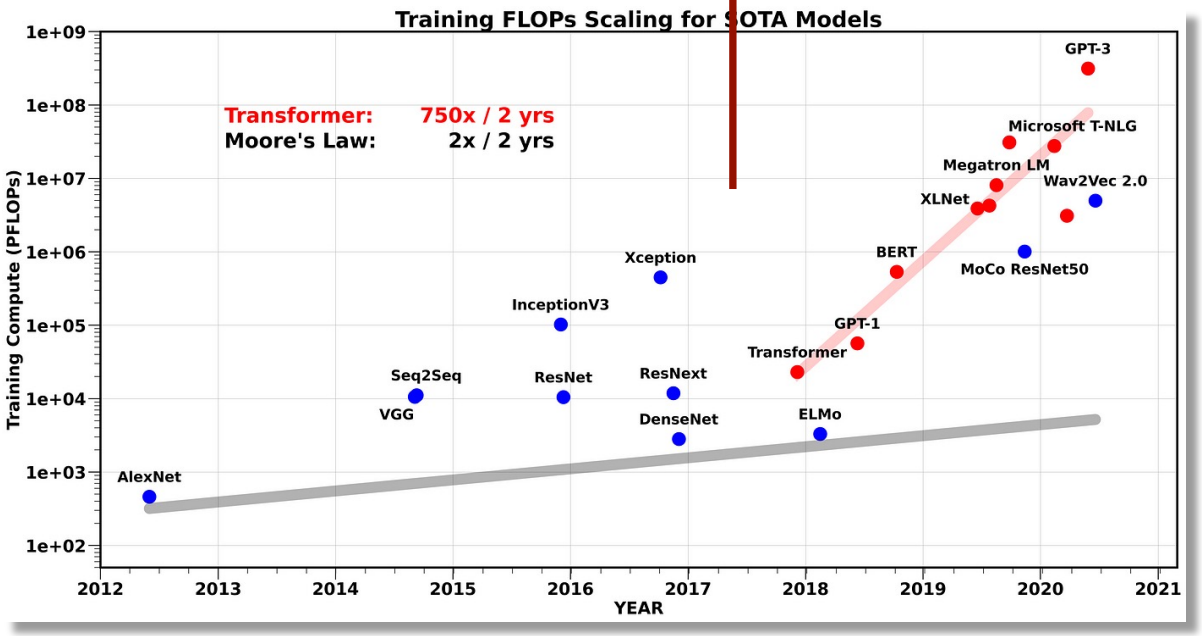


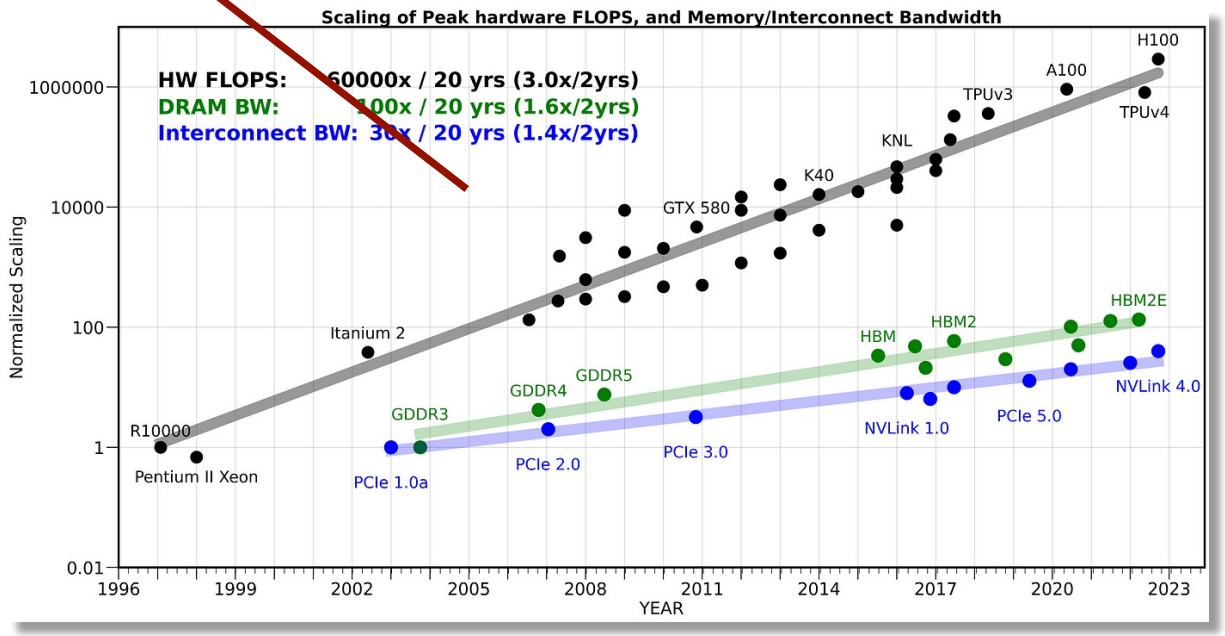
TPU vs GPU：设计哲学对比

维度	GPU	TPU/DSA
设计目标	通用并行计算	仅为矩阵乘法优化
计算单元	CUDA Core + Tensor Core	超大脉动阵列（256×256）
内存访问	访问 HBM，仍有搬运	数据复用极高，内存访问最小化
可编程性	高	较低
能效比	中（通用设计代价）	极高（专用设计收益）
适用场景	训练+推理全场景	大规模矩阵训练/推理

智能芯片架构演进：新兴架构





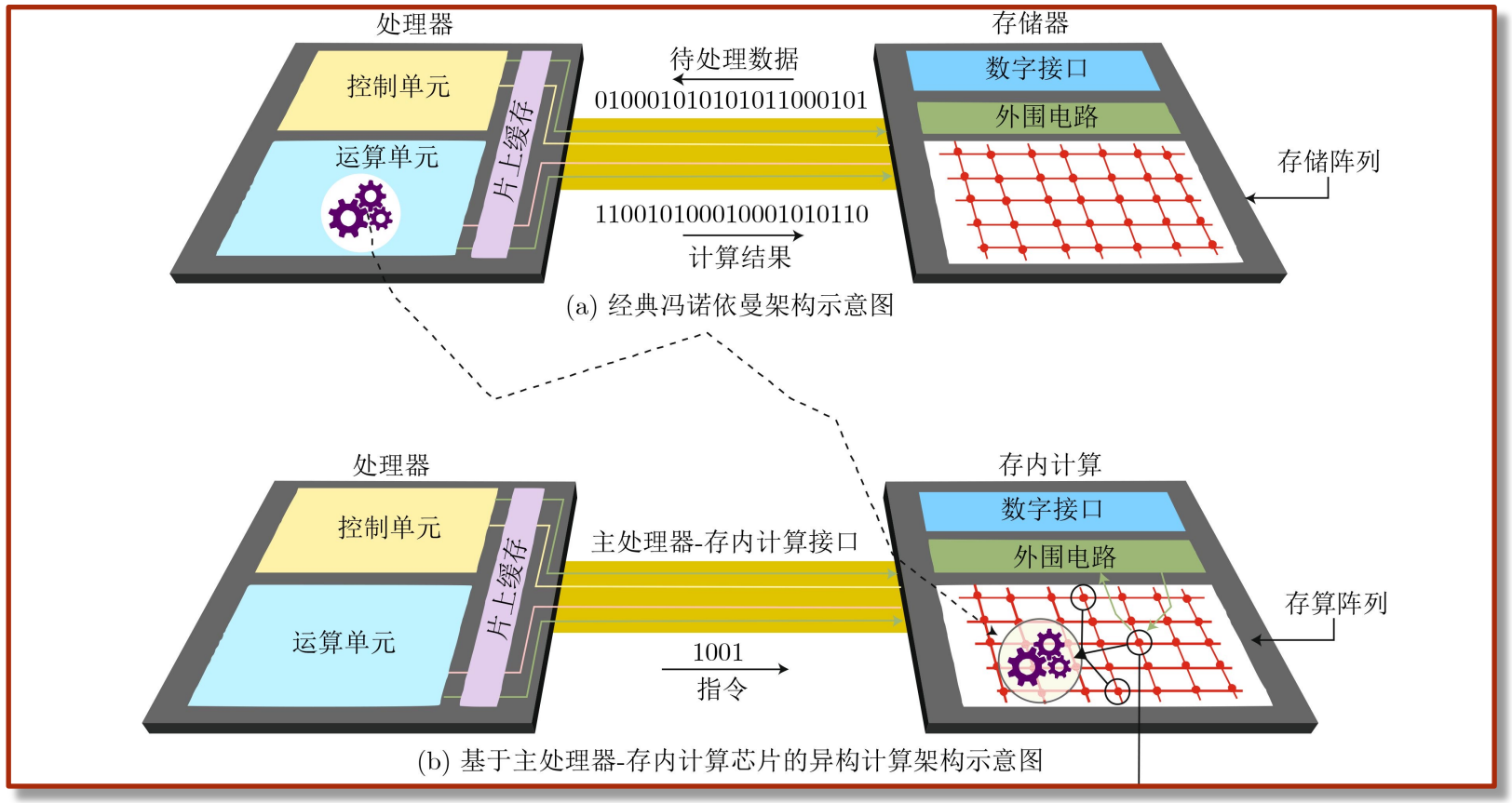


GPU、TPU仍然遵循**冯•诺依曼架构**，存储和计算分离，造成严峻的带宽瓶颈，称为**存储墙**问题

智能芯片架构演进：新兴架构

存算一体（PIM）

把计算搬进内存，消除数据搬运的能耗



核心优势

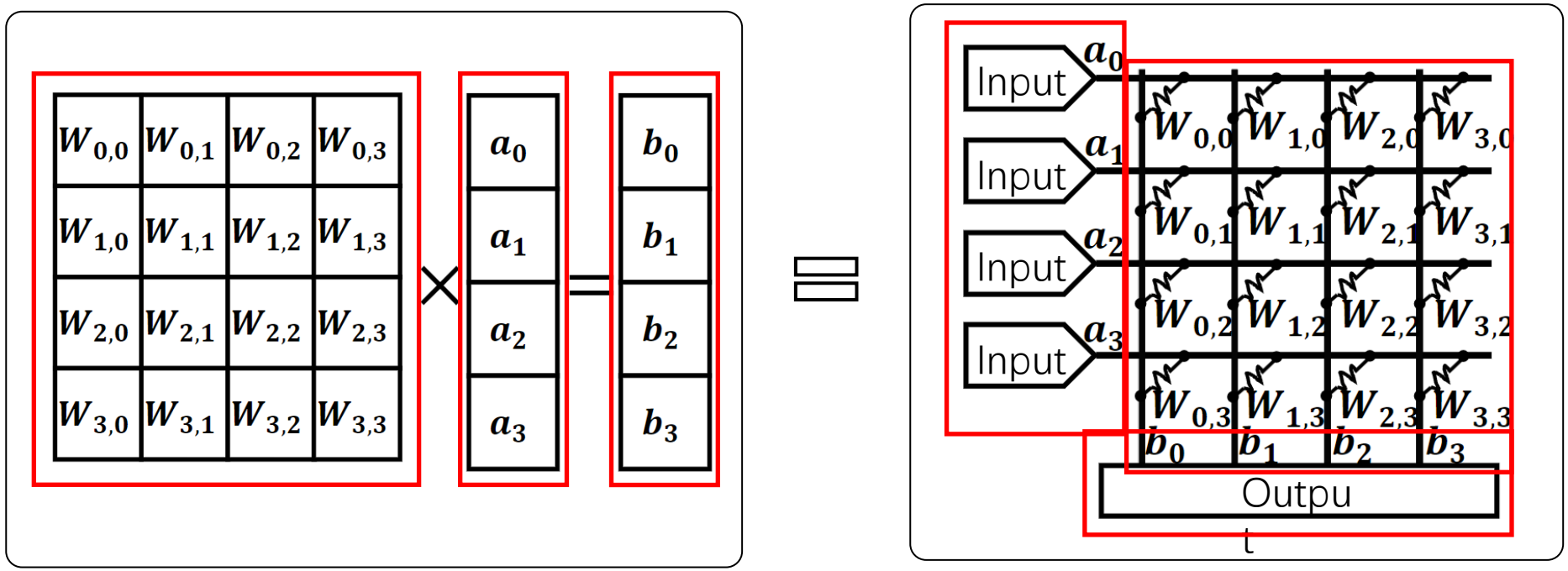
- 带宽提升 10-100x，功耗大幅降低
- 解决 Memory Wall，特别适合大模型推理

智能芯片架构演进：新兴架构

存算一体（PIM）

把计算搬进内存，消除数据搬运的能耗

- 存内计算：设计高密度、高能效的矩阵-向量乘（GEMV）计算的Tensor Core



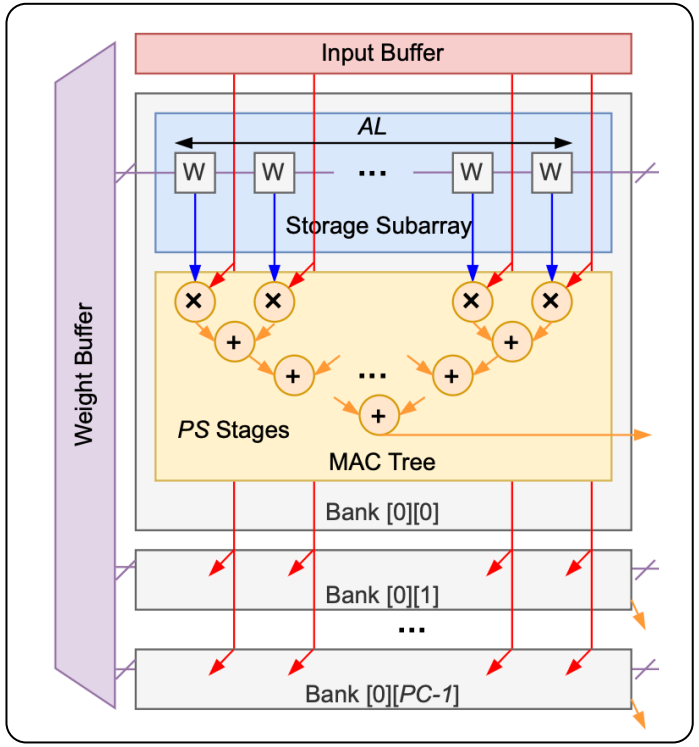
(a) 矩阵-向量乘操作　　(b) 对应的存内计算阵列

智能芯片架构演进：新兴架构

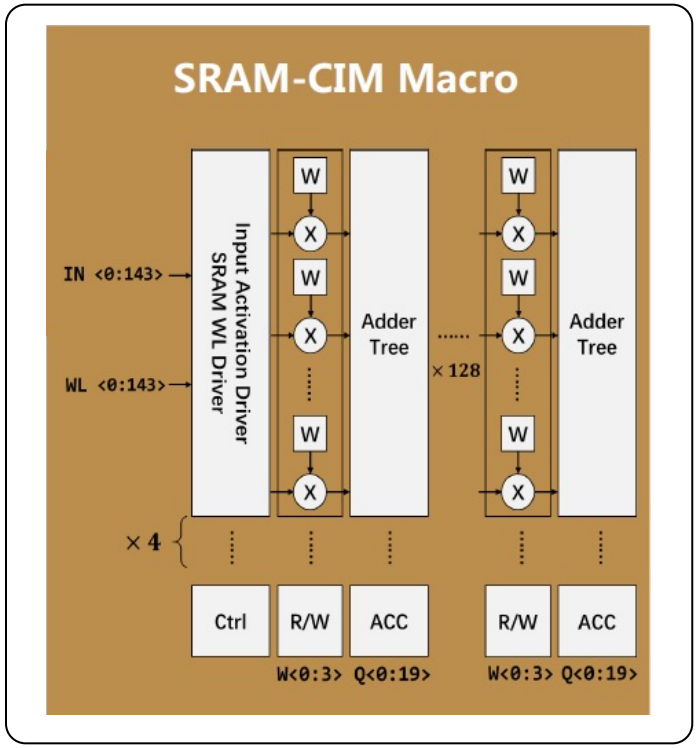
存算一体（PIM）

把计算搬进内存，消除数据搬运的能耗

- 定制的1bit乘法、加法树、累加器、SRAM替代寄存器
- 多个单元融合，规整的版图：面效提升 2x+、能效提升 5x+



(a) 数据广播、加法树



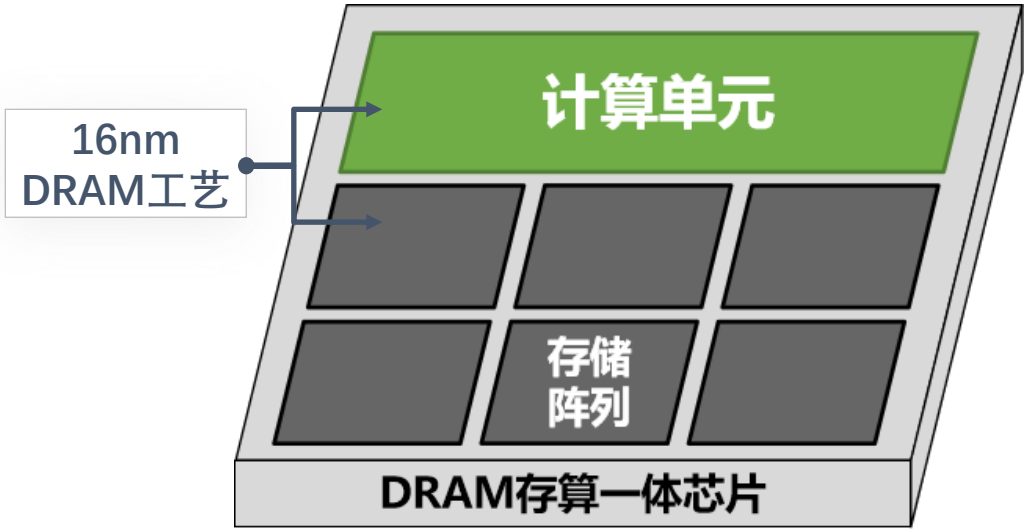
(b) 同规格CIM阵列

智能芯片架构演进：新兴架构

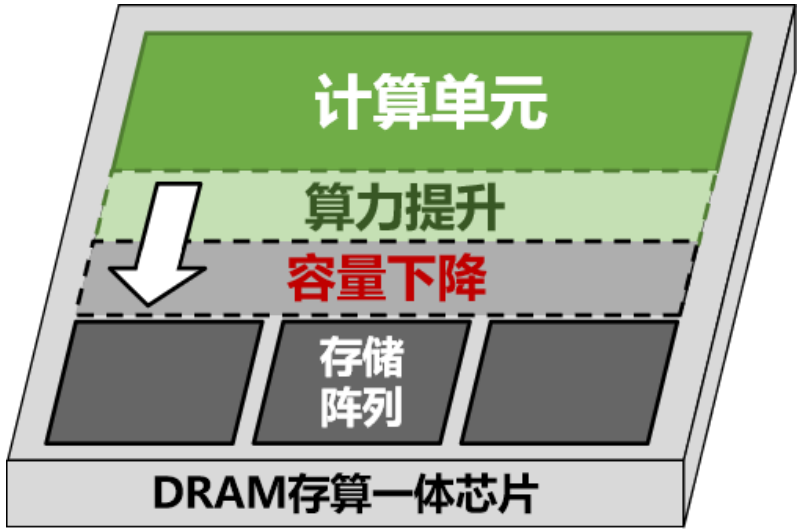
存算一体（PIM）

把计算搬进内存，消除数据搬运的能耗

- 近存：将计算单元置于Bank附近，利用DRAM内部并行提升带宽



(a) DRAM工艺限制



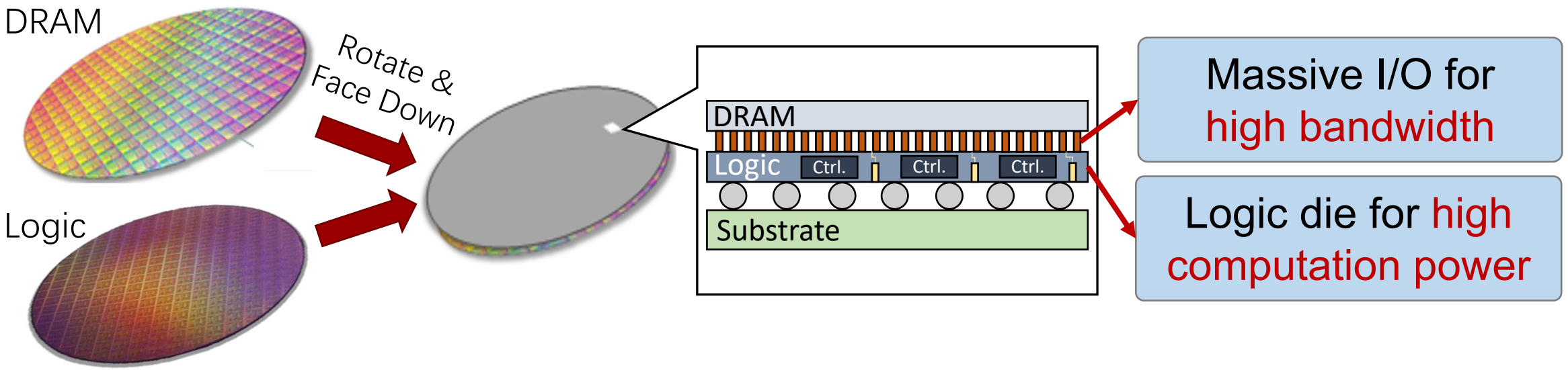
(b) 算力与容量相互制约

智能芯片架构演进：新兴架构

存算一体（PIM）

把计算搬进内存，消除数据搬运的能耗

- **近存：将计算单元置于Bank附近，利用DRAM内部并行提升带宽**



智能芯片架构演进：新兴架构

其他方向：FPGA 加速、光子芯片与类脑计算

FPGA 可编程逻辑加速

是什么	现场可编程门阵列，电路级别可重配置
优势	灵活性极高、功耗低、延迟确定性强
挑战	开发门槛高、算力密度低于GPU/DSA
应用	金融风控、通信基站、视频编解码
代表	赛灵思（AMD）、英特尔Altera、国内复旦微

光子芯片（Photonic）

是什么	用光子传输和计算，光速传输几乎无延迟、超低功耗
优势	带宽极高（THz 级）、理论功耗趋零、无电磁干扰
挑战	精度控制难、与电路集成挑战大、制造复杂
应用	光互联已商用；光计算仍在实验室验证
代表	Lightmatter、Luminous、国内曦智科技

类脑芯片（Neuromorphic）

是什么	模拟生物神经元脉冲，事件驱动，异步处理，极低功耗
优势	静态功耗趋零、生物启发型时序处理、边缘节能
挑战	与现有 AI 框架完全不兼容、编程模型全新
应用	IoT 传感器融合、脑机接口、低功耗边缘推理
代表	Intel Loihi 2、IBM TrueNorth、北大PAICORE

智能芯片架构演进：国内外主流智能芯片

国际主流

NVIDIA H100/B200

数据中心训练
推理全覆盖

CUDA 生态 + HBM3 + NVLink

Google TPU v5

自用训练 +
Cloud TPU

脉动阵列 + XLA 软硬协同

AMD MI300X

数据中心 GPU
CPU 混合架构

192GB HBM 显存容量最大

国产主力

华为 昇腾910B/C

数据中心训练
国内最强

达芬奇架构 + HCCS 互联

寒武纪 MLU370

云端推理
政务场景

指令集自研，生态建设中

地平线 征程6

自动驾驶
边缘推理

BPU 专用架构，车规量产

阿里 真武810E

大模型训练
数据中心

阿里平头哥自研芯片

智能芯片架构演进：不同架构横向对比

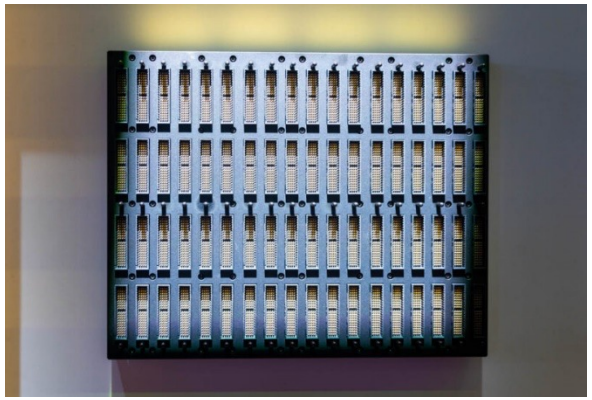
架构	并行度	内存带宽	能效比	可编程性	AI 适用场景	代表芯片
CPU	低（数十核）	中 96GB/s	中	最高	推理辅助、控制	Intel Xeon
GPU	高（数千核）	高 3TB/s	中	高	训练+推理全场景	NVIDIA H100
FPGA	中（可配置）	中高	高	中	特定推理加速	Xilinx U280
TPU/DSA	极高（专用）	极高	极高	低	大规模矩阵训练	Google TPU v5
PIM	极高（分布）	超高 10-100×	极高	极低	LLM 推理	Samsung HBM-PIM

没有最好的架构，只有最合适的架构。理解 Trade-off，是理解整个 AI 芯片产业格局的钥匙。

智能芯片架构演进：超节点多卡集群

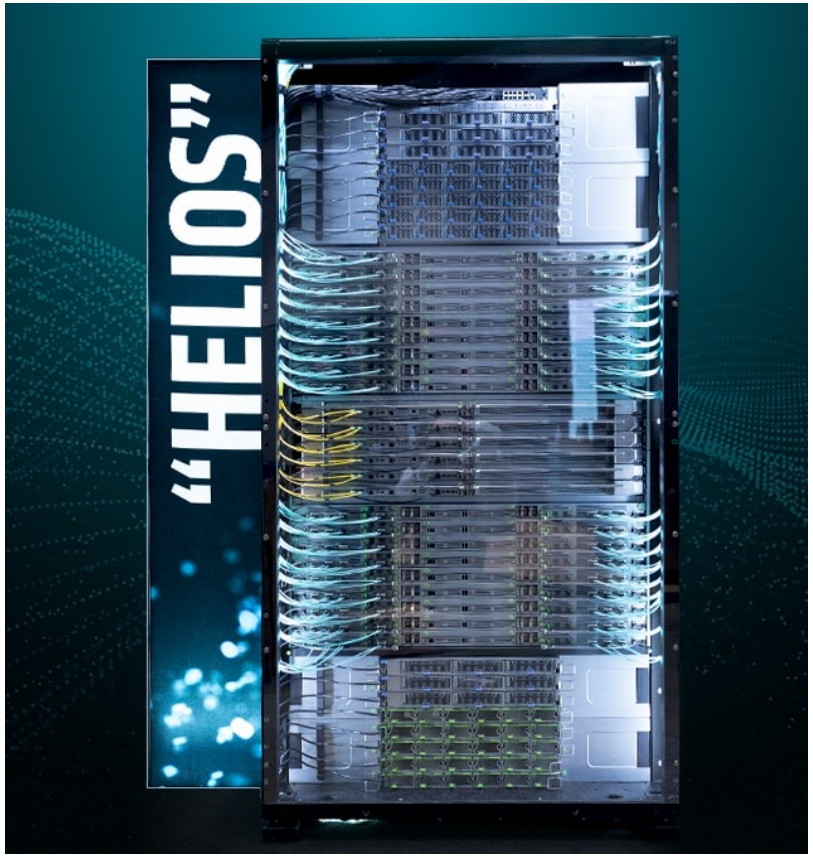
目标：超大容量、超大存储的“一台智能电脑”

NVL72



NVL576

Huawei CloudMatrix384



AMD Helios

提纲

- 人工智能计算需求
- 智能芯片基础架构与指标
- 智能芯片架构演进
- **总结**

本次课总结

章节一 · 神经网络的计算本质

- AI 计算核心为大规模矩阵乘法
- 大模型参数量：亿→万亿，算力需求呈指数增长

章节二 · 芯片性能的评估框架

- 五大指标：算力、带宽、显存、能效比、互联
- Nvidia GPU 以 HBM3+NVLink 组合确立全球领先地位

章节三 · 架构演进的设计哲学

- CPU → GPU → DSA：专用化程度换取效率提升
- TPU 脉动阵列：为矩阵乘法而生的极致优化
- 存算一体：从根本上突破冯诺依曼瓶颈（新兴）

算力竞争的本质是人才 + 资本 + 时间的高度整合。新兴架构+ 软件生态建设 + 细分场景应用落地为算力竞争提供了不同于主流GPU赛道的新的机会。

谢谢！