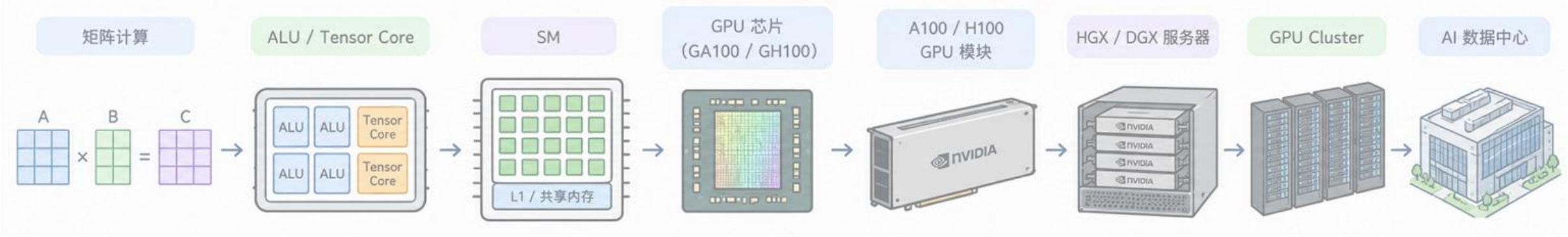
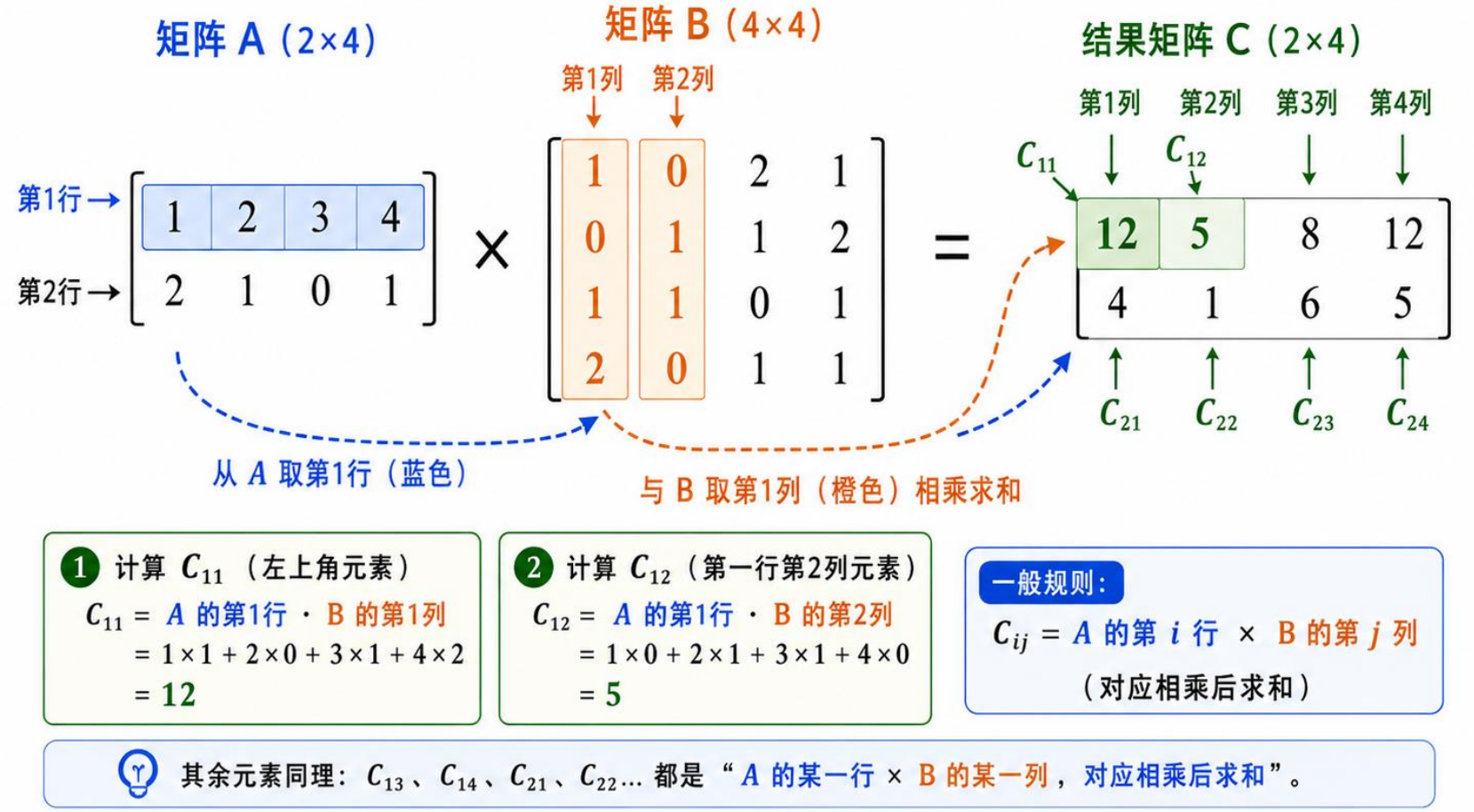
算力基础知识

以NVIDIA为例

湖北社保科创基金管理中心张泽锟

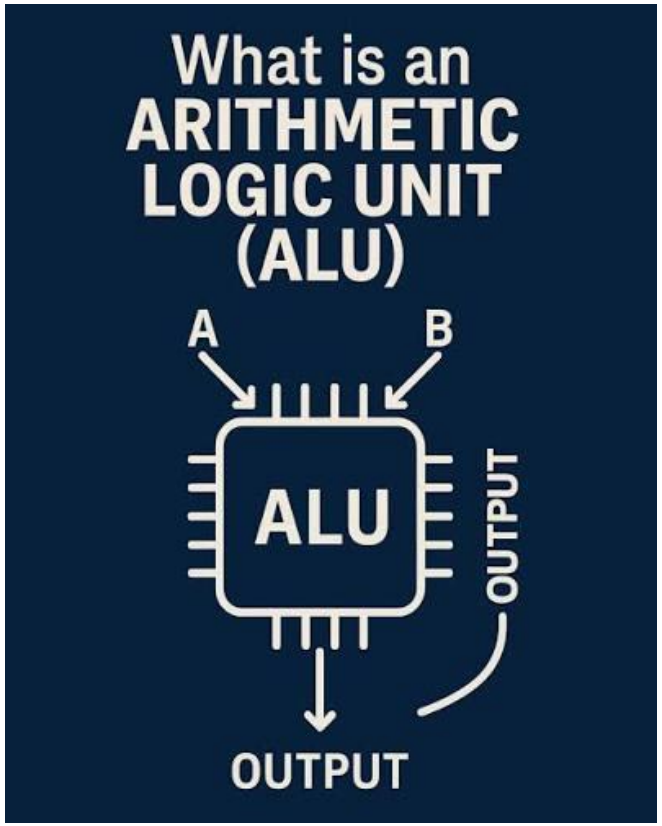


从线性代数说起



首先回忆一下线性代数里基础的矩阵乘法：A×B=C，计算过程由AI生成

ALU最小计算单元



矩阵 A（2×4）

1	2	3	4
2	1	0	1

×

矩阵 B（4×4）

1	0	2	1
0	1	1	2
1	1	0	1
2	0	1	1

=

结果矩阵 C（2×4）

12	5	8	12
4	1	6	5

2 C11 的 4 个乘加步骤（示例）

C11 = A第1行 · B第1列 = 1×1 + 2×0 + 3×1 + 4×2 = 12

周期	操作（acc 初值为 0）	acc 结果
周期 1	acc = 0 + 1×1	1
周期 2	acc = 1 + 2×0	1
周期 3	acc = 1 + 3×1	4
周期 4	acc = 4 + 4×2	12

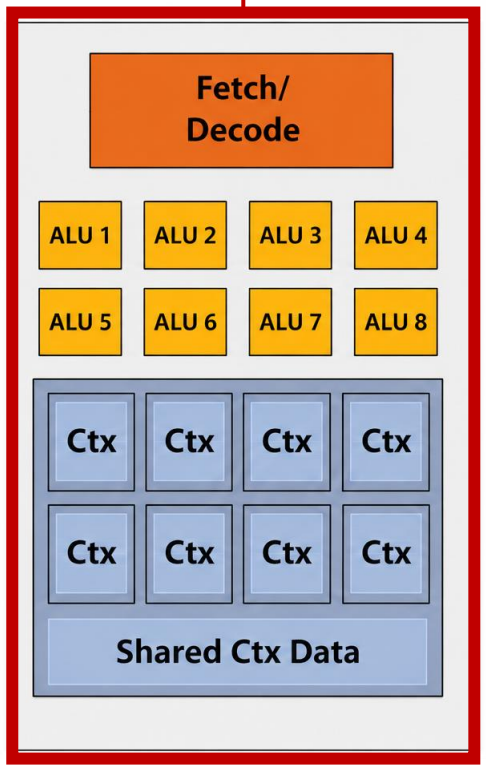
1 8 个输出元素，依次串行计算（每个输出需要 4 个周期）

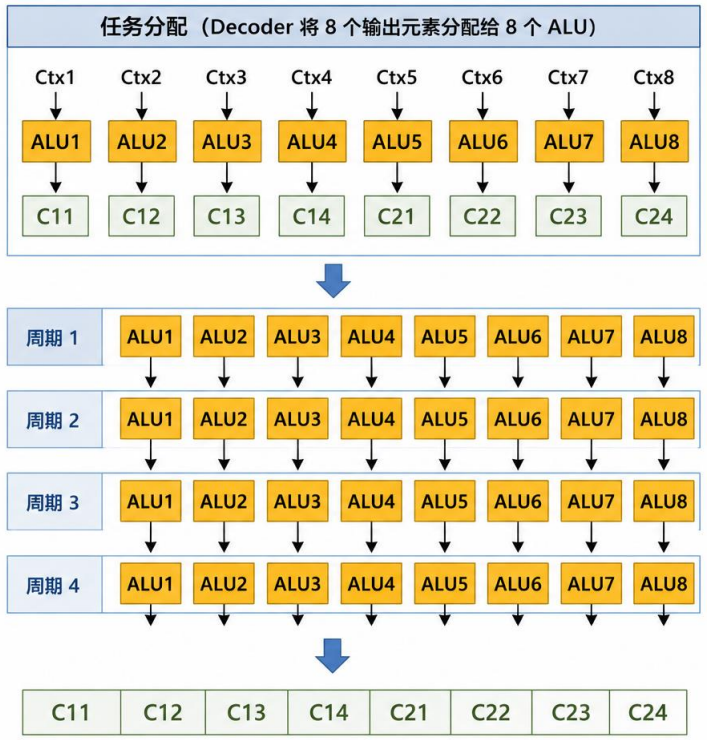
C11	C12	C13	C14	C21	C22	C23	C24
周期 1	周期 5	周期 9	周期 13	周期 17	周期 21	周期 25	周期 29
周期 2	周期 6	周期 10	周期 14	周期 18	周期 22	周期 26	周期 30
周期 3	周期 7	周期 11	周期 15	周期 19	周期 23	周期 27	周期 31
周期 4	周期 8	周期 12	周期 16	周期 20	周期 24	周期 28	周期 32

共 8 个输出
每个 4 个周期
总计
8 × 4 = 32
个周期

关注C_{11}的的计算过程：乘加运算（MAC，Multiply-Accumulate），AI生成
MAC过程通常由GPU的一个计算单元ALU（Arithmetic Logic Unit）来完成，需要 8×4=32 个计算周期

基础并行计算





矩阵计算过程由GPU的8个计算单元ALU并行计算

ALU1 → C_{11}

ALU2 → C_{12}

……

ALU8 → C_{24}

每一个ALU需要4个周期完成计算

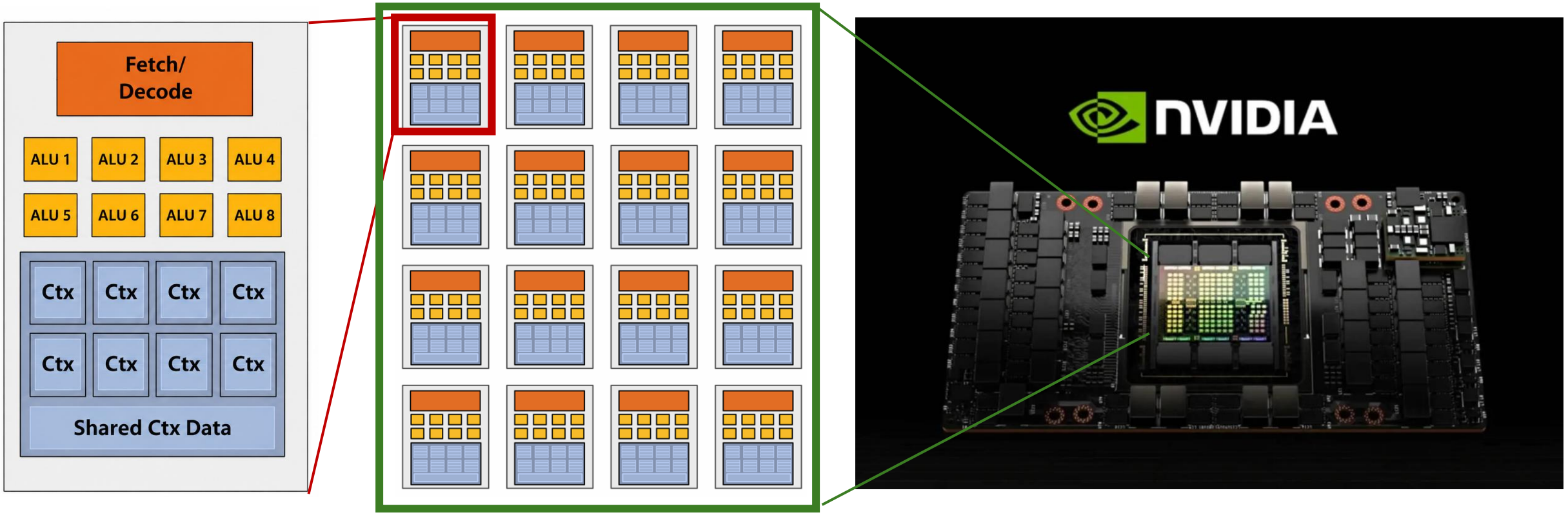
例如$C_{11}=1\times1+2\times0+3\times1+4\times2$

最终并行计算的总时间为4个周期

1核 Core
NVIDA：SM（Streaming Multiprocessor）
AMD：CU（Compute Unit）

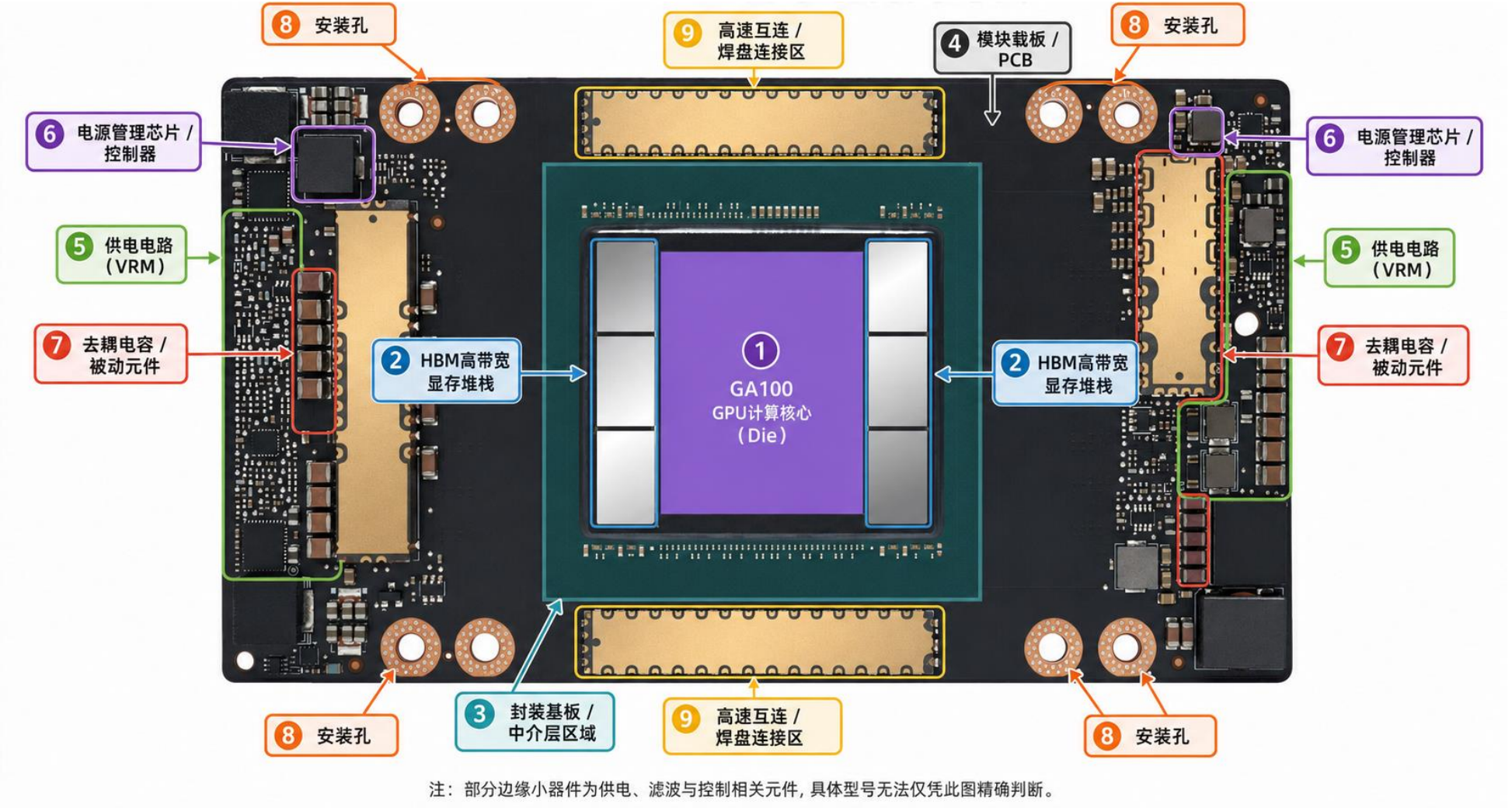
1FLOPS = ALU的一次浮点计算（加、乘）
1 TFLOPS = 10^{12}FLOPS = 每秒 1 万亿次浮点运算
1 TFLOPS的芯片理论上一秒钟可以完成大约**156亿次矩阵乘法运算**

GPU芯片



多核Core组织在一起计算就构成了GPU芯片
例：英伟达A100 GPU芯片简易示例（A100有128个SM）

模块级芯片A100



注：部分边缘小器件为供电、滤波与控制相关元件，具体型号无法仅凭此图精确判断。

核心元件：

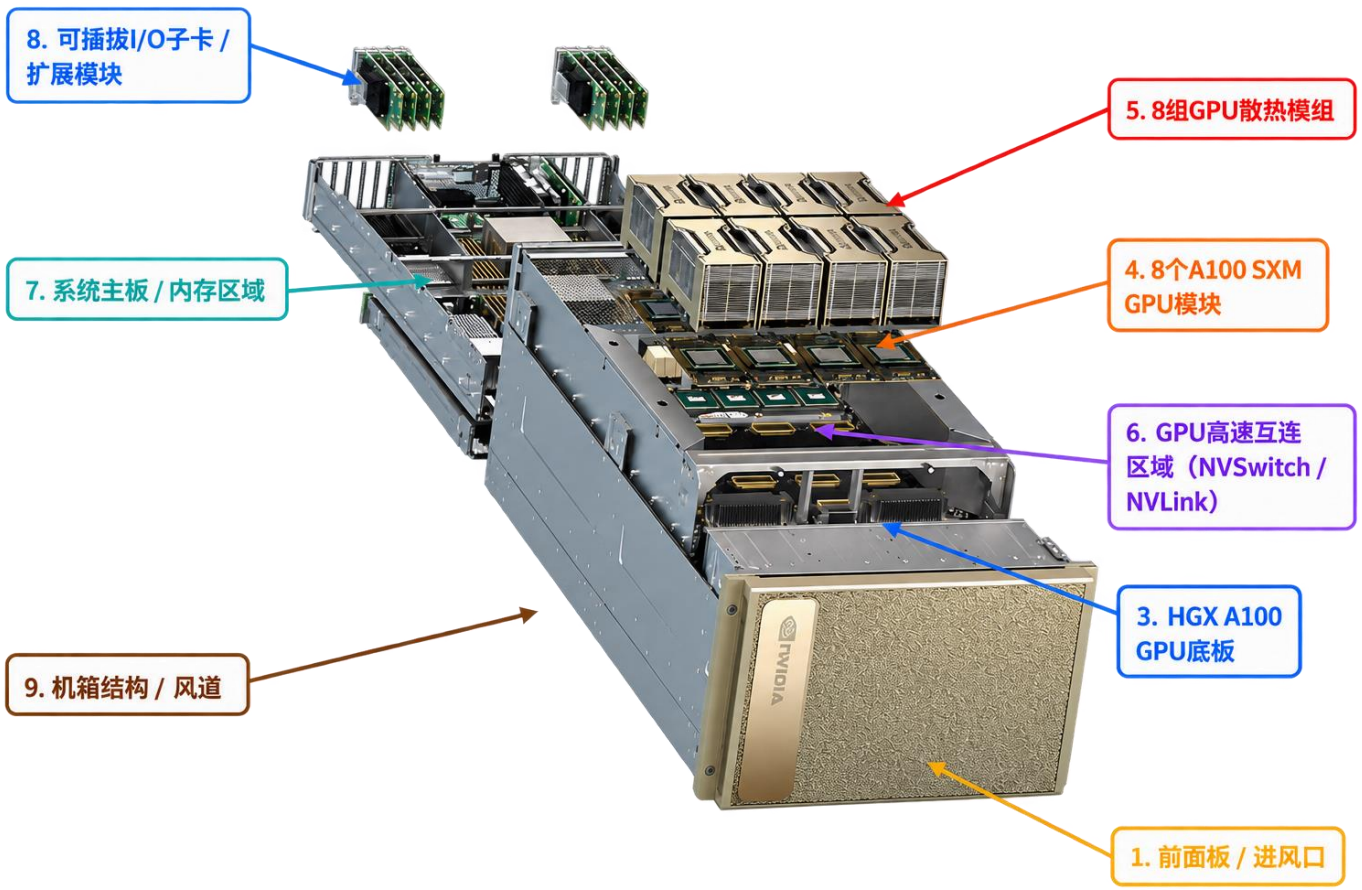
1 GA100
整张计算卡的数据处理核心GPU

2 HBM显存
High Bandwidth Memory：堆叠式动态随机存取存储器（DRAM）

4 PCB
Printed Circuit Board：承载上述所有电子元器件的底层平台

NVIDIA A100 主要可见部件示意图 AI生成

系统级DGX服务器



核心元件：

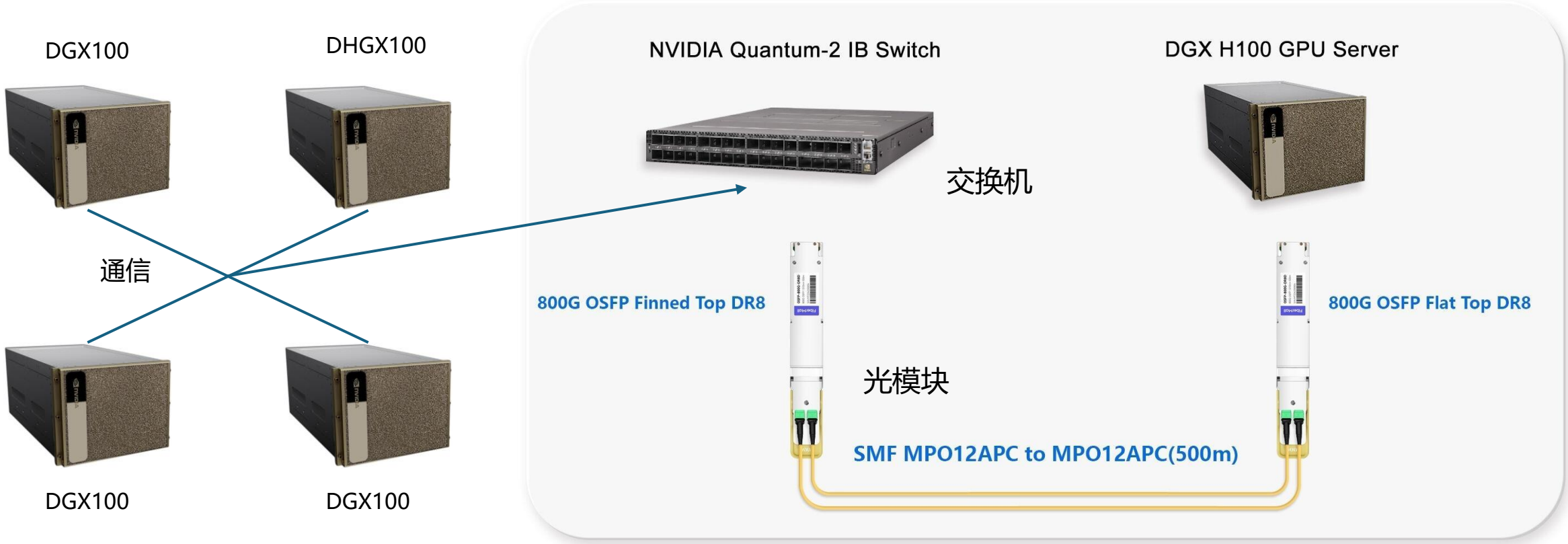
1 A100
GA100 Die、HBM、SM、CUDA Core、Tensor Core

6 NVSwitch/Link
负责 GPU 之间的高速数据传输和协同计算。

7 CPU
负责操作系统、任务调度、数据预处理、I/O 管理，并把任务喂给 GPU。

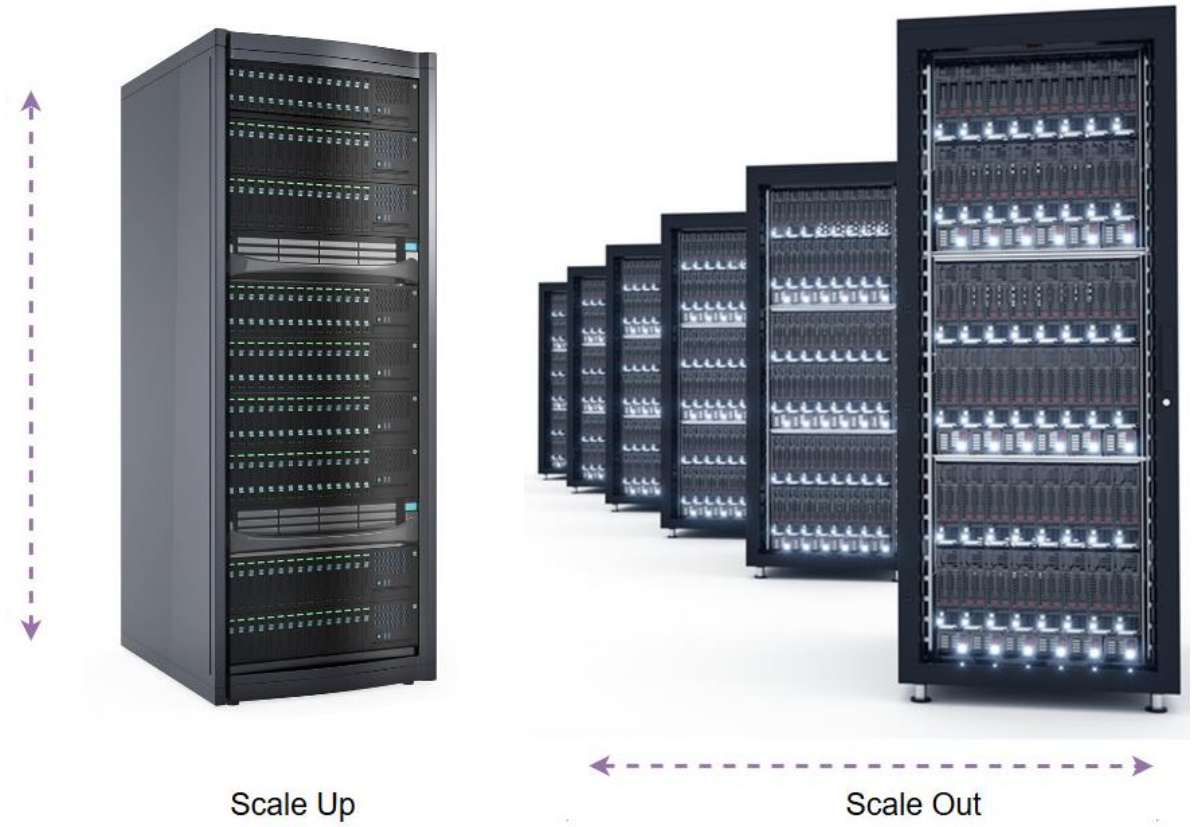
NVIDIA DGX主要可见部件示意图 AI生成

集群级光互联



数据高速传输是目前最大的计算瓶颈之一

集群级数据中心



- **高速网络**：网卡、DPU、交换机，负责服务器间高速通信。
- **供配电**：PSU、PDU、UPS 等，保障高功率集群稳定供电。
- **存储系统**：NVMe SSD、分布式存储，持续向 GPU 提供训练数据。
- **机柜基础设施**：机柜、背板、连接器、线缆，支撑高密度部署。
- **集群管理软件**：负责 GPU 调度、任务编排和故障管理。
- **CPU / DPU / SmartNIC**：承担数据处理、网络卸载和系统控制。