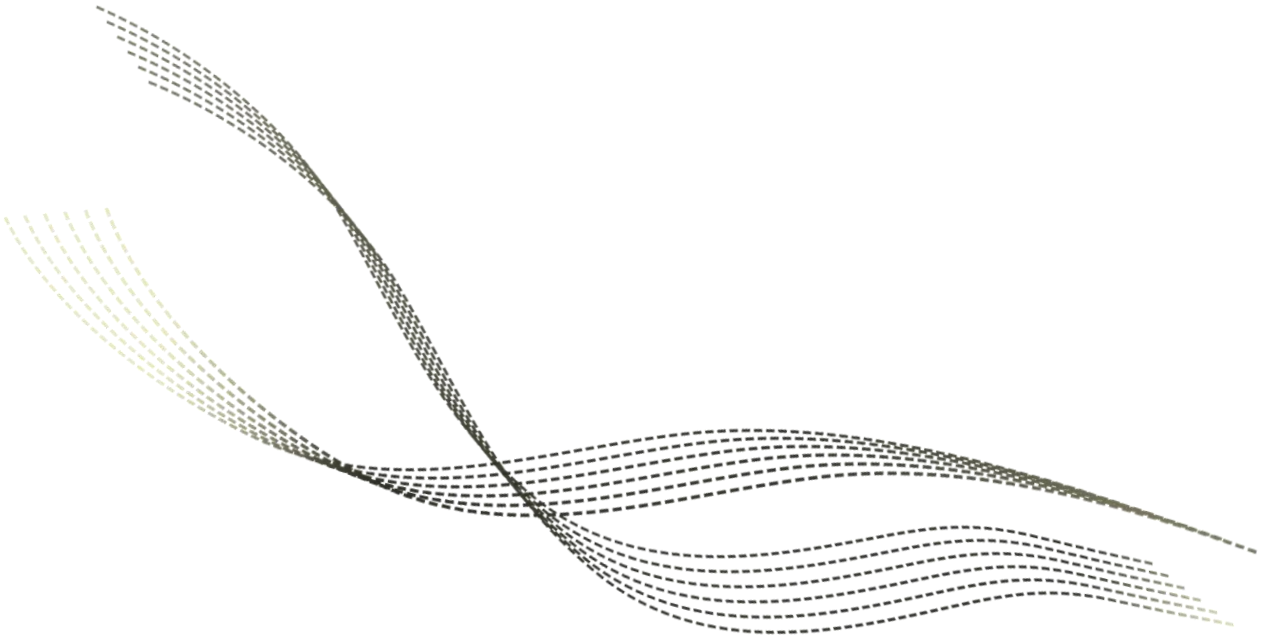
高级计算机体系结构

Advanced Computer Architecture



集成电路现状

沈明华

目录

CONTENTS

01 芯片的分类

02 CPU和GPU

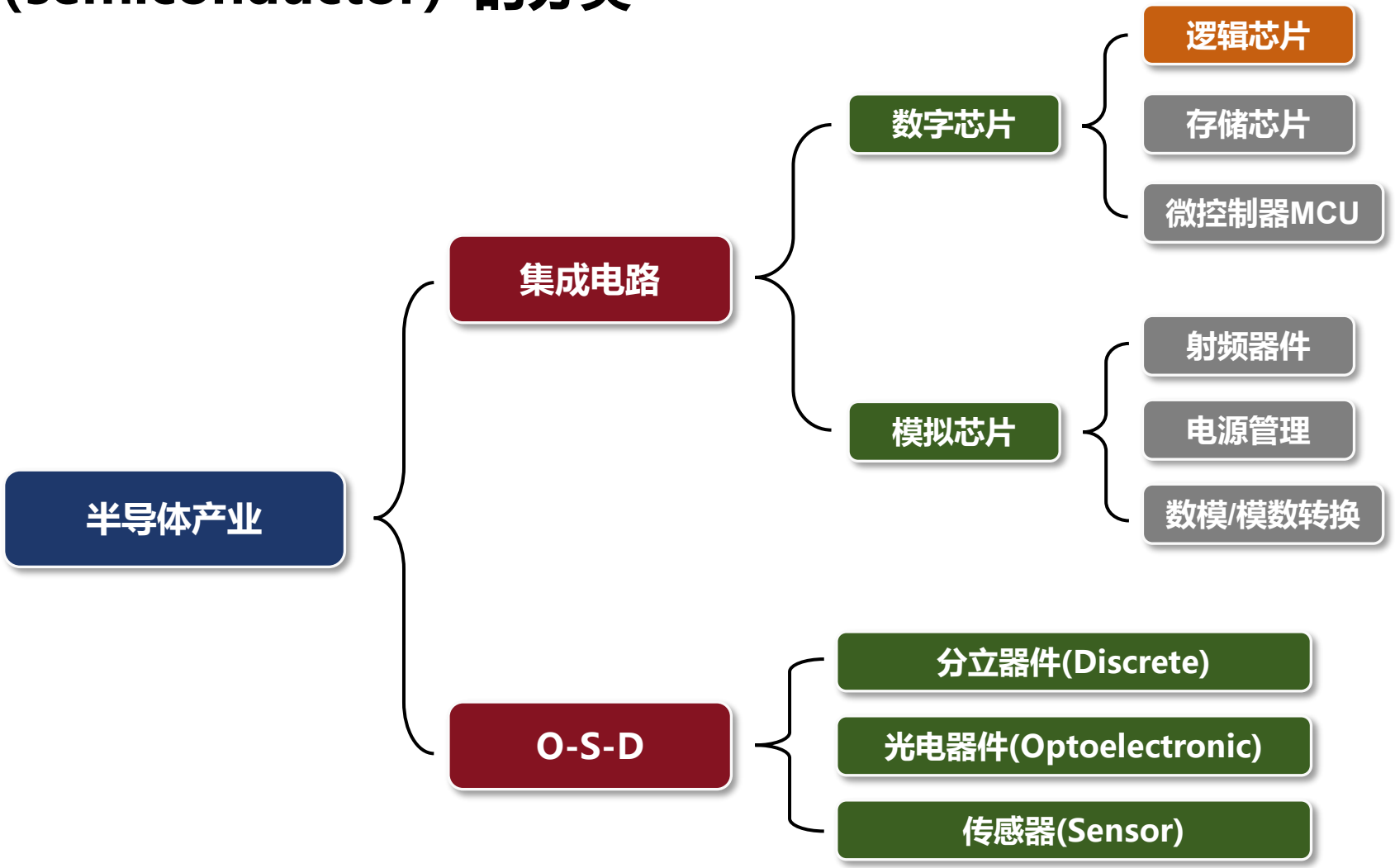
03 ASIC和FPGA

04 总结对比

PART 01 芯片的分类

■ 芯片的分类

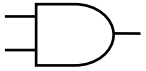
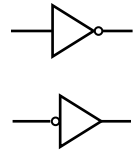
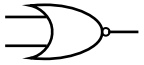
□ 半导体（semiconductor）的分类



■ 芯片的分类

□ 逻辑芯片（Logic Chip）

- 逻辑芯片是一类用于执行特定逻辑运算的集成电路，是现代电子系统的核心组件之一。
- 逻辑芯片利用晶体管来构建各种逻辑门电路（例如：与门AND、或门OR、非门NOT、异或门XOR等），进而组成更为复杂的电路，实现不同类别的逻辑运算。
- 逻辑芯片能够实现数据处理、控制和其他各种功能，广泛应用于消费电子、工业制造、教育医疗、国防军事等各个领域。

名称	图形符号
与门	
或门	
非门	
与非门	
或非门	

■ 芯片的分类

□ 逻辑芯片的分类

– 根据功能和用途的不同，逻辑芯片可以分为以下几大类别：

PART 02

CPU和GPU

■ CPU和GPU

□ CPU的定义

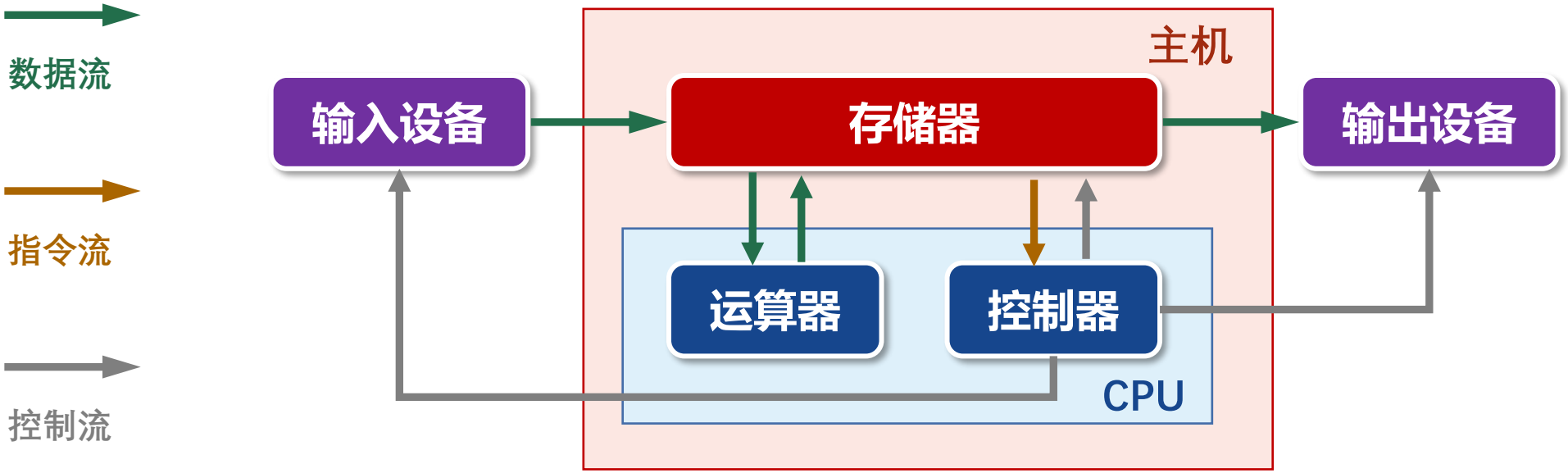
- CPU，就是Central Processing Unit，中央处理器。
- CPU是计算机系统的运算和控制核心，是信息处理、程序运行的最终执行单元。
- CPU的主要作用是解释计算机指令以及处理计算机软件中的数据。
- CPU是计算机中负责读取指令，对指令译码并执行指令的核心部件。

■ CPU和GPU

□ 冯•诺依曼架构

– 现代计算机，都是基于1940年代诞生的冯·诺依曼架构。

– 在这个架构中，包括了运算器（也叫逻辑运算单元，ALU）、控制器（CU）、存储器、输入设备、输出设备等组成部分。运算器和控制器这两个核心功能，都由CPU负责承担。



■ CPU和GPU

□ 冯·诺依曼架构

– 处理流程：数据先存在存储器。然后，控制器会从存储器拿到相应数据，再交给运算器进行运算。运算完成后，再把结果返回到存储器。

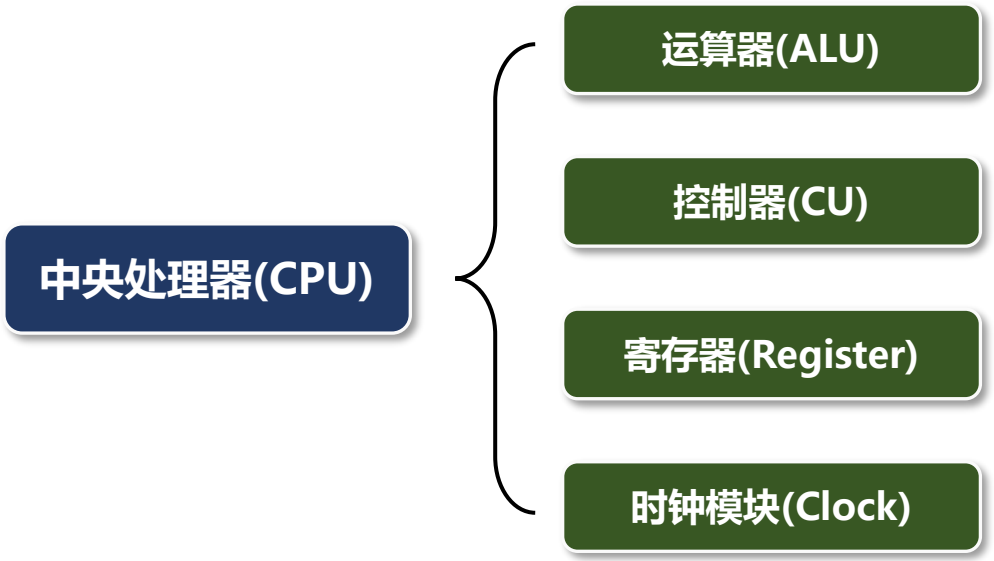
■ CPU和GPU

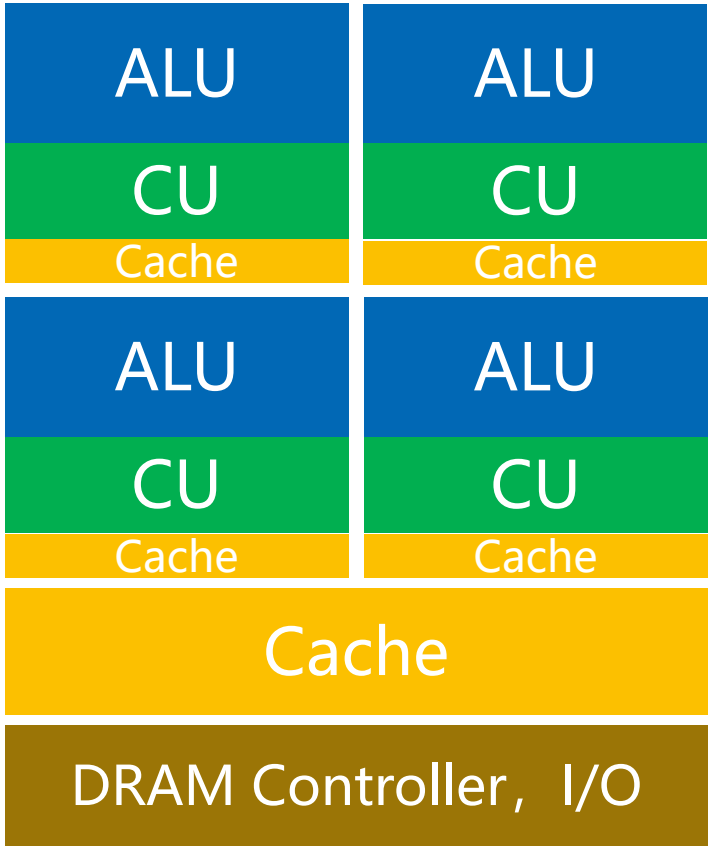
□ CPU的主要组成

- **运算器（也叫算术逻辑单元Arithmetic Logic Unit，ALU）**：包括加法器、减法器、乘法器、除法器等，负责执行所有数学计算和逻辑判断。
- **控制器（控制单元Control Unit）**：负责协调整个CPU的操作，包括取指、解码、执行和写回四个阶段。负责从内存中读取指令、解码指令、执行指令。还负责生成各种控制信号来指导其他硬件组件的工作。
- **寄存器（高速缓存）**：是CPU中的高速存储器，存储最近使用过的数据或即将使用的指令。通常分为L1、L2和L3三级缓存。它的CPU与内存（RAM）之间的“缓冲”，速度比一般的内存更快，避免内存“拖累”CPU的工作。
- **时钟模块**：负责管理CPU的时间，为CPU提供稳定的时基。它通过周期性地发出信号，驱动CPU中的所有操作，调度各个模块的工作。

■ CPU和GPU

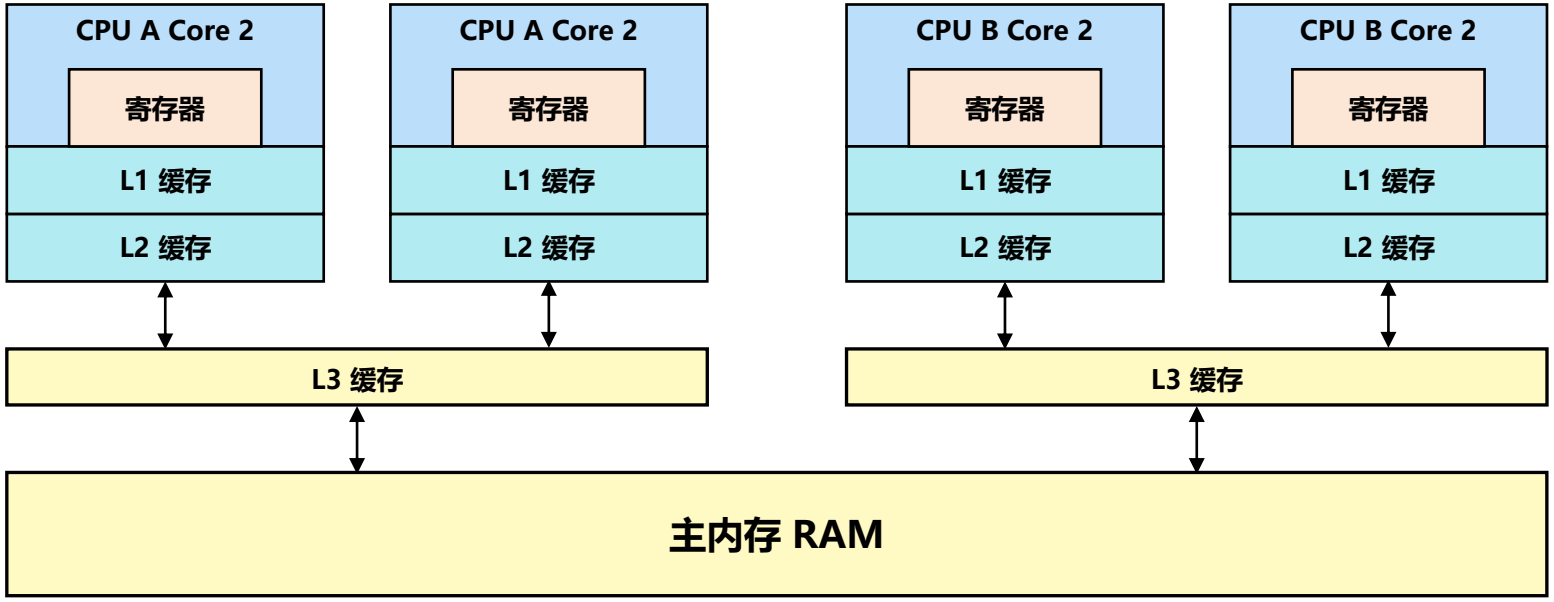
□ CPU的主要组成





■ CPU和GPU

□ 多核CPU



CPU多核硬件架构示例

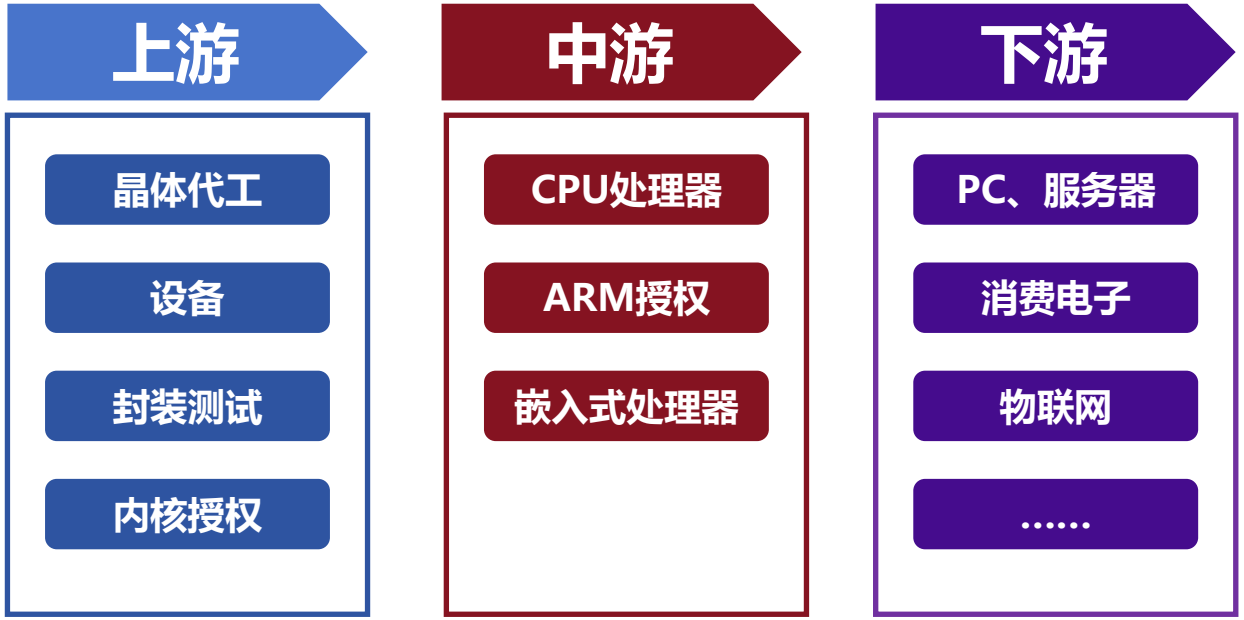
■ CPU和GPU

□ CPU的类别（按指令集）

指令集名称	类型	推出时间	推出公司/机构	采用此指令集主要授权商
x86	CISC 复杂指令集	1978年	美国Intel、美国AMD	兆芯、众志、海光等
ARM	RISC 精简指令集	1985年	英国Arm (被日本软银公司收购)	苹果、三星、AMD、TI、东芝、微芯、高通、联发科、展讯、飞腾、海思、瑞芯微、晶晨、全志等
MIPS (终止更新)		1980年代	美国MIPS	瑞昱、炬力等
SPARC (终止更新)		1985年	美国SUN (被甲骨文收购)	德州仪器、Cypress(被Infineon收购)、富士通等
PowerPC (被迫开源)		1991年	美国IBM	曾用：苹果、任天堂、微软、索尼、中晟
Alpha (终止更新)		1992年	美国DEC (被惠普并购)	申威
RISC-V (开源)		2010年	美国加州大学伯克利分校 (RISC-V基金会运营)	GreenWaves、Imagination、平头哥、晶心科技、芯源股份、中天微、睿思芯科、香山处理器

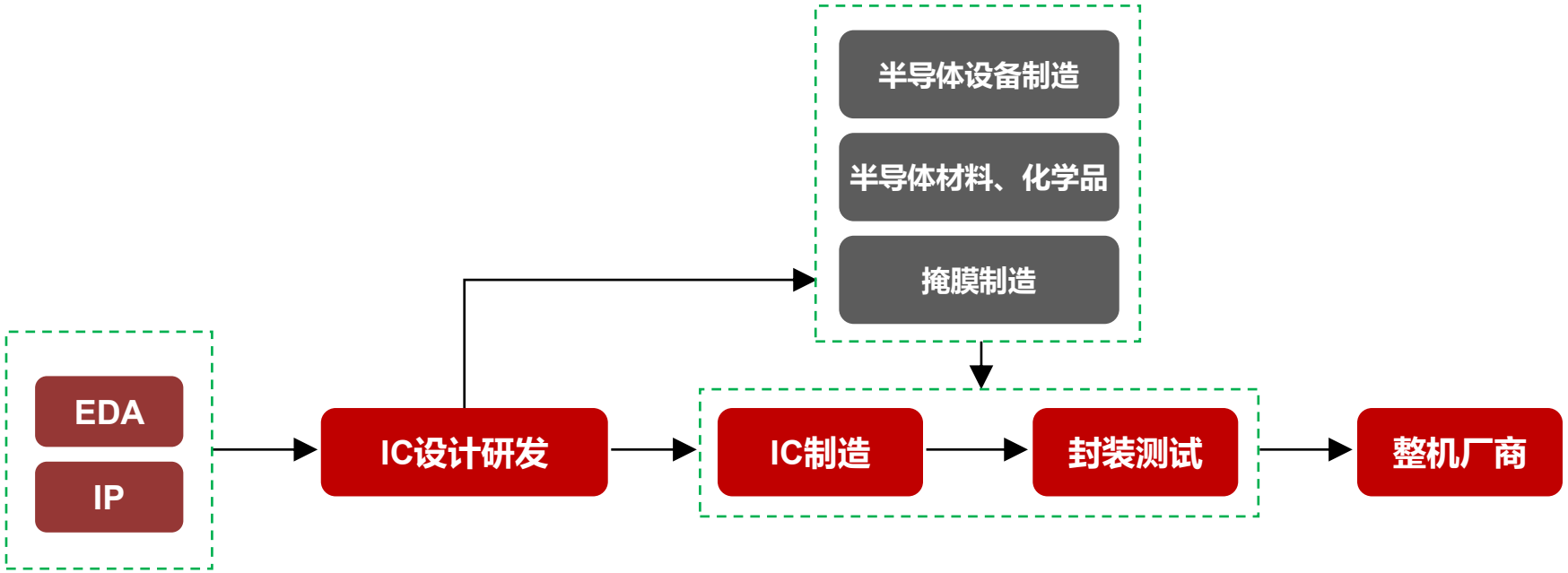
■ CPU和GPU

□ 半导体产业链



■ CPU和GPU

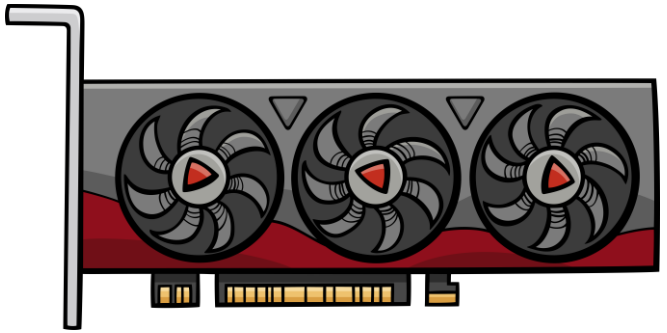
□ 半导体产业链



■ CPU和GPU

□ GPU的定义

- GPU，Graphics Processing Unit，图形处理单元。
- GPU最初是为了加速计算机图形渲染而设计的专用处理器。它能够高效地执行大量并行计算任务，在3D图形渲染、视频编码解码、科学计算等领域表现出色。
- 随着技术的发展，GPU的应用范围已经远远超出了传统的图形处理，成为通用计算的重要组成部分。



■ CPU和GPU

□ GPU的特点

– **高度并行架构**

GPU拥有成百上千个简单的处理核心，可以同时处理多个数据流，具有强大的并行计算能力。

– **专为浮点运算优化**

由于图形渲染涉及大量的矩阵乘法和向量运算，因此GPU在浮点运算方面进行了特别优化，提供了比CPU更高的吞吐量。

– **内存带宽高**

GPU通常配备有高速的专用显存（VRAM），其带宽远高于普通系统内存，这有助于快速读取和写入大量图像数据。

– **低延迟响应**

在图形渲染过程中，GPU需要即时生成每一帧画面，因此它被设计成能够在极短的时间内完成复杂的计算任务。

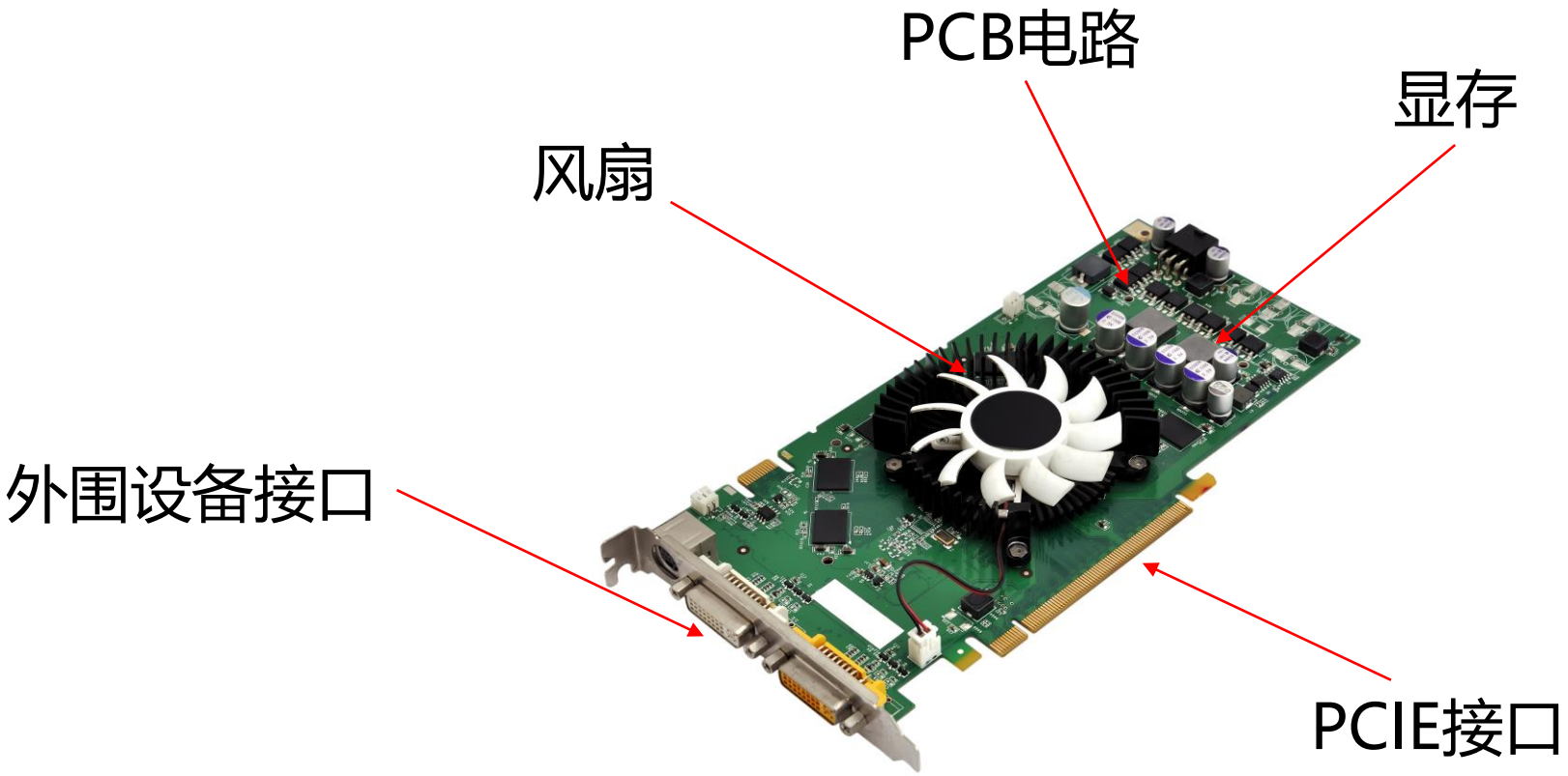
– **支持多种编程接口**

现代GPU不仅限于使用OpenGL、DirectX等图形API，还广泛支持CUDA、OpenCL、Vulkan等通用计算API。

■ CPU和GPU

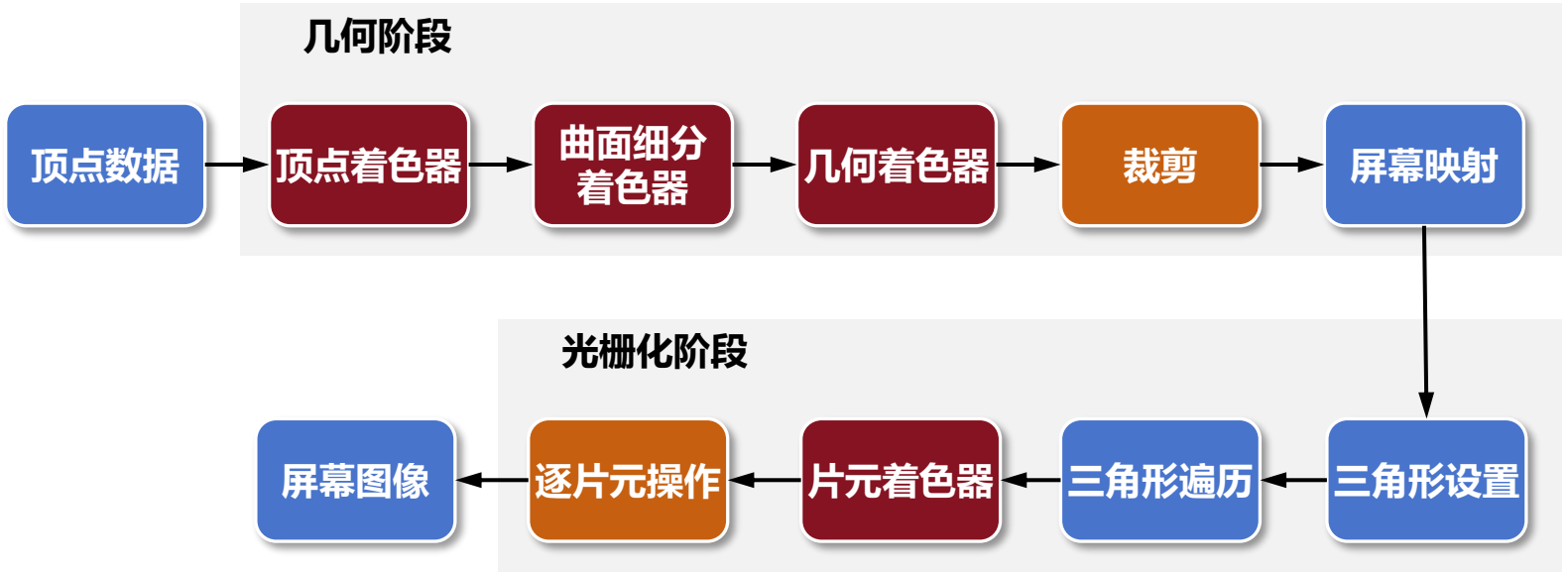
□ 显卡的组成

– 显卡除了GPU之外，还包括显存、VRM稳压模块、MRAM芯片、总线、风扇、外围设备接口等。



■ CPU和GPU

□ 图形渲染的流程



■ CPU和GPU

□ 显卡产业的发展历程

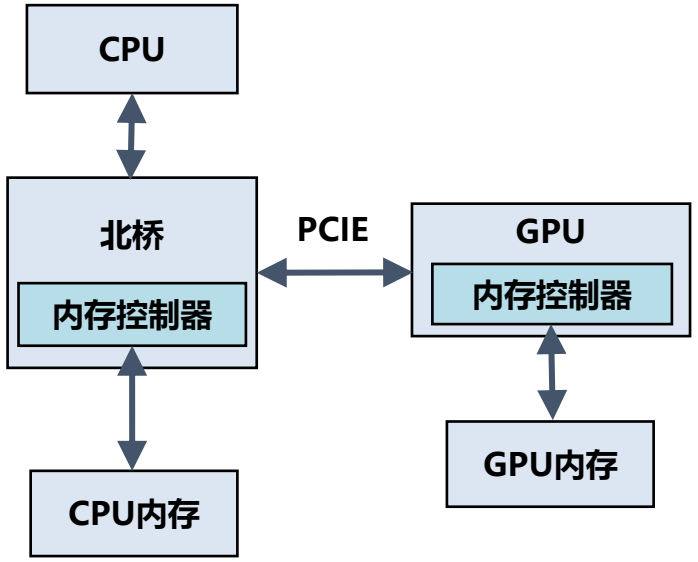
- 1962年，麻省理工学院博士伊凡·苏泽兰（Ivan Sutherland）奠定了计算机图形学基础。
- 1984年，SGI公司推出了面向专业领域的高端图形工作站，俗称图形加速器，是首个专门的图形处理硬件。
- 1994年，3DLabs发布GLINT300SX，是PC最早的3D硬件加速图形芯片，从此开启3D显卡时代。
- 1995年，3Dfx发布Voodoo图形加速卡，是真正意义第一款消费级3D显卡。
- 1999年8月，NVIDIA（英伟达）公司发布图形芯片Geforce256，首次提出GPU的概念。
- 1999年，NVIDIA崛起，击败并收购3Dfx。
- 2006年，AMD以54亿美元收购ATI。

■ CPU和GPU

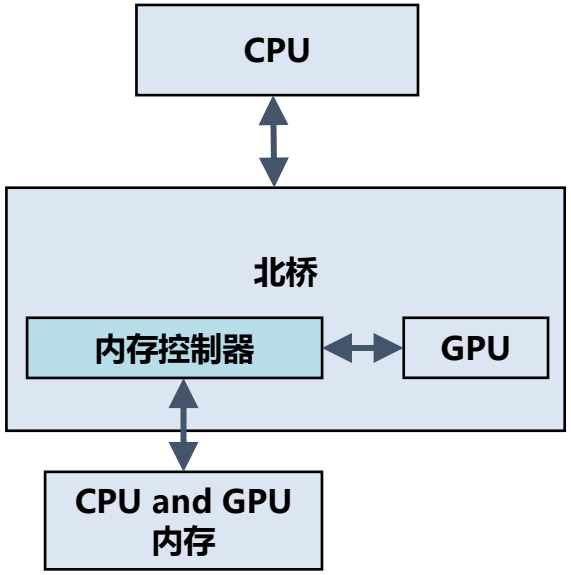
□ GPU（显卡）的分类

– 独立GPU（dGPU，discrete/dedicated GPU），常说的独立显卡（独显）。

– 集成GPU（iGPU，integrated GPU），常说的集成显卡（集显）。



独立GPU



集成GPU

■ CPU和GPU

□ GPU（显卡）的主要参数

- 制程：GPU的制造工艺和设计规则，代表不同电路特性，通常以生产精度nm表示
- 图形处理器单元数量：包含了光栅单元ROP，纹理单元TMU的数量，数量越多可执行指令越多
- CUDA核数：CUDA是执行函数的重要部件，CUDA核数越多，性能运行越好
- Tensor核数：指张量处理单元的数量，Tensor Core核数越多，性能越好
- 核心频率：指显示核心的工作频率，能反映显示核心的性能优良
- 显存容量：显存容量越大，GPU能够处理的数据量越大
- 显存位宽：指显存在单位时钟周期内所传送数据的位数，位数越大瞬间传送数据量越大
- 显存带宽：等于显存频率X显存位宽/8，与显存频率、位宽成正比
- 显存频率：反映显存速度，以MHz为衡量单位，越高端的显存，频率越高

■ CPU和GPU

□ 英伟达消费级显卡经典型号

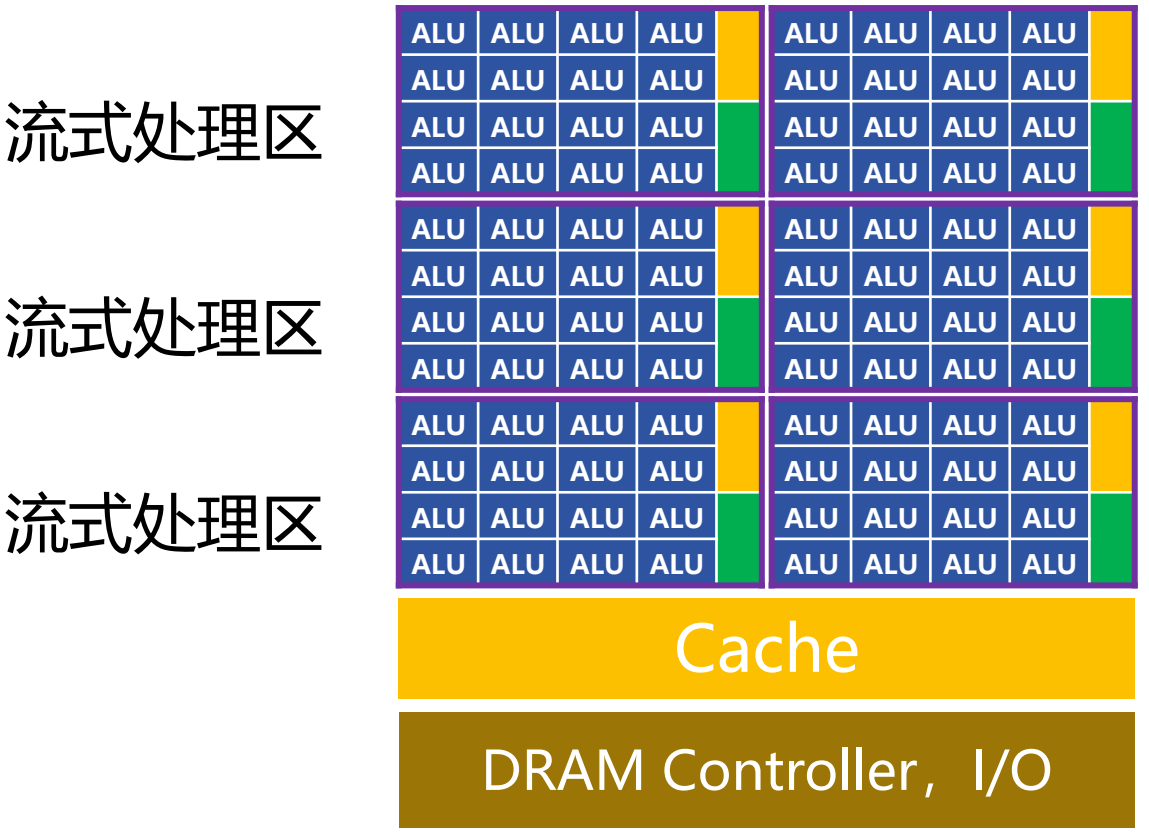
时间	发布型号	制程
1995	STG-2000X	500nm
1998	RIVA 128	350nm
1999	Riva TNT2	250nm
1999	GeForce 256	220nm
2001	GeForce 3	180nm
2002	GeForce 4 Ti 4200	150nm
2004	GeForce 6800	130nm
2006	GeForce 8800 GTX	90nm
2010	GeForce GTX 480	40nm
2013	GeForce GTX Titan	28nm
2014	GeForce GTX 970	28nm
2016	GeForce GTX 1080	16nm
2018	GeForce RTX 2080	12nm
2020	GeForce RTX 3090	三星 8nm
2022	GeForce RTX 40系列	台积电 5nm
2025	GeForce RTX 50系列	台积电 3nm

■ CPU和GPU

□ GPU（显卡）的组成

– GPU的核，称为流式多处理器（Stream Multi-processor，SM），是一个独立的任务处理单元。

– 在整个GPU中，会划分为多个流式处理区。每个处理区，包含数百个内核。



■ CPU和GPU

□ GPGPU

- GPGPU，General Purpose computing on GPU，基于GPU的通用计算。
- GPGPU利用GPU的计算能力，在非图形处理领域进行更通用、更广泛的科学计算。
- GPGPU在传统GPU的基础上，进行了进一步的优化设计，使之更适合高性能并行计算。

	主要执行任务	功能	国内主要公司
GPU	图形渲染	图形渲染、图形计算	景嘉微、摩尔线程、象帝先、芯动科技、格兰菲、励算、深流微、芯瞳、绘智微
GPGPU	并行计算	AI相关计算，科学计算和通用计算	壁仞、沐曦、登临、天数智芯、红山微电子、瀚博

■ CPU和GPU

□ 英伟达GPU架构演进

架构代号	中文代号	年代	工艺制程	晶体管数量	代表型号
Tesla	特斯拉	2008	90nm	约6.84亿	G80
Fermi	费米	2010	40/28nm	30亿	Quadro 7000
Kepler	开普勒	2012	28nm	71亿	K80、K40M
Maxwell	麦克斯韦	2014	28nm	80亿	M5000、M4000
Pascal	帕斯卡	2016	16nm	153亿	P100、GTX1080、P6000
Volta	伏特	2017	12nm	211亿	V100、TiTan V
Turing	图灵	2018	12nm	186亿	T4、2080TI、RTX 5000
Ampere	安培	2020	7nm	283亿	A100、A30、3090
Hopper	赫柏	2022	5nm	800亿	H100
Blackwell	布莱克威尔	2024	5nm	2080亿	B200、B100

■ CPU和GPU

□ 英伟达GPU架构演进

- 2007年，Tesla架构：是第一代真正用于并行运算的GPU架构，标志用于计算的GPU产品线正式独立。
- 2010年，Fermi架构：首个完整GPU架构，是第一个可支持与共享存储结合纯cache层次的GPU架构。
- 2012年，Kepler架构：首次在GPU中引入了动态并行技术。
- 2014年，Maxwell架构：可解决视觉计算领域中最复杂的光照和图形难题，优化功耗。
- 2016年，Pascal架构：采用了HBM2的CoWoS技术。首次引入了3D内存及NVLink高速互联总线。
- 2017年，Volta架构：首次引入Tensor（张量）运算单元。
- 2018年，Turing架构：架构最大的变革，引入了RTX追光技术总线。
- 2020年，Ampere架构：包含540亿个晶体管，大幅提升了人工智能和高效能运算。
- 2022年，Hopper架构：第一个真正的异构加速平台，适用于高性能计算（HPC）和 AI 工作负载。
- 2024年，Blackwell架构：专门用于处理数据中心规模的生成式 AI 工作流，能效是 Hopper 的 25 倍。

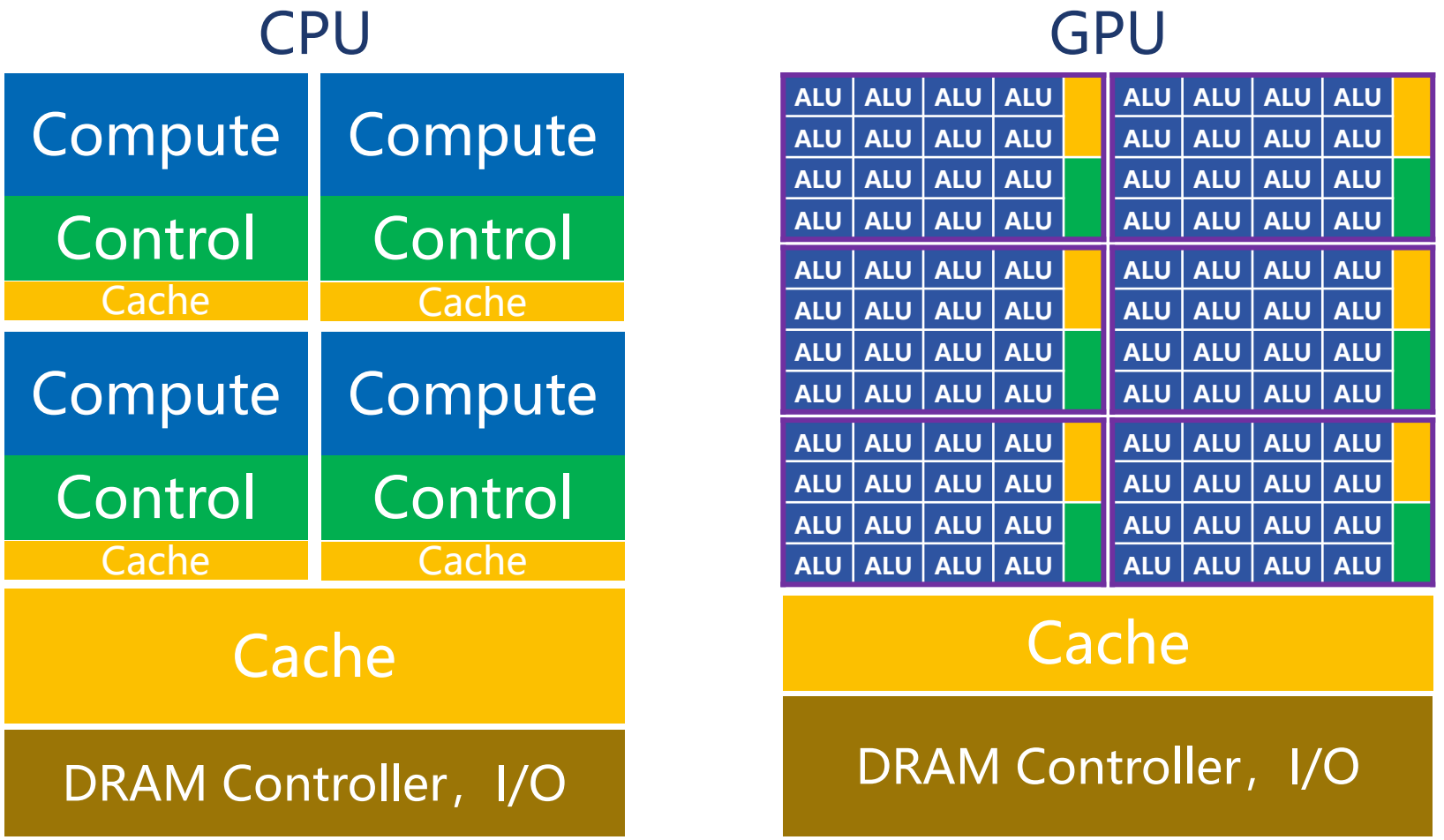
■ CPU和GPU

□ 部分国产GPU设计厂商及产品

厂商名称	代表型号
景嘉微电子	JM5系列、JM7系列和JM9系列
壁仞科技	BR100
摩尔线程	MTT S60、MTT S80、MTT S3000
燧原科技	邃思2.0
寒武纪-U	思元220、思元290、思元370等
沐曦集成电路	MXN系列GPU(曦思)、MXC系列GPU(曦云)、MXG系列GPU(曦彩)
昆仑芯	昆仑芯2代芯片
芯动科技	风华1号、风华2号

■ CPU和GPU

□ CPU和GPU的区别



■ CPU和GPU

□ CPU和GPU的区别

- CPU的内核（包括了ALU）数量比较少，最多只有几十个。但是，CPU有大量的缓存（Cache）和复杂的控制器（CU）。
- CPU是一个通用处理器。作为计算机的主核心，它的任务非常复杂，既要应对不同类型的数据计算，还要响应人机交互。复杂的条件和分支，还有任务之间的同步协调，会带来大量的分支跳转和中断处理工作。
- CPU需要更大的缓存，保存各种任务状态，以降低任务切换时的时延。它也需要更复杂的控制器，进行逻辑控制和调度。
- CPU的强项是管理和调度。真正干活的功能，反而不强（ALU占比大约5%~20%）。

■ CPU和GPU

□ CPU和GPU的区别

- GPU的内核数，远远超过CPU，可以达到几千个甚至上万个（也因此被称为“众核”）。
- GPU为图形处理而生，任务非常明确且单一。它要做的，就是图形渲染。图形是由海量像素点组成的，属于类型高度统一、相互无依赖的大规模数据。
- GPU的任务，是在最短的时间里，完成大量同质化数据的并行运算。所谓调度和协调的“杂活”，反而很少。
- GPU的控制器功能简单，缓存也比较少。它的ALU占比，可以达到80%以上。
- 虽然GPU单核的处理能力弱于CPU，但是数量庞大，非常适合高强度并行计算。同等晶体管规模条件下，它的算力，反而比CPU更强。

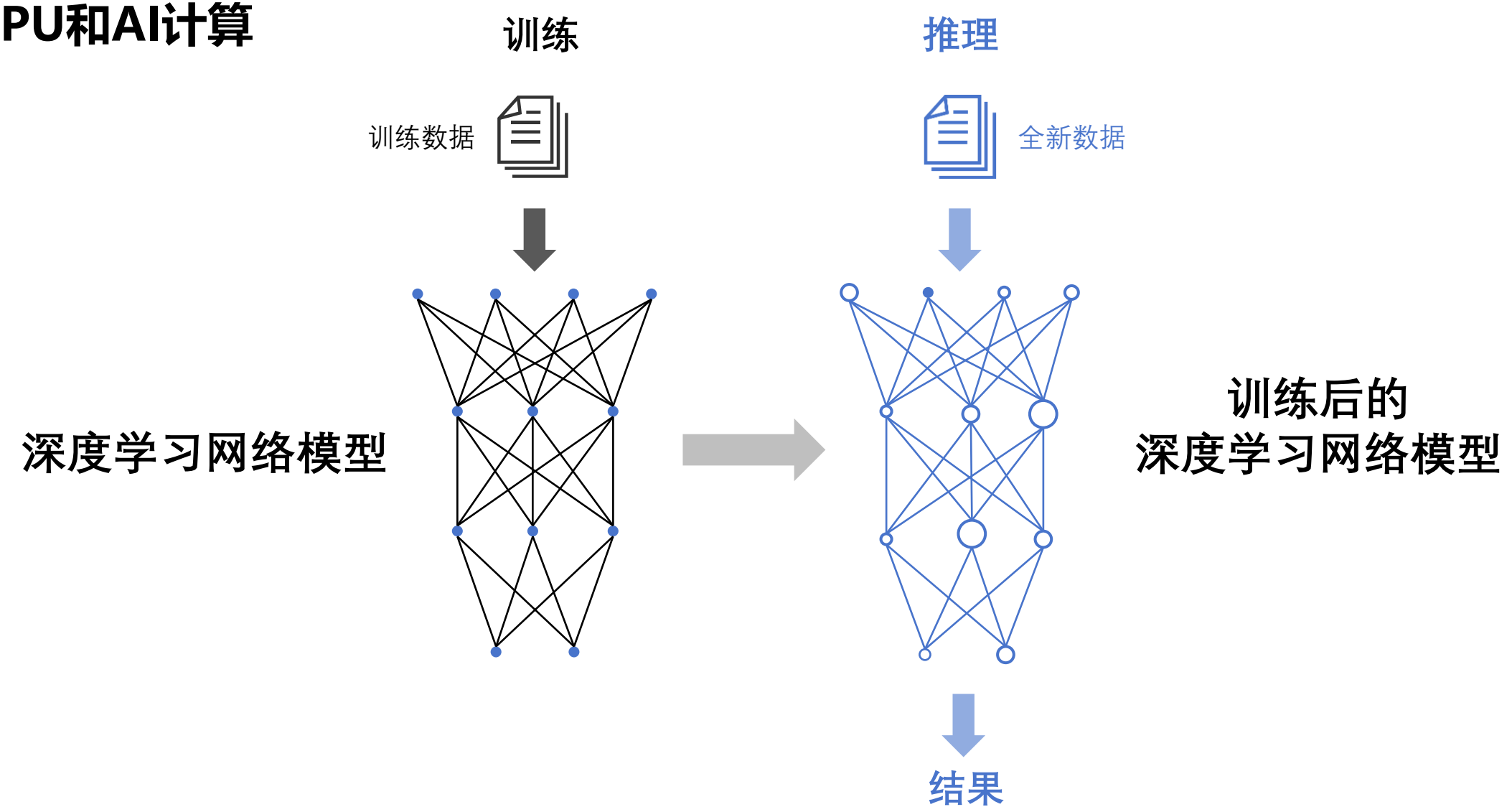
■ CPU和GPU

□ GPU和AI计算

- AI计算和图形计算一样，也包含了大量的高强度并行计算任务。
- 深度学习是目前最主流的人工智能算法，包括训练（training）和推理（inference）两个环节。
- 在训练环节，通过投喂大量的数据，训练出一个复杂的神经网络模型。在推理环节，利用训练好的模型，使用大量数据推理出各种结论。
- 它们所采用的具体算法，包括矩阵相乘、卷积、循环层、梯度运算等，分解为大量并行任务，可以有效缩短任务完成的时间。
- GPU凭借自身强悍的并行计算能力以及内存带宽，可以很好地应对训练和推理任务，已经成为业界在深度学习领域的首选解决方案。
- 目前，大部分企业的AI训练，采用的是英伟达的GPU集群。如果进行合理优化，一块GPU卡，可以提供相当于数十其至上百台CPU服务器的算力。

■ CPU和GPU

□ GPU和AI计算



■ CPU和GPU

□ 英伟达主流AI算卡参数

项目	A100	H100	L40S	H200
架构	Ampere	Hopper	AdaLovelace	Hopper
发布时间	2020	2022	2023	2024
FP16 TensorCore	312 TFLOP	756.5 TFLOPS	366.5 TFLOPS	1979 TFLOPS
INT8 TensorCore	624 TOPS	1513 TOPS	733 TOPS	3958 TOPS
FP64	9.7 TFLOPS	34 TFLOPS	25.7 TFLOPS	34 TFLOPS
FP32	19.5 TFLOPS	67 TFLOPS	91.6TFLOPS	67 TFLOPS
GPU内存	80 GB HBM2e	80 GB	48 GB GDDR6，带有ECC	141 GB HBM3e
GPU内存带宽	2.039 Tbps	3.35 Tbps	0.864 Tbps	4.8 Tbps
最高TDP	400 W	700 W	350 W	700 W
互联技术	NVLink:600GB/s PCIeGen4:64GB/s	NVLink:900GB/s PCIeGen5:128GB/s	PCIeGen4x16: 64GB/s bidirectional	NVLink:900GB/s PCIe Gen5:128GB/s

■ CPU和GPU

□ CUDA

- CUDA（Compute Unified Device Architecture）是英伟达推出的一种并行计算平台和编程模型。
- CUDA 通过提供一系列的工具、库和 API，使开发人员可以编写能够在NVIDIA GPU上高效运行的代码。
- CUDA 广泛应用于科学计算、机器学习、深度学习、图像处理、视频编码等多个领域。
- CUDA和cuDNN（CUDA 深度神经网络库）已经成为训练复杂神经网络不可或缺的一部分。

感谢！