AI芯片基础知识

2025年1月16日 R1.0版

鲜枣课堂

目录

CONTENTS

PART 01

芯片的分类

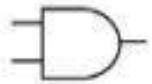
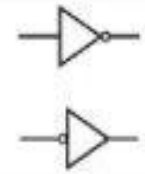
芯片的分类

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

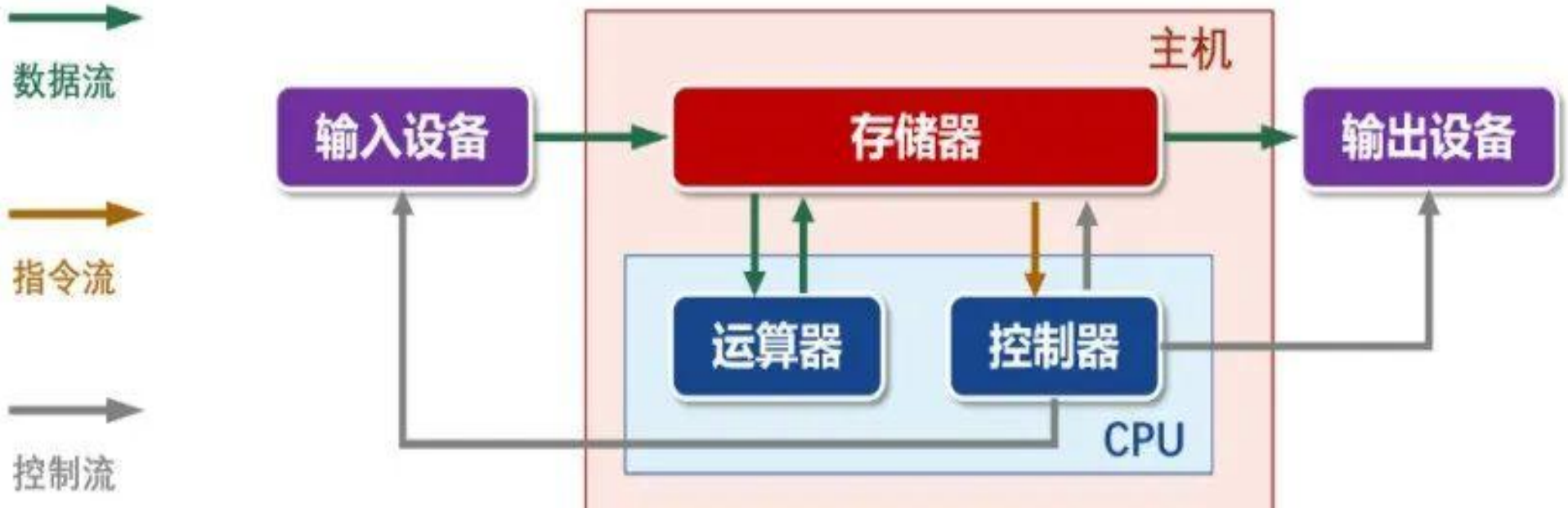
CPU的定义

- 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。

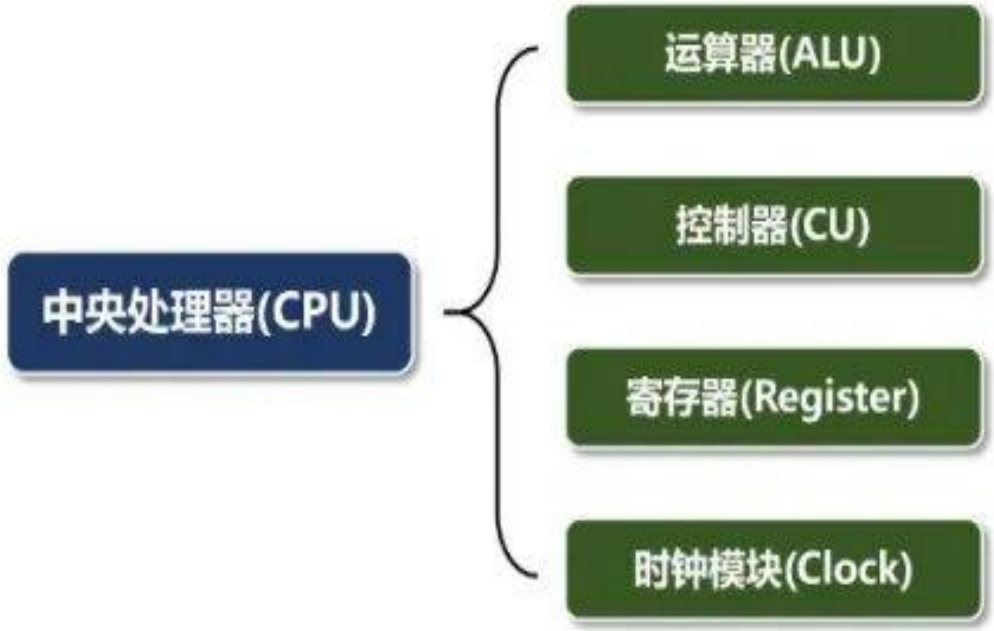
CPU和GPU

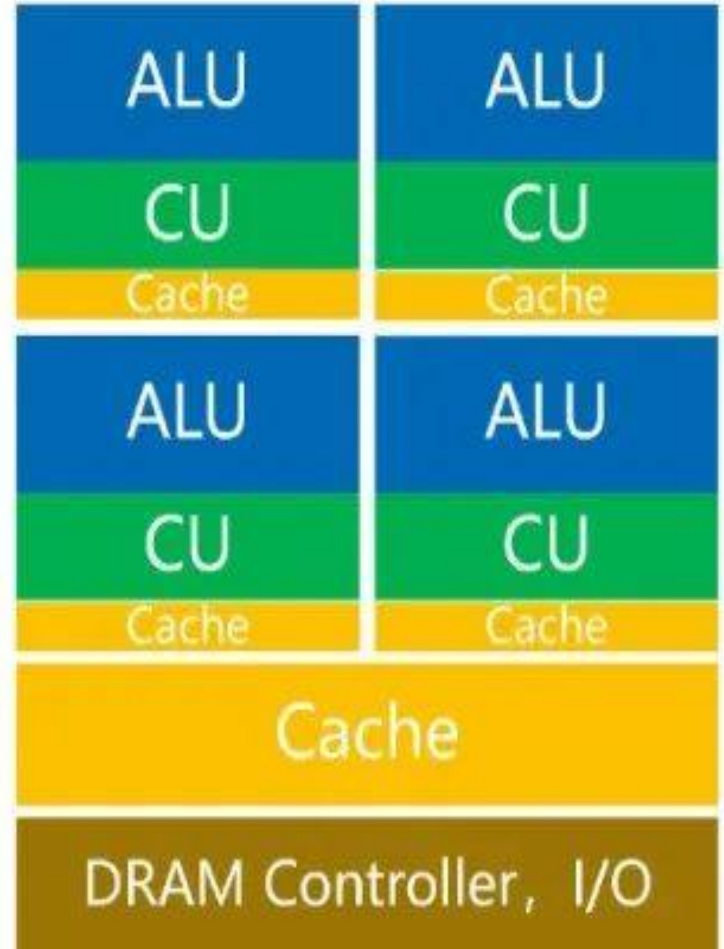
CPU的主要组成

- **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

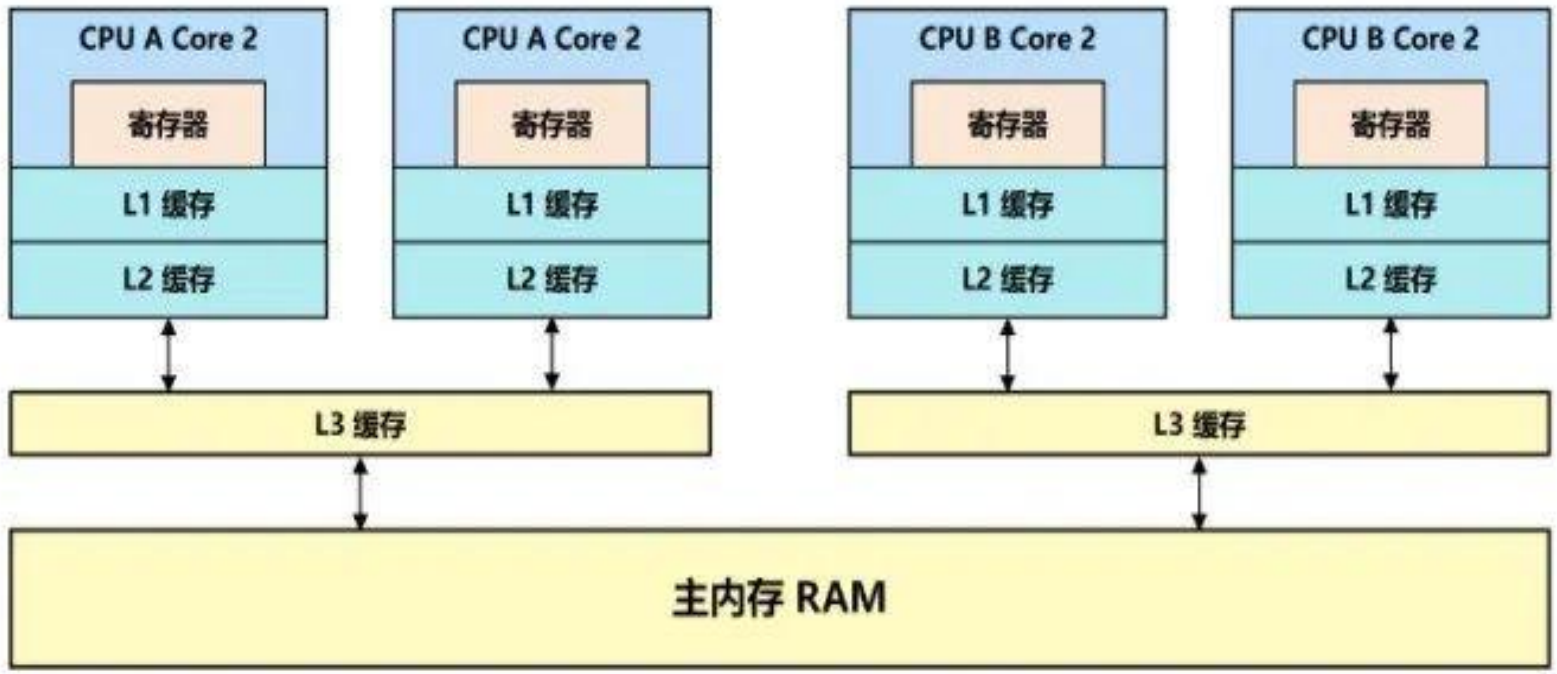
CPU的主要组成





CPU和GPU

多核CPU



CPU多核硬件架构示例

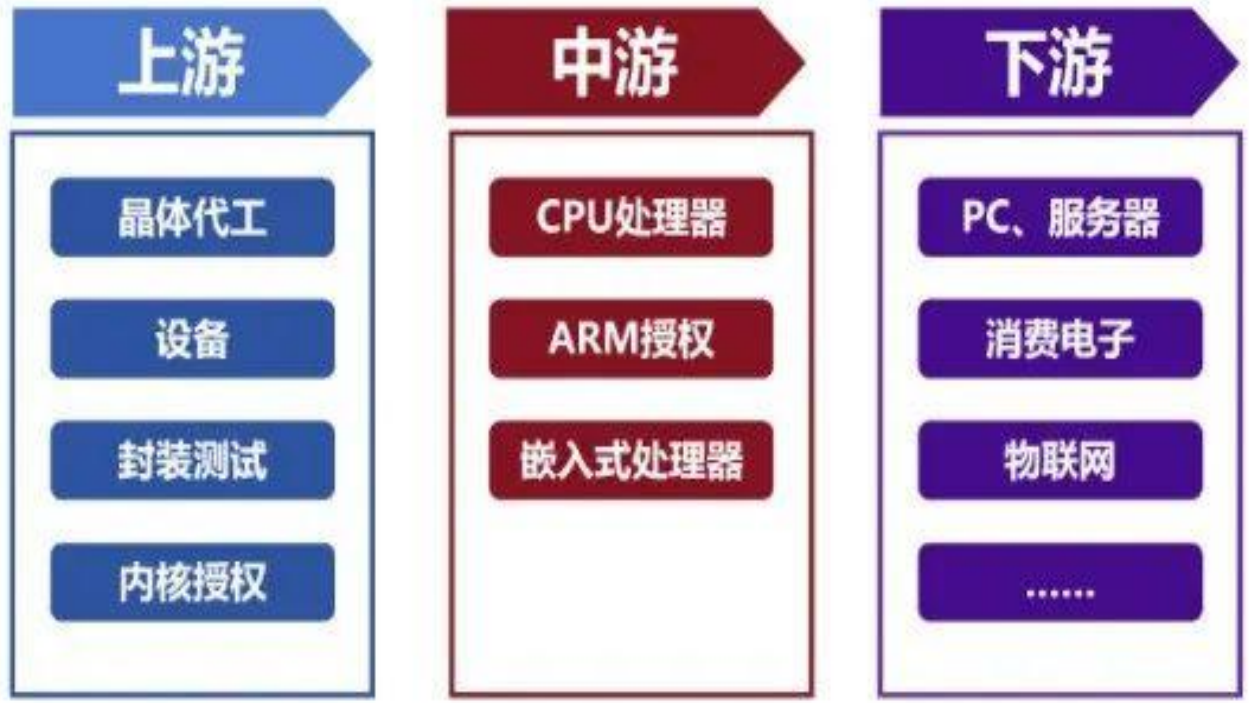
CPU和GPU

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

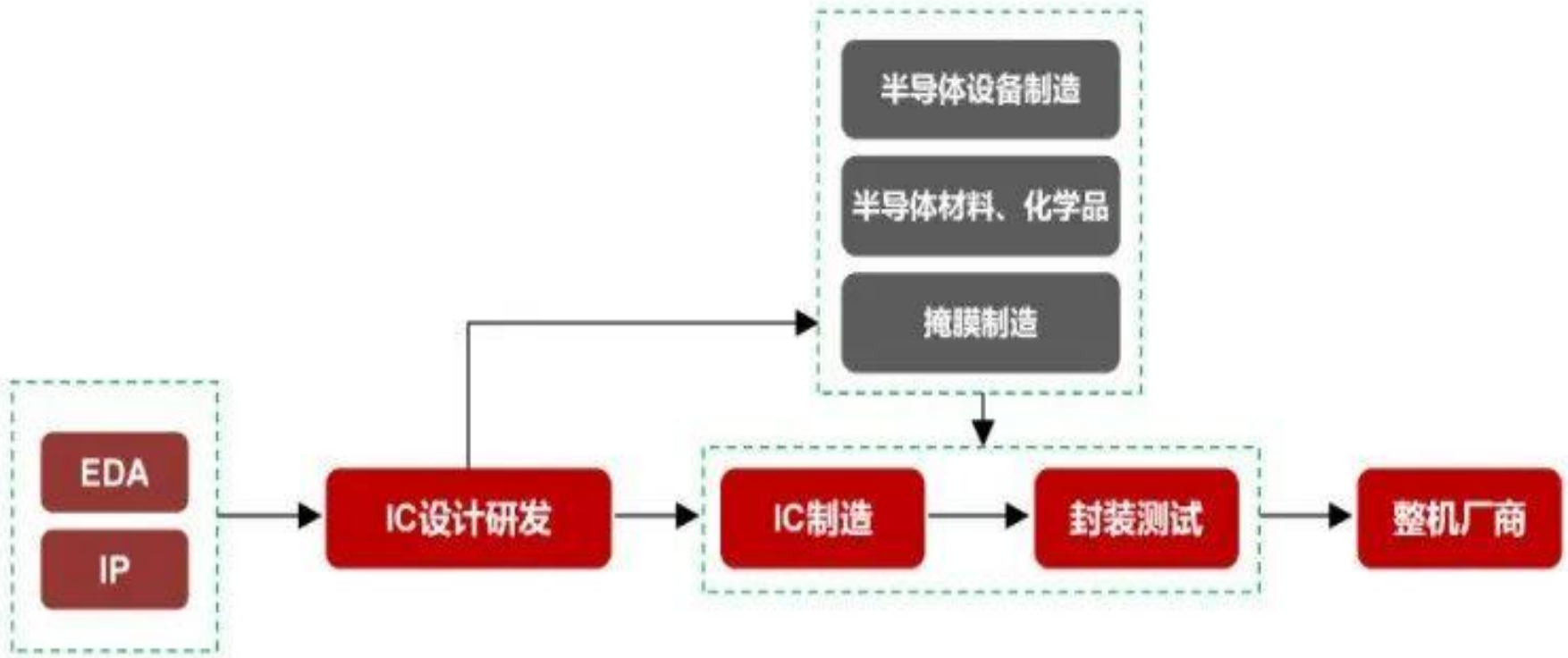
CPU和GPU

半导体产业链



CPU和GPU

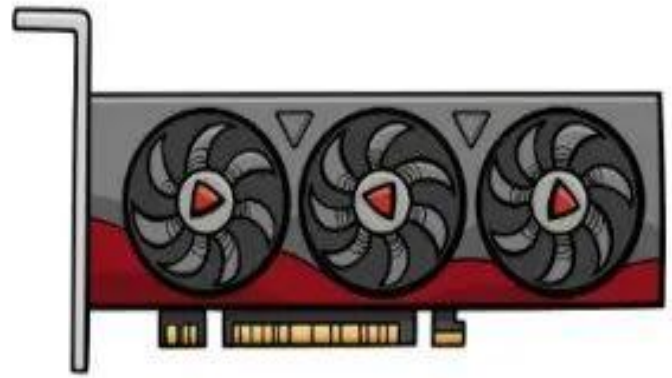
半导体产业链



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。

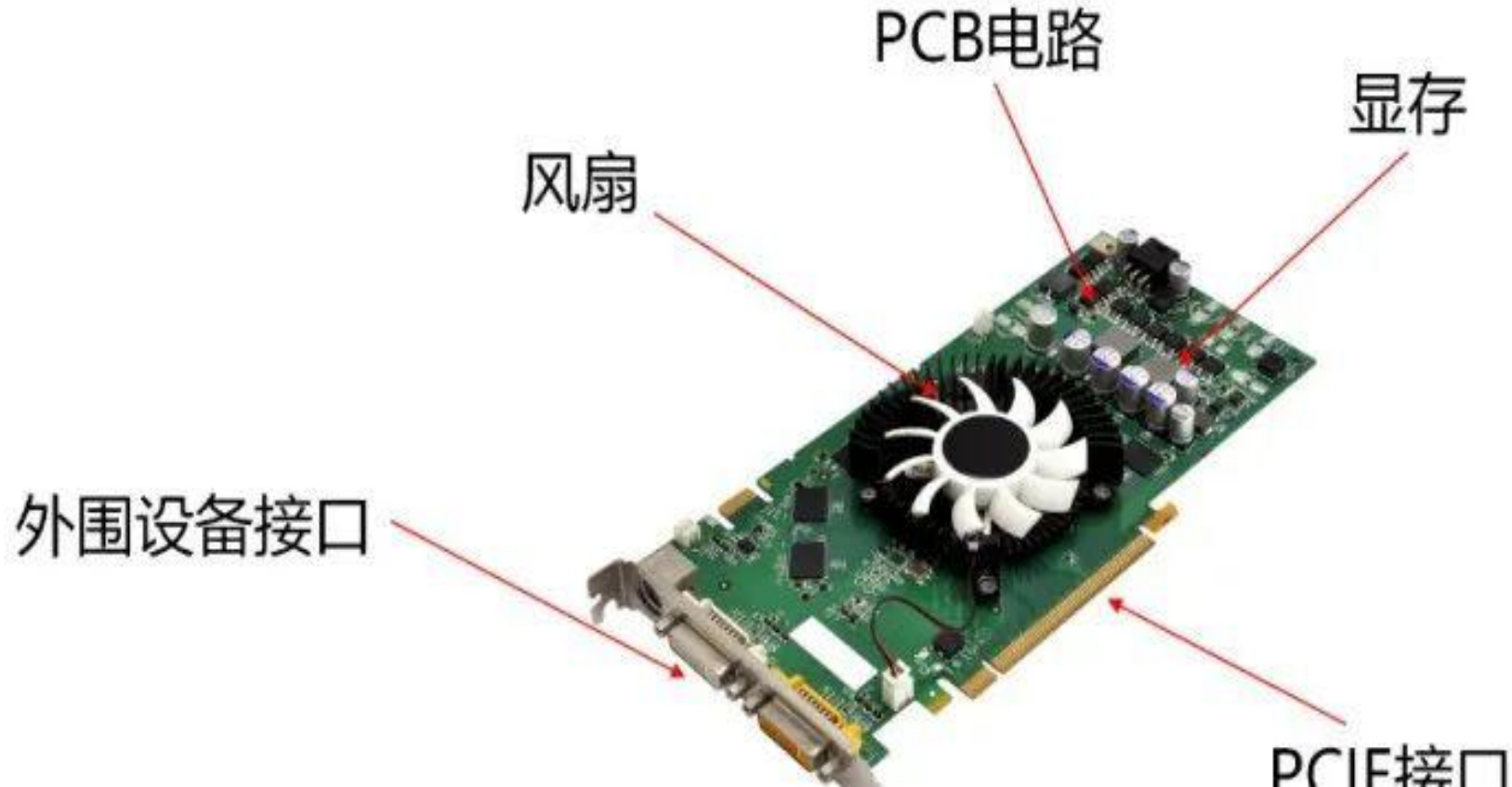
- **支持多种编程接口**

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

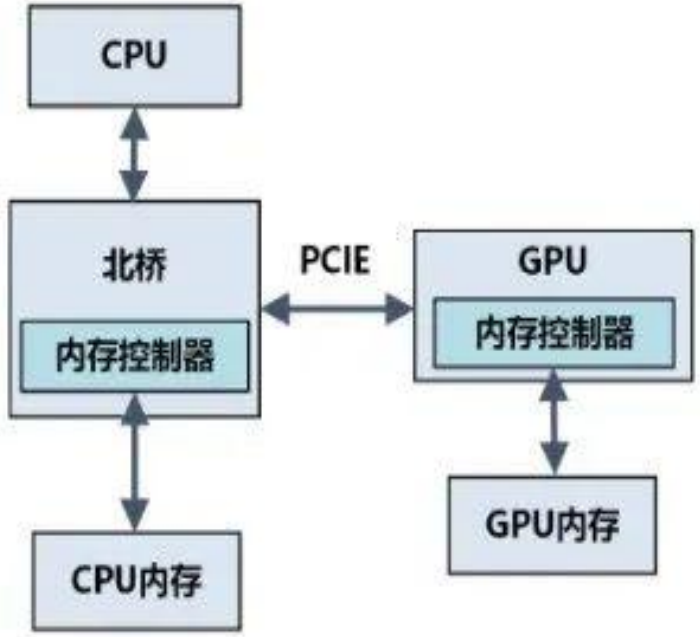
显卡产业的发展历程

- 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。

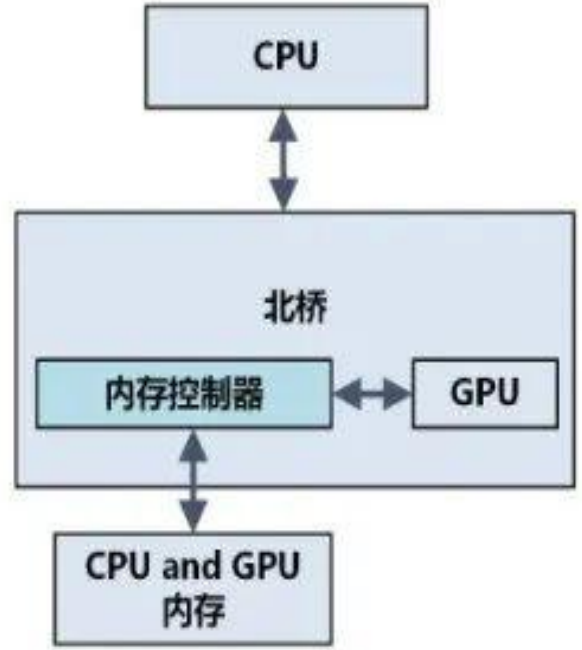
CPU和GPU

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

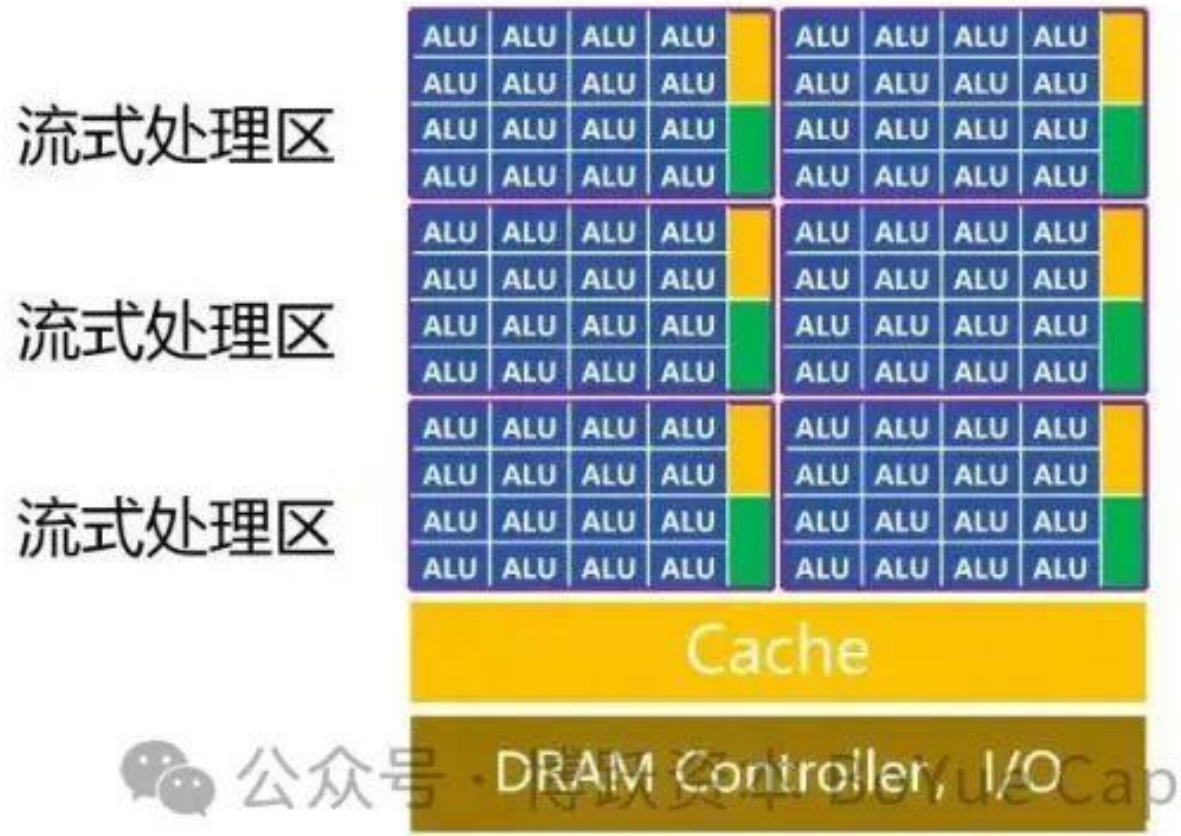
英伟达消费级显卡经典型号

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。

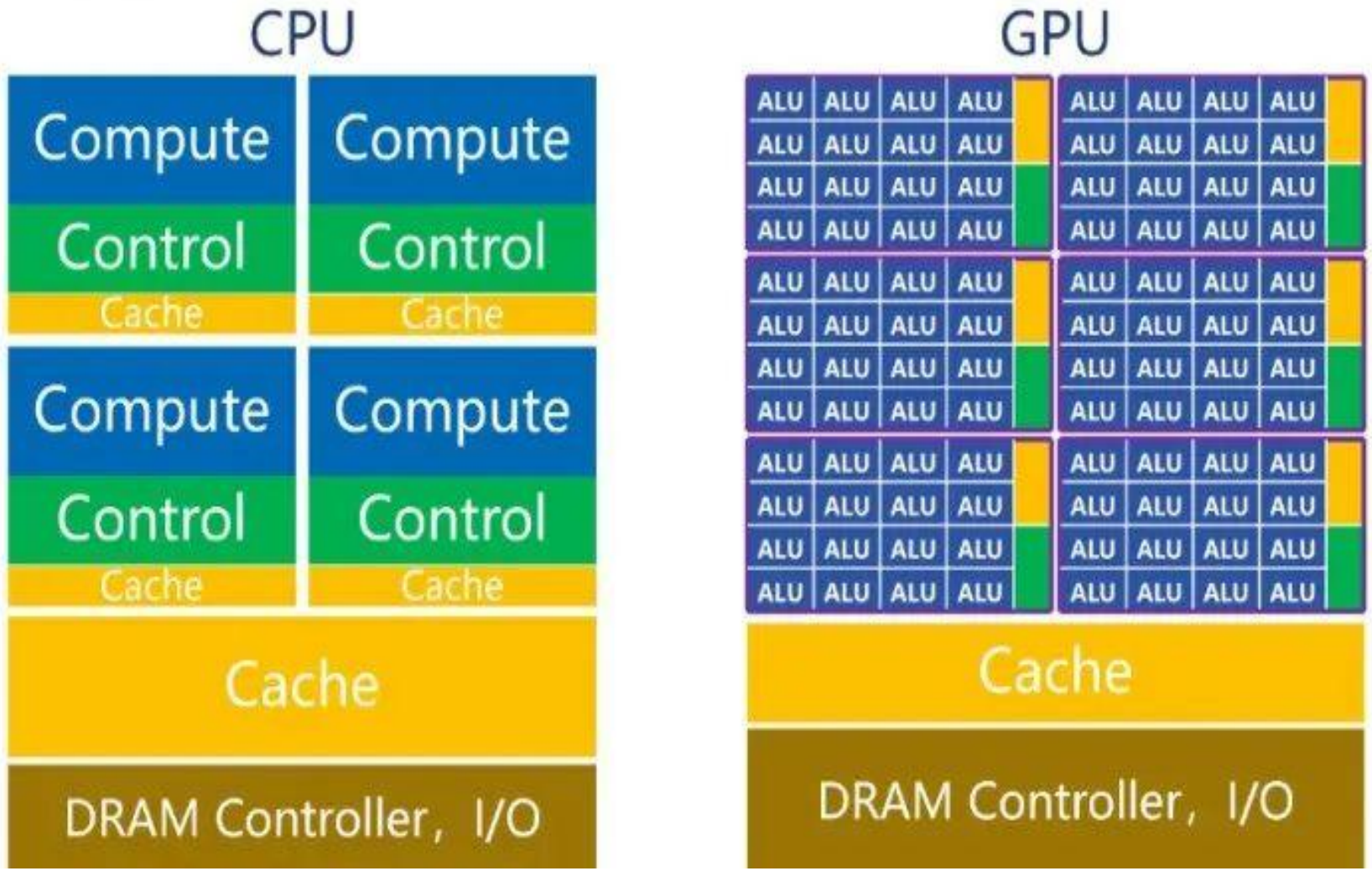
CPU和GPU

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。

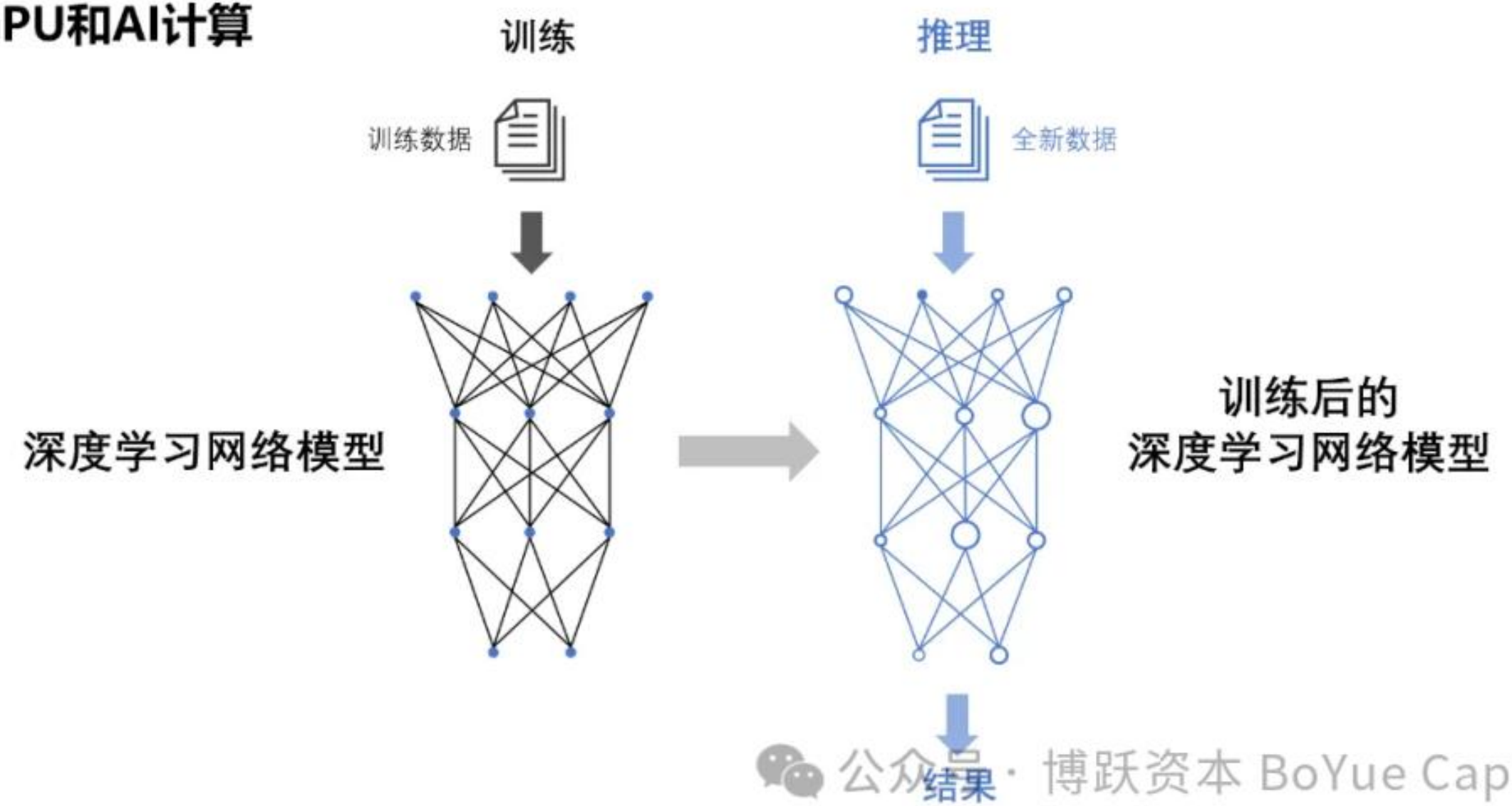
CPU和GPU

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。

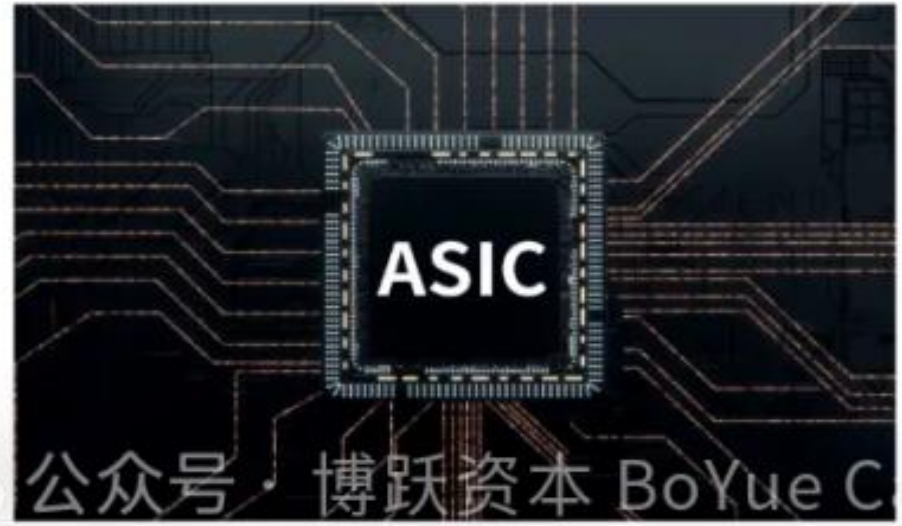
PART 03

ASIC和FPGA

ASIC和FPGA

ASIC的定义

- ASIC（Application Specific Integrated Circuit，专用集成电路），就是一种专用于特定任务的芯片。
- ASIC的官方定义：应特定用户的要求，或特定电子系统的需要，专门设计、制造的集成电路。
- ASIC芯片面向专项任务，计算能力和计算效率都严格匹配于任务算法。
- ASIC芯片的核心数量、逻辑计算单元和控制单元比例，以及缓存等，都是精确定制的。



ASIC和FPGA

ASIC的优点

- 高性能：由于ASIC是为特定应用定制的，它们可以在这些应用中提供比通用芯片更高的性能。
- 低功耗：ASIC可以通过优化电路设计来降低功耗，这在移动设备和嵌入式系统中尤为重要。
- 高成本效益：尽管ASIC的设计和制造初期成本较高，但单位成本会随着产量增加而显著下降。
- 小尺寸：ASIC可以集成更多功能于更小的芯片面积内，有助于减小产品体积。
- 高安全性：由于ASIC是专为特定用途设计的，它能够提供更强的安全特性，防止逆向工程和攻击。

ASIC和FPGA

ASIC的缺点

- 成本高昂，技术难度大。对芯片进行定制设计，对一家企业的研发技术水平要求极高，且耗资极为巨大。
- 研发一款ASIC芯片，首先要经过代码设计、综合、后端等复杂的设计流程，再经过几个月的生产加工以及封装测试，才能拿到芯片来搭建系统。
- 研发ASIC需要“流片（Tape-out）”。像流水线一样，通过一系列工艺步骤制造芯片，就是流片。简单来说，就是试生产。14nm工艺，流片一次需要300万美元左右。5nm工艺，更是高达4725万美元。流片一旦失败，将损耗大量的经费，耽误大量的时间和精力。

ASIC和FPGA

ASIC的应用领域

- 消费电子产品：例如智能手机中的基带处理器、图像信号处理器等。
- 网络通信：包括路由器、交换机中的交换芯片等。
- 加密货币挖矿：比特币等加密货币的挖矿机常采用ASIC来提高哈希率并减少电力消耗。
- 汽车电子：用于高级驾驶辅助系统（ADAS）、动力总成控制单元等。
- 医疗设备：如超声波成像仪、心电图仪等需要高效处理大量传感器数据的设备。
- 人工智能/机器学习加速器：一些公司开发了专门用于AI推理和训练的ASIC，如Google的TPU（张量处理单元）。

ASIC和FPGA

TPU

- TPU，全称Tensor Processing Unit，张量处理单元。
- TPU专为加速TensorFlow框架中的张量运算而设计，显著提高了深度学习模型训练和推理的速度与效率。
- 所谓“张量（tensor）”，是一个包含多个数字（多维数组）的数学实体。
- 目前，几乎所有的机器学习系统，都使用张量作为基本数据结构。所以，张量处理单元，我们可以简单理解为“AI处理单元”。

ASIC和FPGA

TPU

- 2015年，为了更好地完成自己的深度学习任务，提升AI算力，Google推出了一款专门用于神经网络训练的芯片，也就是TPU v1。
- 相比传统的CPU和GPU，在神经网络计算方面，TPU v1可以获得15~30倍的性能提升，能效提升更是达到30~80倍，给行业带来了很大震动。
- 2017年和2018年，Google又推出了能力更强的TPU v2和TPU v3，用于AI训练和推理。
- 2021年，Google推出了TPU v4，采用7nm工艺，晶体管数达到220亿，性能相较上代提升了10倍，比英伟达的A100还强1.7倍。

ASIC和FPGA

其它ASIC

- 除了Google之外，还有很多头部企业这几年也在研发ASIC。
- 英特尔公司在2019年底收购了以色列AI芯片公司Habana Labs，2022年，发布了Gaudi 2 ASIC芯片。
- 2022年底，IBM研究院发布了AI ASIC芯片AIU。
- 三星早几年也推出过ASIC，当时做的是矿机专用芯片。

ASIC和FPGA

NPU

- NPU，Neural Processing Unit，神经网络处理单元。
- 在电路层模拟人类神经元和突触，并用深度学习指令集处理数据。
- NPU专门用于神经网络推理，能够实现高效的卷积、池化等操作。
- 经常集成在手机SoC芯片中，提供端侧的AI计算能力。

ASIC和FPGA

DPU

- DPU，Data Processing Unit，数据处理单元。
- 被设计用于加速数据中心内的特定任务，特别是那些与网络、存储和安全相关的任务。
- 用于卸载、加速和隔离关键基础设施服务，从而提高效率、性能和安全性。
- 部分DPU制造商和技术：
  - 英伟达BlueField DPU
  - Fungible DPU
  - 英特尔 Infrastructure Processing Units（IPU）

ASIC和FPGA

华为昇腾

- 华为昇腾（Ascend）系列处理器是华为自主研发的AI芯片，也属于ASIC芯片。
- 采用了先进的架构和制程技术，提供卓越的计算性能。
- 支持多种主流的AI框架，如TensorFlow、PyTorch、Caffe等，并且与华为自己的MindSpore框架紧密集成。

芯片	异腾910 Ascend910	异腾310 Ascend310
功能	训练	推理
架构	达芬奇	达芬奇
工艺	7nm	12nm
算力	INT8 640TOPS FP16 320TFLOPS	INT8 22TOPS FP16 11TFLOPS
功耗	310W	8W
内存	HBM2E	2*LPDDR4x

ASIC和FPGA

FPGA的定义

- FPGA，英文全称Field Programmable Gate Array，现场可编程门阵列。
- FPGA是在PAL（可编程阵列逻辑）、GAL（通用阵列逻辑）等可编程器件的基础上发展起来的产物，属于一种半定制电路。
- 简单来说，FPGA就是可以重构的芯片。它可以根据用户的需要，在制造后，进行无限次数的重复编程，以实现想要的数字逻辑功能。

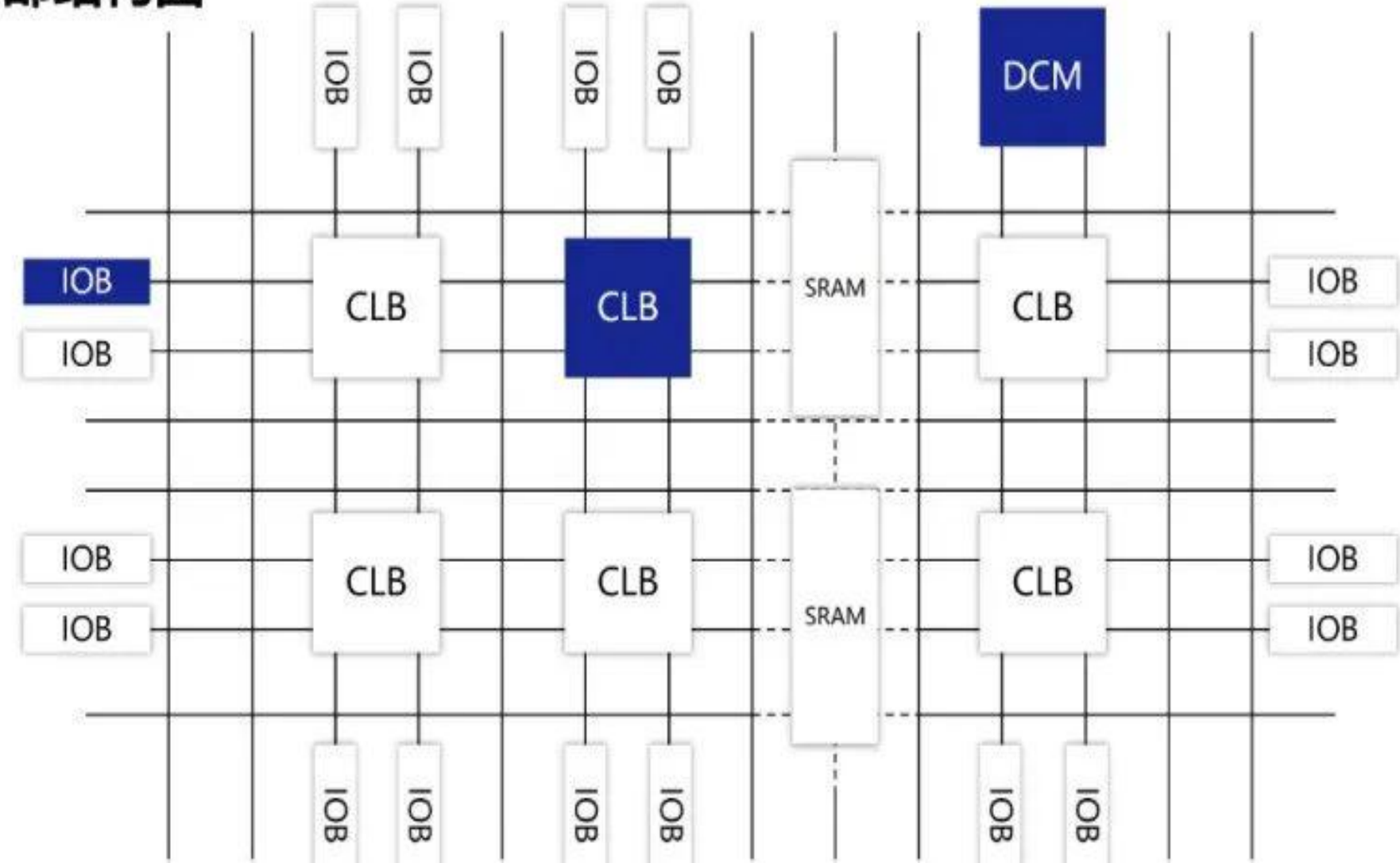
ASIC和FPGA

FPGA的组成部分

- **三种可编程电路：**
  - **可编程逻辑块（Configurable Logic Blocks，CLB）：**最重要的部分，是实现逻辑功能的基本单元，承载主要的电路功能。它们通常规则排列成一个阵列（逻辑单元阵列，LCA，Logic Cell Array），散布于整个芯片中。
  - **输入/输出模块（I/O Blocks，IOB）：**主要完成芯片上的逻辑与外部引脚的接口，通常排列在芯片的四周。
  - **可编程互连资源（Programmable Interconnect Resources，PIR）：**提供了丰富的连线资源，包括纵横网状连线、可编程开关矩阵和可编程连接点等。它们实现连接的作用，构成特定功能的电路。
- **静态存储器SRAM：**
  - 用于存放内部IOB、CLB和PIR的编程数据，并形成对它们的控制，从而完成系统逻辑功能。

ASIC和FPGA

FPGA内部结构图



ASIC和FPGA

FPGA的组成部分

- CLB本身，又主要由查找表（Look-Up Table，LUT）、多路复用器（Multiplexer）和触发器（Flip-Flop）构成。它们用于承载电路中的一个个逻辑“门”，可以用来实现复杂的逻辑功能。
- 我们可以把LUT理解为存储了计算结果的RAM。当用户描述了一个逻辑电路后，软件会计算所有可能的结果，并写入这个RAM。每一个信号进行逻辑运算，就等于输入一个地址，进行查表。LUT会找出地址对应的内容，返回结果。这种“硬件化”的运算方式，显然具有更快的运算速度。
- FPGA的逻辑单元功能在编程时已确定，属于用硬件来实现软件算法。对于保存状态的需求，FPGA中的寄存器和片上内存（BRAM）属于各自的控制逻辑，不需要仲裁和缓存。

ASIC和FPGA

FPGA的使用

- 用户使用FPGA时，可以通过硬件描述语言（Verilog或VHDL），完成的电路设计，然后对FPGA进行“编程”（烧写），将设计加载到FPGA上，实现对应的功能。
- 加电时，FPGA将EPROM（可擦编程只读存储器）中的数据读入SRAM中，配置完成后，FPGA进入工作状态。掉电后，FPGA恢复成白片，内部逻辑关系消失。如此反复，就实现了“现场”定制。
- FPGA的功能非常强大。理论上，如果FPGA提供的门电路规模足够大，通过编程，就能够实现任意ASIC的逻辑功能。

ASIC和FPGA

FPGA厂商

- **海外四巨头：**
  - Xilinx公司（赛灵思）：2020年，AMD以350亿美元收购了Xilinx。
  - Altera（阿尔特拉）：2015年5月，Intel 以167亿美元的天价收购了Altera，后来收编为PSG（可编程解决方案事业部）部门。2023年10月，Intel宣布计划拆分PSG部门，独立业务运营。
  - Lattice（莱迪思）
  - Microsemi（美高森美）
- **国内厂商：**
  - 复旦微电、紫光国微、安路科技、东土科技、高云半导体、京微齐力、京微雅格、智多晶、遨格芯等。

ASIC和FPGA

FPGA产业链

ASIC和FPGA

ASIC和FPGA的区别

- ASIC和FPGA，本质上都是芯片。AISC是全定制芯片，功能写死，没办法改。而FPGA是半定制芯片，功能灵活，可修改性强。
- 类比：
  - ASIC：模具玩具。事先要进行开模，比较费事。一旦开模之后，就没办法修改了。如果要做新玩具，就必须重新开模。
  - FPGA：乐高积木。上手就能搭，花一点时间就可以搭好。如果不满意，或者想搭新玩具，可以拆开，重新搭。

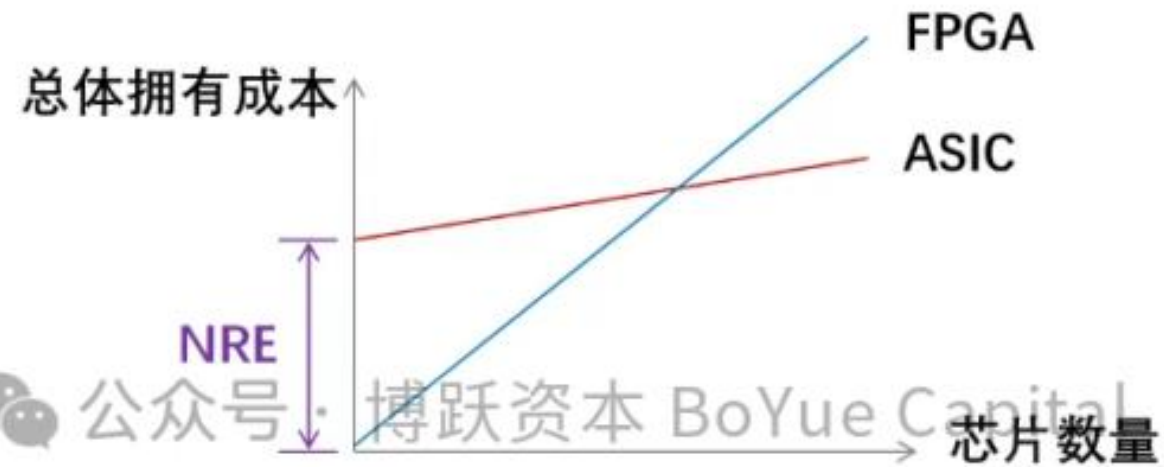
ASIC

FPGA

ASIC和FPGA

ASIC和FPGA的区别

- 设计：ASIC与FPGA的很多设计工具是相同的。在设计流程上，FPGA没有ASIC那么复杂，去掉了一些制造过程和额外的设计验证步骤，大概只有ASIC流程的50%-70%。
- 流片：ASIC需要流片。FPGA不需要流片。
- 开发周期：开发ASIC，可能需要几个月甚至一年以上的时间。开发FPGA，只需要几周或几个月的时间。
- 成本：FPGA可以在实验室或现场进行预制和编程，不需要一次性工程费用 （NRE）。但是，作为“通用玩具”，它的成本是ASIC（压模玩具）的10倍。如果生产量比较低，那么，FPGA会更便宜。如果生产量高，ASIC的一次性工程费用被平摊，那么，ASIC反而便宜。



ASIC和FPGA

ASIC和FPGA的区别

- 性能和功耗：作为专用定制芯片，ASIC比FPGA强。
- FPGA是通用可编辑的芯片，冗余功能比较多。无论怎么设计，都会多出来一些部件。
- FPGA和ASIC，不是简单的竞争和替代关系，而是各自的定位不同。
- FPGA现在多用于产品原型的开发、设计迭代，以及一些低产量的特定应用。它适合那些开发周期必须短的产品。FPGA还经常用于ASIC的验证。
- ASIC用于设计规模大、复杂度高的芯片，或者是成熟度高、产量比较大的产品。

ASIC和FPGA

FPGA的应用场景

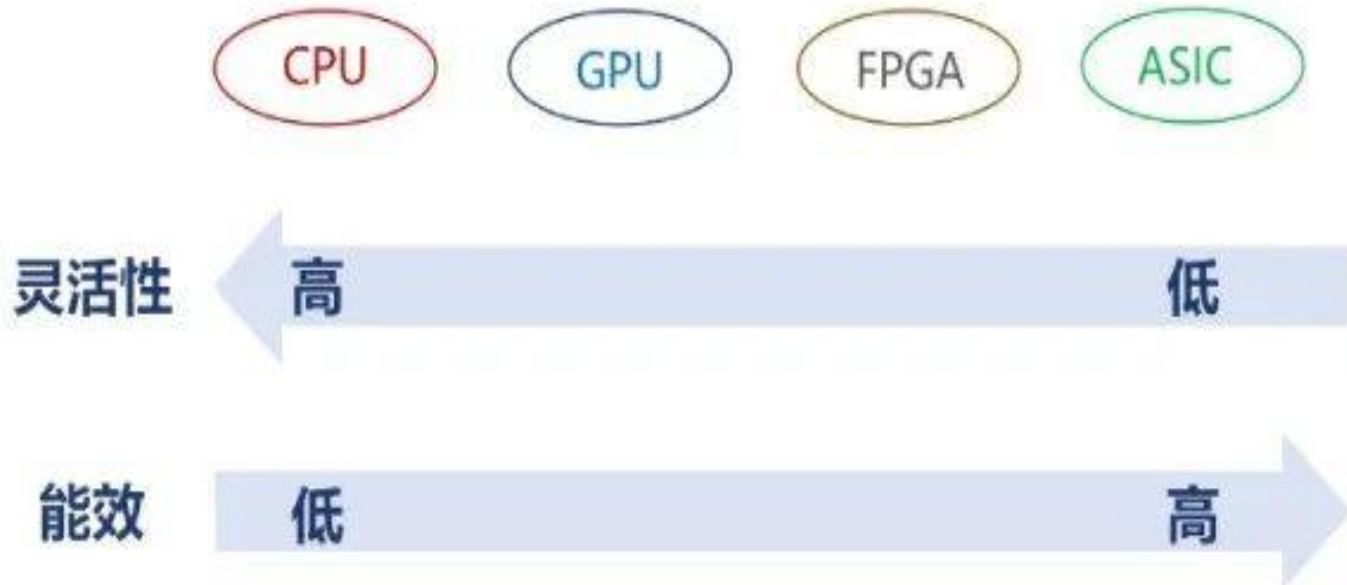
- FPGA特别适合初学者学习和参加比赛。现在很多大学的电子类专业，都在使用FPGA进行教学。
- 从商业化的角度来看，FPGA的主要应用领域是通信、国防、航空、数据中心、医疗、汽车及消费电子。
- FPGA在通信领域用得很早。很多基站的处理芯片（基带处理、波束赋形、天线收发器等），都是用的FPGA。核心网的编码和协议加速等，也用到它。数据中心之前在DPU等部件上，也用。后来，很多技术成熟了、定型了，通信设备商们就开始用ASIC替代，以此减少成本。

PART 04

总结对比

总结对比

整体对比



总结对比

整体对比

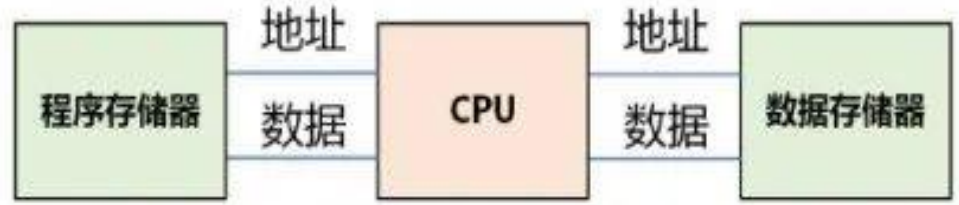
	CPU	GPU	FPGA	ASIC
定制化程度	通用	半通用	半定制化	全定制化
灵活性	高	高	高	低
成本	较低	高	较高	低
功耗	较高	高	较高	低
主要优点	通用性最强	计算能力强 生态成熟	灵活强较高	能效最高
主要缺点	并行算力弱	功耗较大 编程难度较大	峰值计算能力弱 编程难度较难	研发时间长 技术风险高
应用场景	较少用于AI	云端训练和推理	云端推理 终端推理	云端训练和推理 终端推理

总结对比

整体对比

- 从理论和架构的角度，ASIC和FPGA的性能和成本，肯定是优于CPU和GPU的。
- CPU、GPU遵循的是冯·诺依曼体系结构，指令要经过存储、译码、执行等步骤，共享内存在使用时，要经历仲裁和缓存。
- 而FPGA和ASIC并不是冯·诺依曼架构（是哈佛架构）。以FPGA为例，它本质上是无指令、无需共享内存的体系结构。

冯诺依曼架构



哈佛架构

总结对比

运算单元对比

- 从ALU运算单元占比来看，GPU比CPU高，并行计算效率更高。
- FPGA因为几乎没有控制模块，所有模块都是ALU运算单元，比GPU更高。

总结对比

功耗对比

- GPU的功耗极高，单片可以达到250W，甚至600W（RTX5090）。而FPGA一般只有30~50W。
- 这主要是因为内存读取。GPU的内存接口（GDDR5、HBM、HBM2）带宽极高，大约是FPGA传统DDR接口的4-5倍。但就芯片本身来说，读取DRAM所消耗的能量，是SRAM的100倍以上。GPU频繁读取DRAM的处理，产生了极高的功耗。
- 另外，FPGA的工作主频（500MHz以下）比CPU、GPU（1~3GHz）低，也会使得自身功耗更低。FPGA的工作主频低，主要是受布线资源的限制。有些线要绕远，无法支持更高的时钟频率。

总结对比

时延对比

- GPU时延高于FPGA。
- GPU通常需要将不同的训练样本，划分成固定大小的“Batch（批次）”，为了最大化达到并行性，需要将数个Batch都集齐，再统一进行处理。
- FPGA的架构，是无批次（Batch-less）的。每处理完成一个数据包，就能马上输出，时延更有优势。

总结对比

目前GPU在AI芯片占比较大的主要原因

- 在英伟达的长期努力下，GPU的核心数和工作频率一直在提升，芯片面积也越来越大，算力非常强劲。
- 功耗方面，GPU依赖先进的工艺制程，以及水冷等被动散热，可以勉强支撑。
- 生态方面，英伟达推出的CUDA编程模型及其相关库（如cuDNN）已经成为行业标准，极大地简化了开发者编写高效GPU代码的过程。几乎所有的主流深度学习框架（TensorFlow、PyTorch等）都内置了对GPU的支持，这进一步促进了其普及。
- 普及性方面，由于GPU已经广泛存在于个人电脑、服务器甚至移动设备中，因此基于GPU的解决方案更容易被采纳，并且可以利用现有的硬件基础设施。

总结对比

AI芯片发展趋势

- ASIC芯片加速崛起，提升市场占比。
- 随着摩尔定律逐渐接近极限，业界正在探索新的计算范式（如量子计算、类脑计算）。
- 结合CPU、GPU、DSP、FPGA等多种芯片的异构计算加速普及，可以根据不同任务的需求灵活调配资源，实现更高的效率和更低的功耗。
- 端侧推理AI芯片发展提速。
- 针对特定行业或应用领域（如自动驾驶、医疗影像、智能家居）的高度定制化AI芯片需求增加。

参考文献

- 《GPU行业深度报告》，华金证券；
- 《GPU研究框架》，中信证券；
- 《中国FPGA行业芯片研究报告》，头豹研究院；
- 百度百科、维基百科。