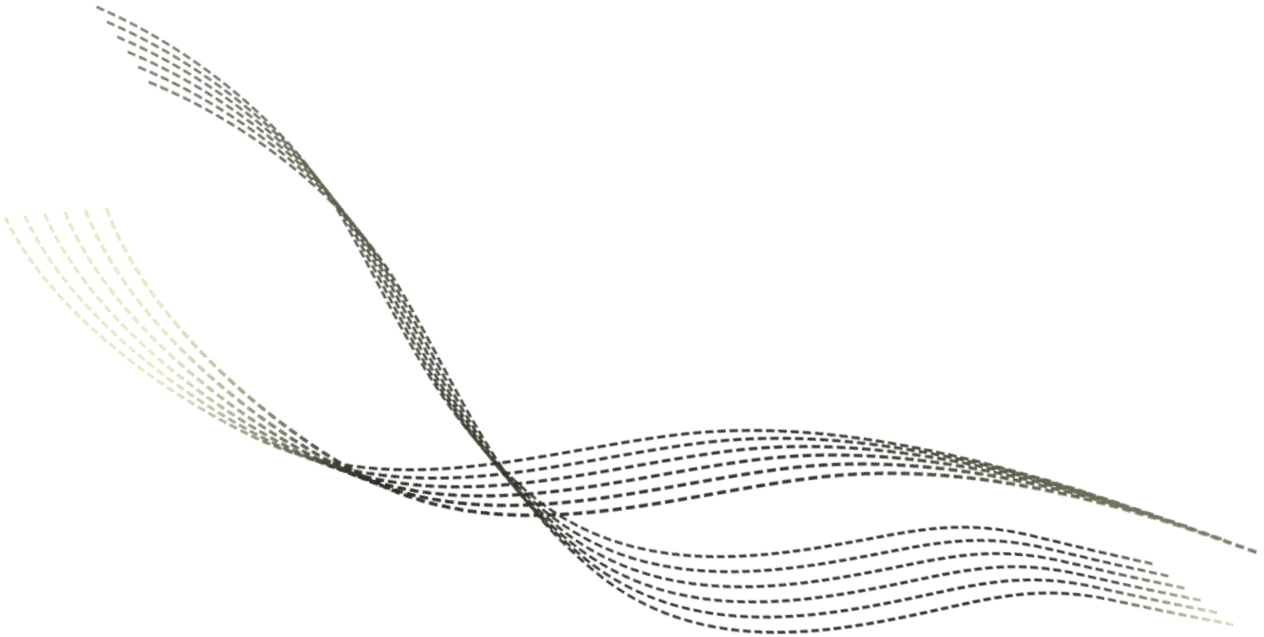
高级计算机体系结构

Advanced Computer Architecture



集成电路现状

沈明华

目录

CONTENTS

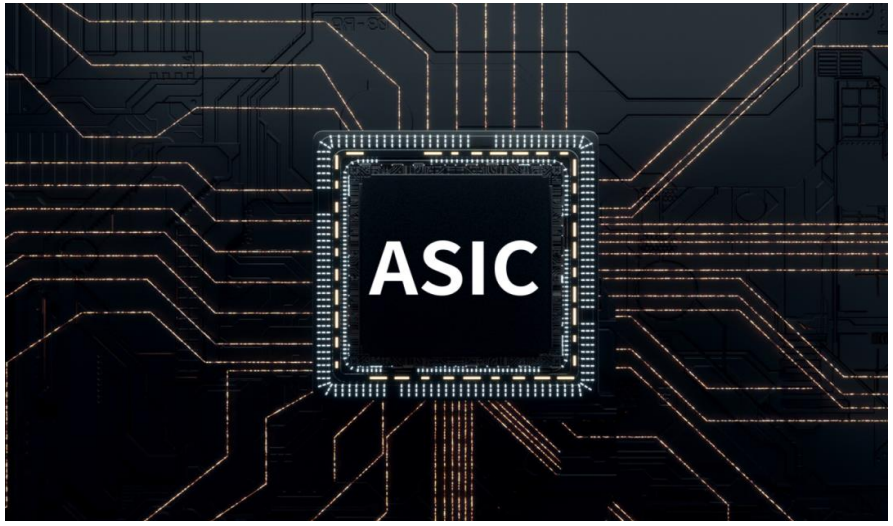
PART 03

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

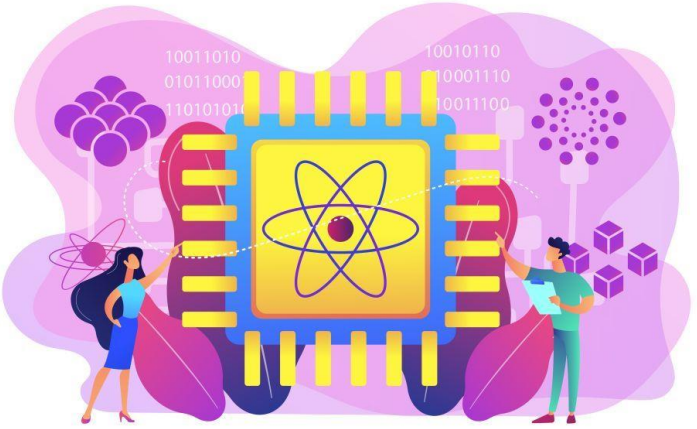
□ TPU

- 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	昇腾910 Ascend910	昇腾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：**

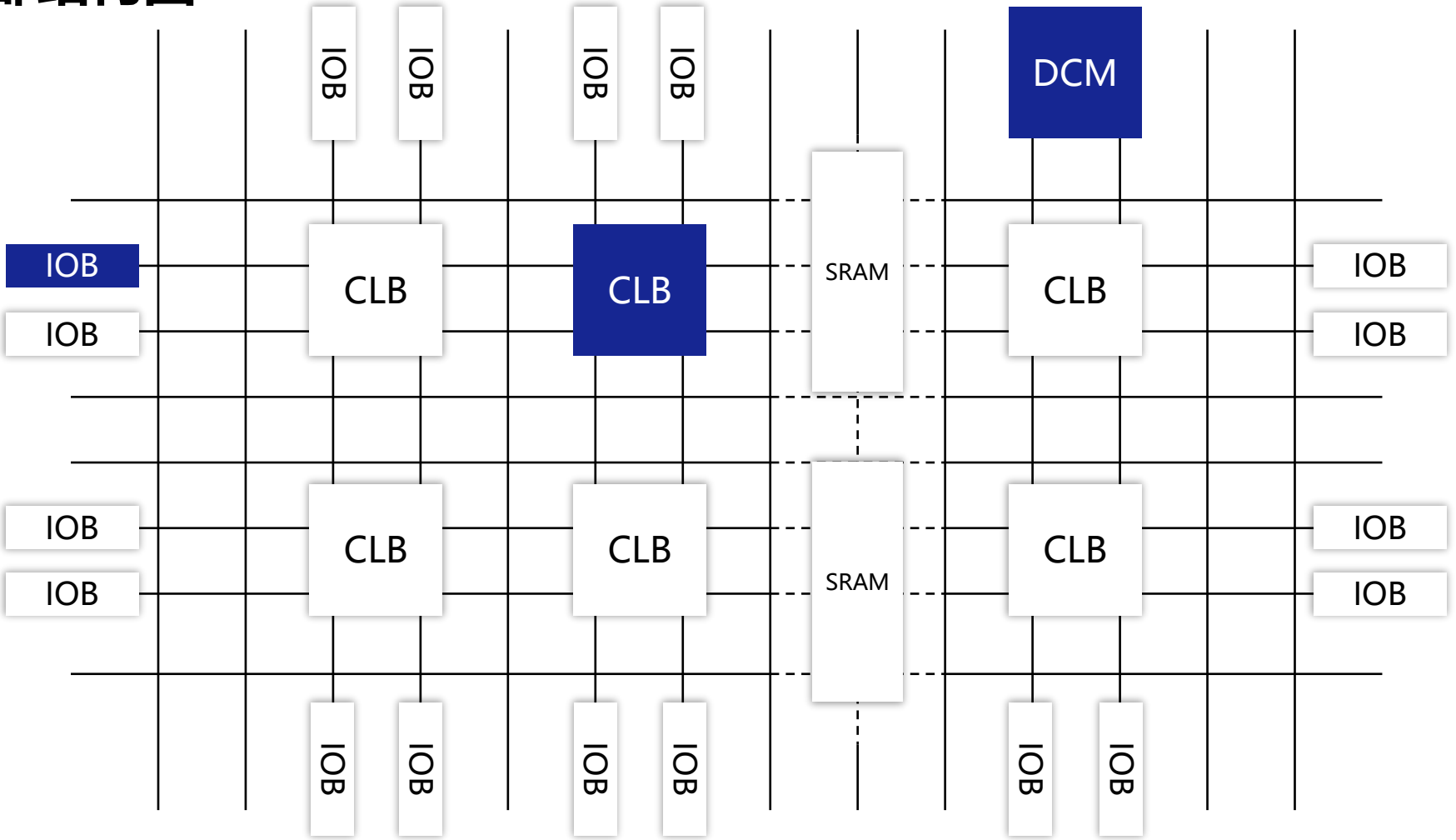
- **可编程逻辑块（Configurable Logic Blocks，CLB）：**最重要的部分，是实现逻辑功能的基本单元，承载主要的电路功能。它们通常规则排列成一个阵列（逻辑单元阵列，LCA，Logic Cell Array），散布于整个芯片中。
- **输入/输出模块（I/O Blocks，IOB）：**主要完成芯片上的逻辑与外部引脚的接口，通常排列在芯片的四周。
- **可编程互连资源（Programmable Interconnect Resources，PIR）：**提供了丰富的连线资源，包括纵横网状连线、可编程开关矩阵和可编程连接点等。它们实现连接的作用，构成特定功能的电路。

– **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。

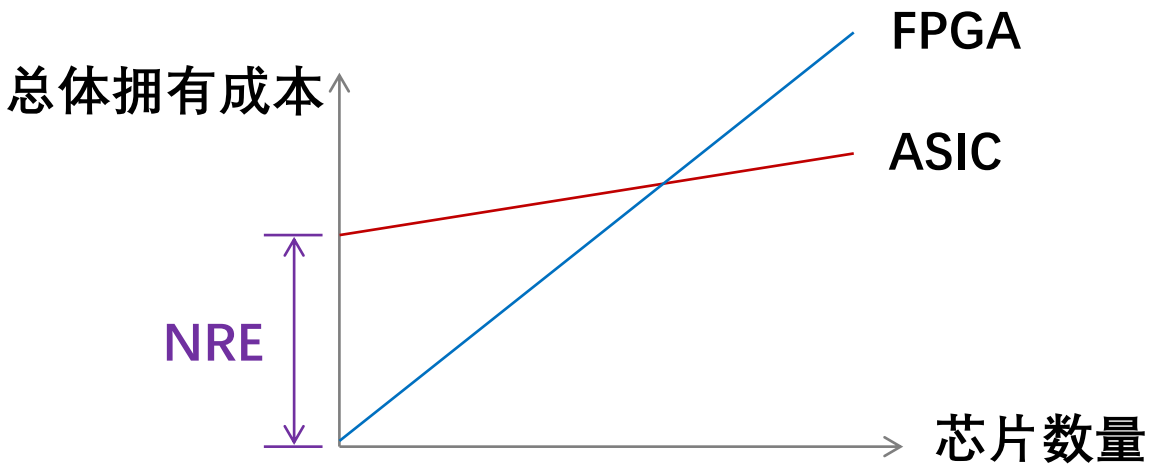
ASIC

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

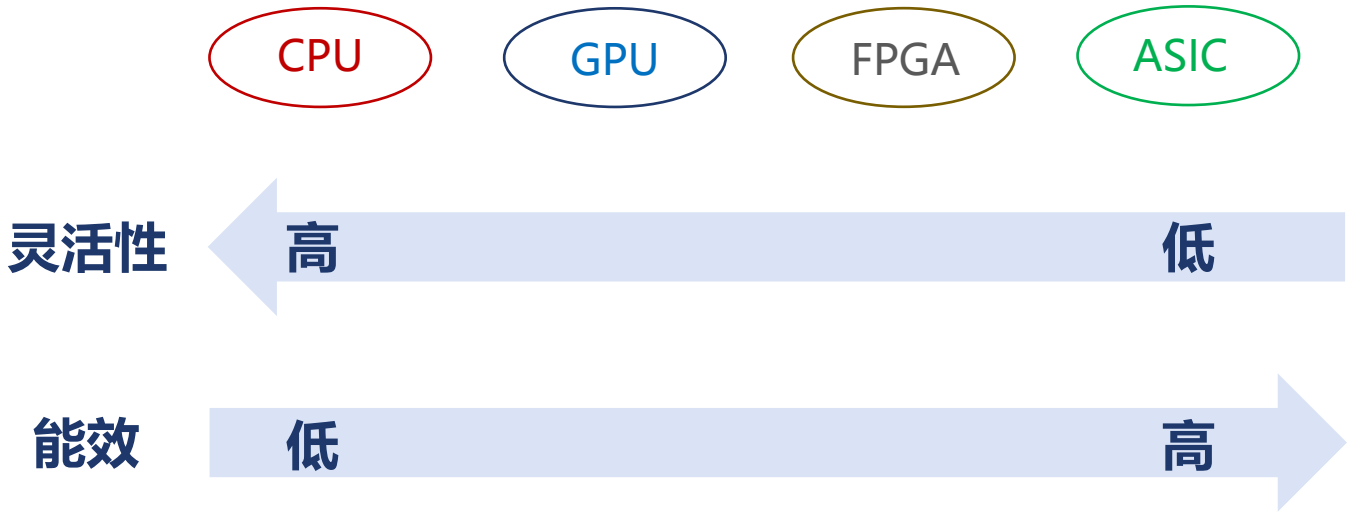
□ FPGA的应用场景

- 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

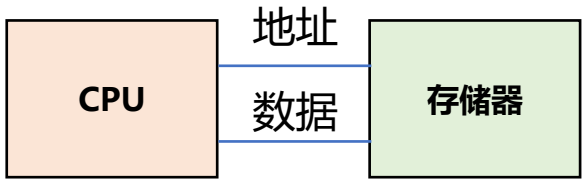
□ 整体对比

	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

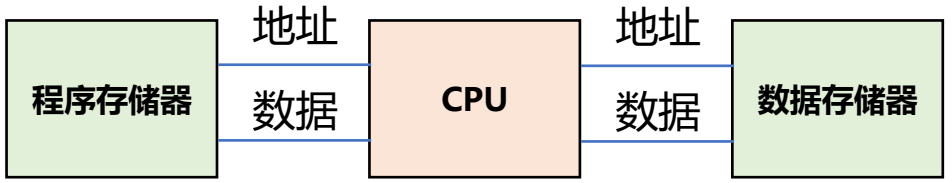
■ 总结对比

□ 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。

感谢！