AI 芯片 – AI 芯片基础

AI专用处理器

ZOMI

Talk Overview

1. AI 计算体系

- 深度学习计算模式
- 计算体系与矩阵运算

2. AI 芯片基础

- 通用处理器 CPU
- 从数据看 CPU 计算
- 通用图形处理器 GPU
- AI专用处理器 NPU/TPU
- 计算体系架构的黄金10年

Talk Overview

1. AI 计算体系
   - 深度学习计算模式
   - 计算体系与矩阵运算
2. **AI 芯片基础**
   - 通用处理器 CPU
   - 从数据看 CPU 计算
   - 通用图形处理器 GPU
   - AI专用处理器 NPU/TPU
   - 计算体系架构的黄金10年

1. **华为昇腾 NPU**
   - 达芬奇架构
   - 昇腾AI处理器
2. **谷歌 TPU**
   - TPU 核心脉动阵列
   - TPU 系列架构
3. **特斯拉 DOJO**
   - DOJO 架构
4. **国内外其他AI芯片**
   - AI芯片的思考

Talk Overview

I. AI专用处理器

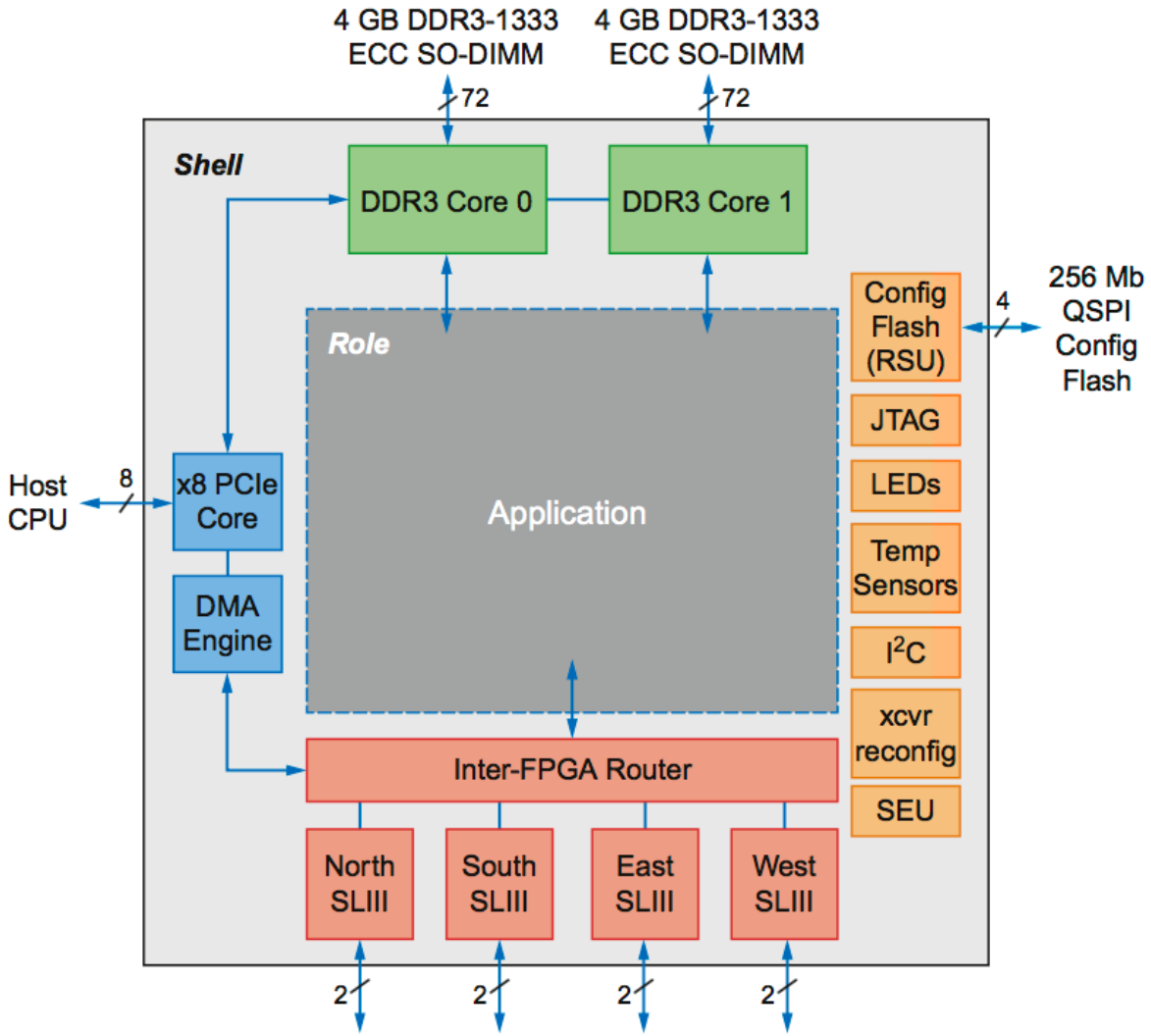
- What is AI Chip - 什么是AI芯片
- Task and Development - AI 芯片任务和部署
- RoadMap of AI Chip - AI 芯片技术路线
- Application scenario - AI芯片应用场景

What is AI Chip

什么是AI芯片

AI 芯片：定义

- **DSA**：特定领域的体系结构，通常被称为加速器，因为与在通用 CPU 上执行整个应用程序相比，它们可以加速某些应用程序。
- **好处**：DSA 可以实现更好的性能，因为它们更贴近应用的实际需求；DSA 例子，包括图形加速单元（即GPU），用于深度学习的神经网络处理器（NPU/TPU），以及软件定义处理器（SDN）。



AI 芯片：定义

- **DSA**：特定领域的体系结构，通常被称为加速器，因为与在通用 CPU 上执行整个应用程序相比，它们可以加速某些应用程序。
- **AI 芯片**：AI 加速器或计算卡，专门用于加速 AI 应用中的大量计算任务的模块。AI 芯片是用于运行 AI 算法的专用处理器，与传统芯片（如CPU）的区别在于专用性或通用性的侧重上。

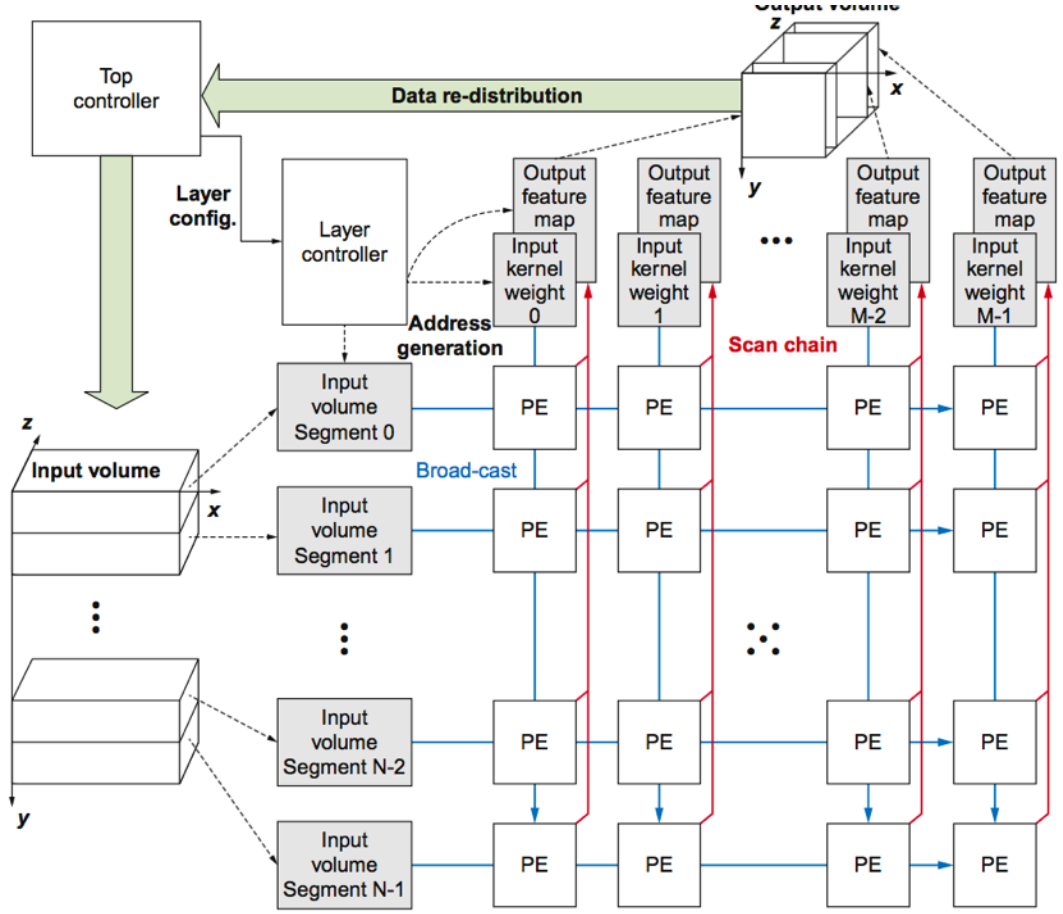
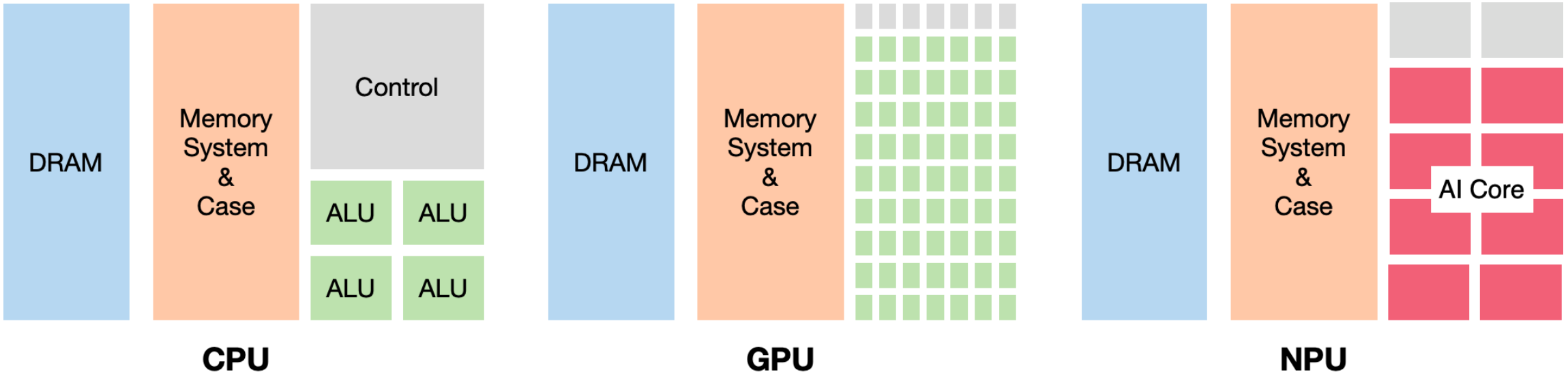


Figure 7.21 CNN Accelerator for Catapult. The Input Volume of the left correspond to Layer[$i-1$] on the left of Figure 7.20, with NumFM[$i-1$] corresponding to y and DimFM[$i-1$] corresponding to z. Output Volume at the top maps to Layer[i], with z mapping to NumFM[i] and DimFM[i] mapping to x. The next figure shows the inside of the Processing Element (PE).

AI Chip vs CPU and GPU

- **AI 芯片**：AI 加速器或计算卡，专门用于加速 AI 应用中的大量计算任务的模块。AI 芯片是用于运行 AI 算法的专用处理器，与传统芯片（如CPU）的区别在于专用性或通用性的侧重上。



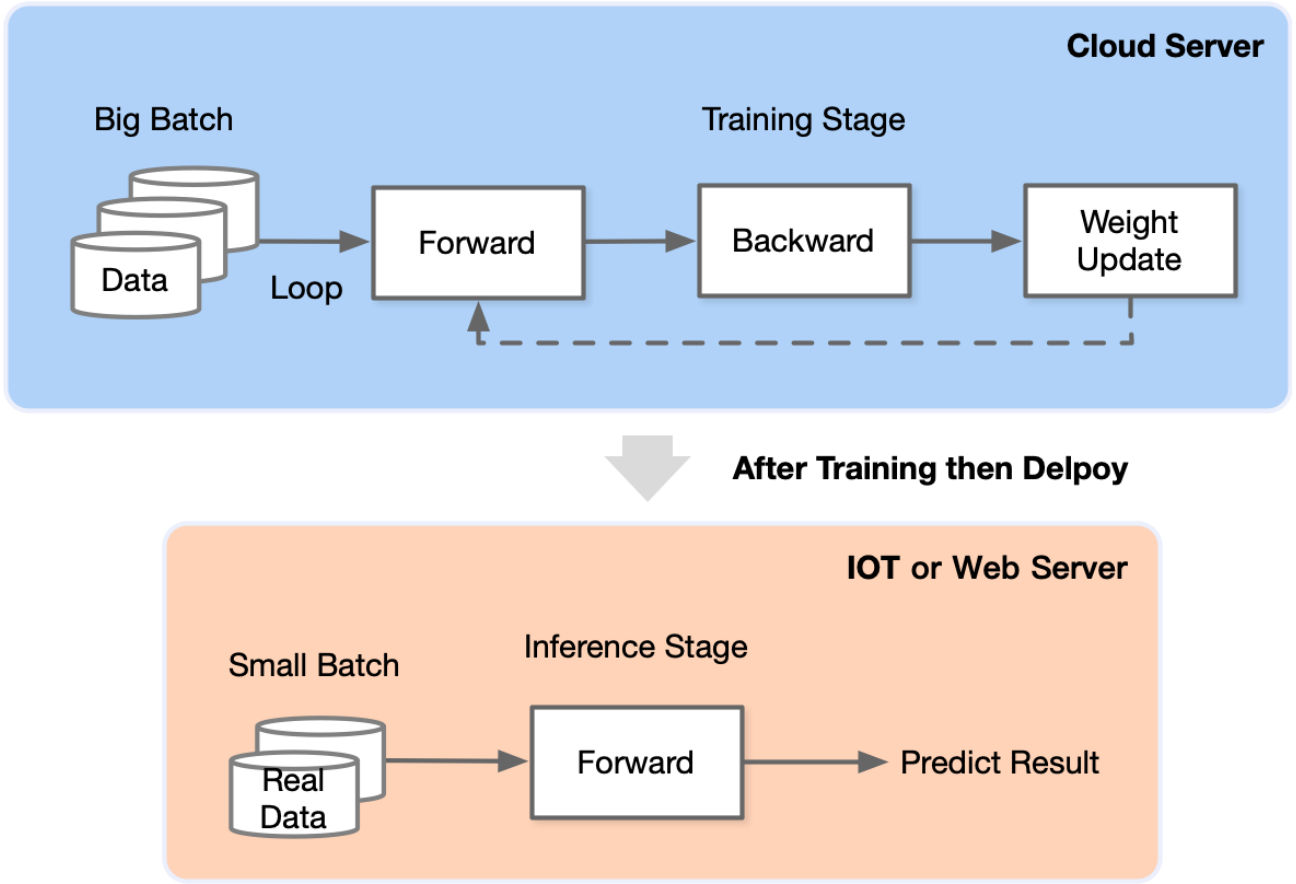
Task & Development

AI 芯片任务和部署

AI 芯片任务：训练 与 推理

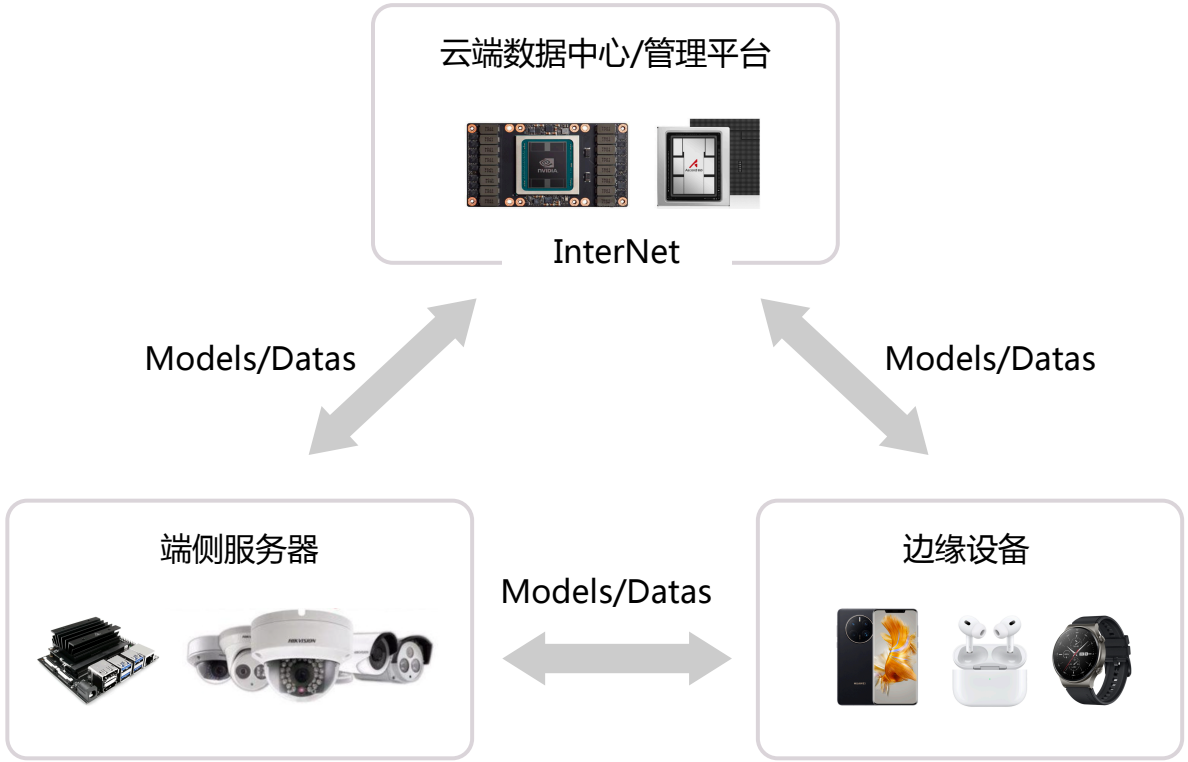
- **训练**：合适的模型结构以及损失函数和优化算法，将数据集以 mini-batch 反复进行前向计算并计算损失，反向计算梯度利用优化函数来更新模型，使得损失函数最小。

- **推理**：训练好的模型结构和参数基础上，一次前向传播得到模型输出过程。最终目标是将训练好的模型部署生产环境中。

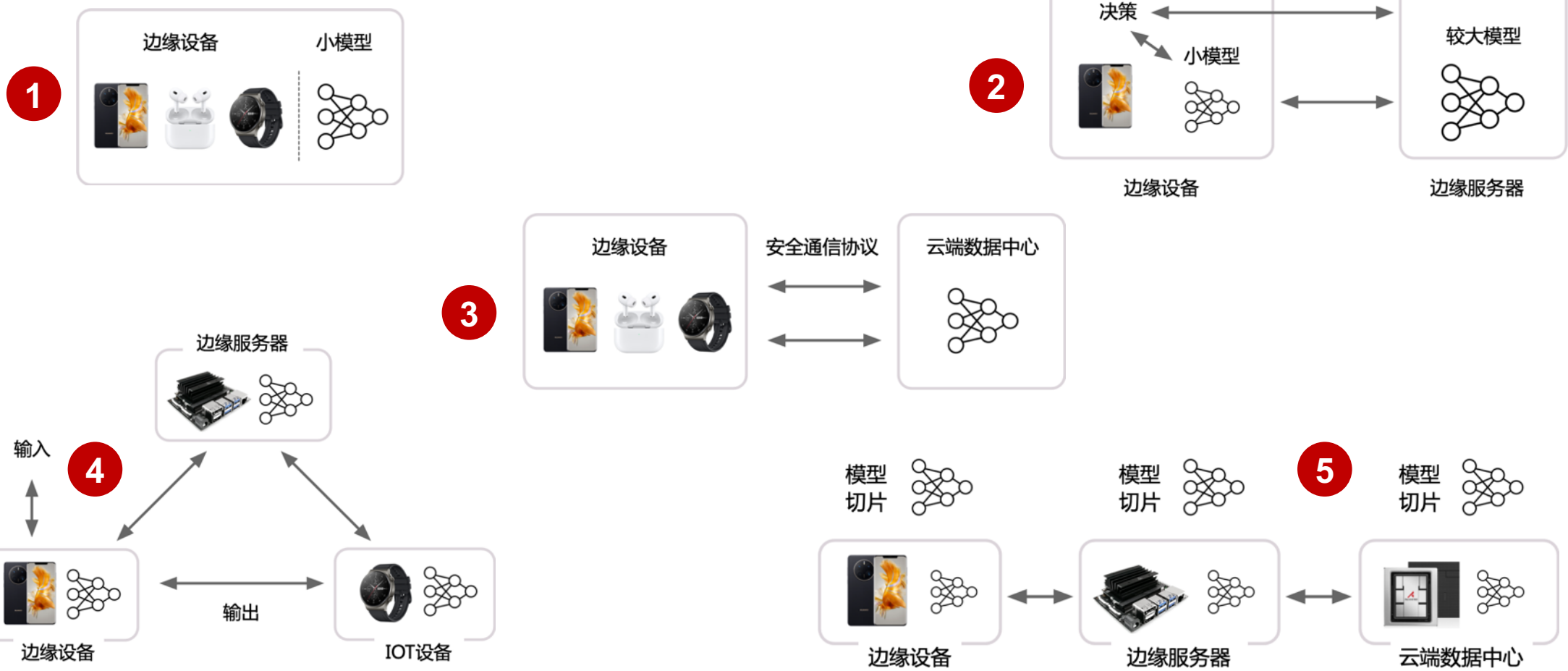


AI 芯片部署方式：云、边、端

- 推理系统一般可以部署在云或者边缘。云端部署的推理系统更像传统 Web 服务，在边缘侧部署的模型更像手机应用和IOT应用系统。



AI 芯片部署方式：云、边、端



RoadMap
AI 芯片技术路线

AI 芯片技术路线

- 作为加速应用的AI芯片，主要的技术路线有三种：**GPU、FPGA、ASIC。**

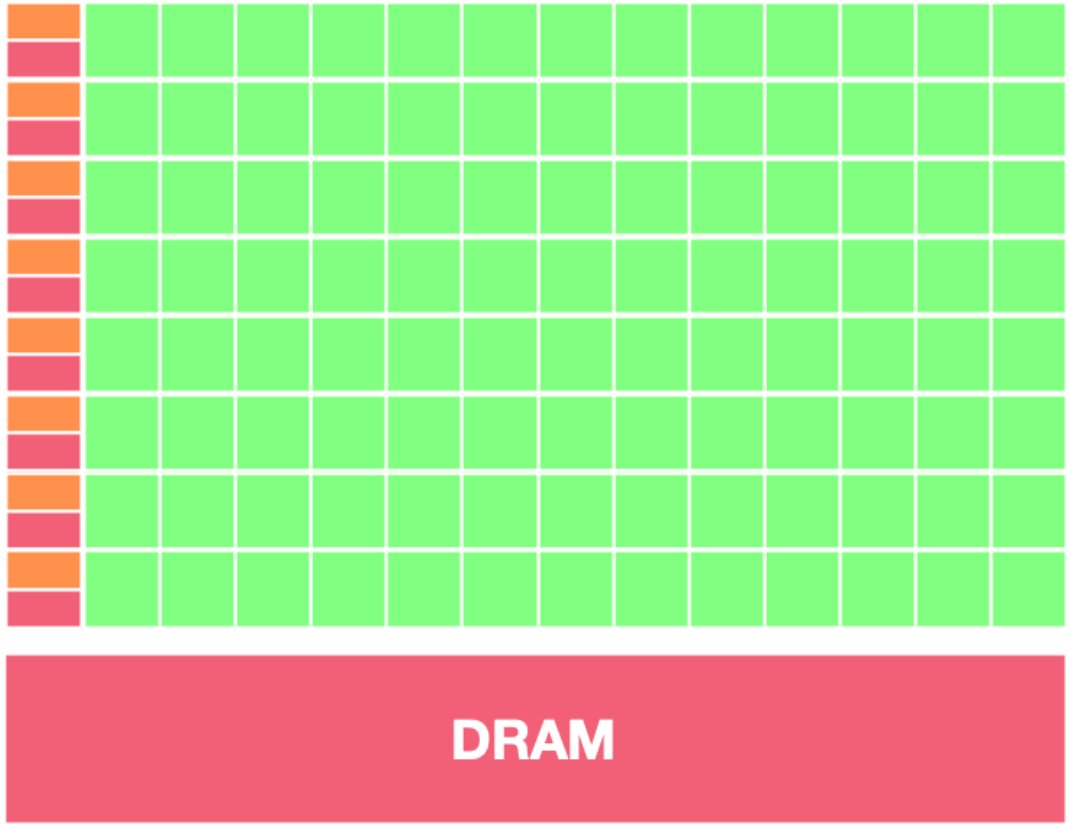
	GPU	FPGA	ASIC
定制化程度	通用	半定制化	定制化
灵活性	好	好	不好
成本	高	较高	低
编程语言	CUDA/OpenCL	Verilog/VHDL，OpenCL/HLS	/
功耗	高	较高	较高
优点	峰值计算能力强，产品成熟	平均性能较高，功耗较低，灵活性强	专用性能强，功耗较低
缺点	整体算力利用率低，功耗高	量产单价高，峰值算力低，上层软件构筑难	上层软件构筑难，针对具体应用，泛化性差
应用场景	云端训练、云端推理	云端推理、终端推理	云端训练与推理、终端推理

GPU

- **图形处理器 GPU（Graphics Processing Unit）**：由大量计算核心组成的大规模并行计算架构，专为同时处理多重任务而设计。
- GPU 专门处理图像计算，包括图像显示、三维渲染等图像并行计算算法，这些算法与深度学习的算法还是有比较大的区别。GPU 非常适合做并行计算，因此可以用来进行 AI 加速。
- GPU 因矩阵计算和并行计算优势，最早在数据中心被用于 AI 计算。由于 GPU 采用并行架构，超过 80% 为运算单元，具备较高性能运算速度。相比较下，CPU 仅有20%为运算单元，更多的是逻辑单元，因此CPU擅长逻辑控制与串行运算，而GPU擅长大规模并行运算。

GPU

- GPU（Graphics Processing Unit），图形处理器

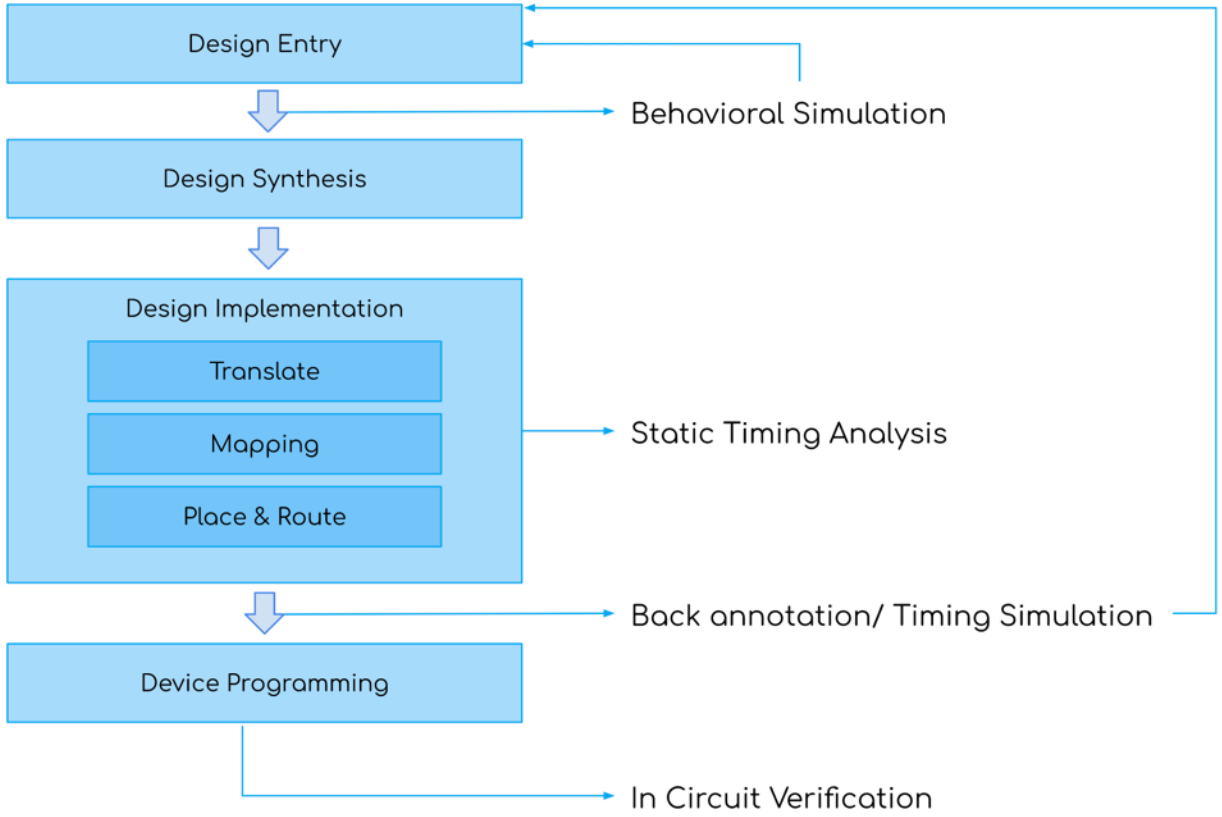


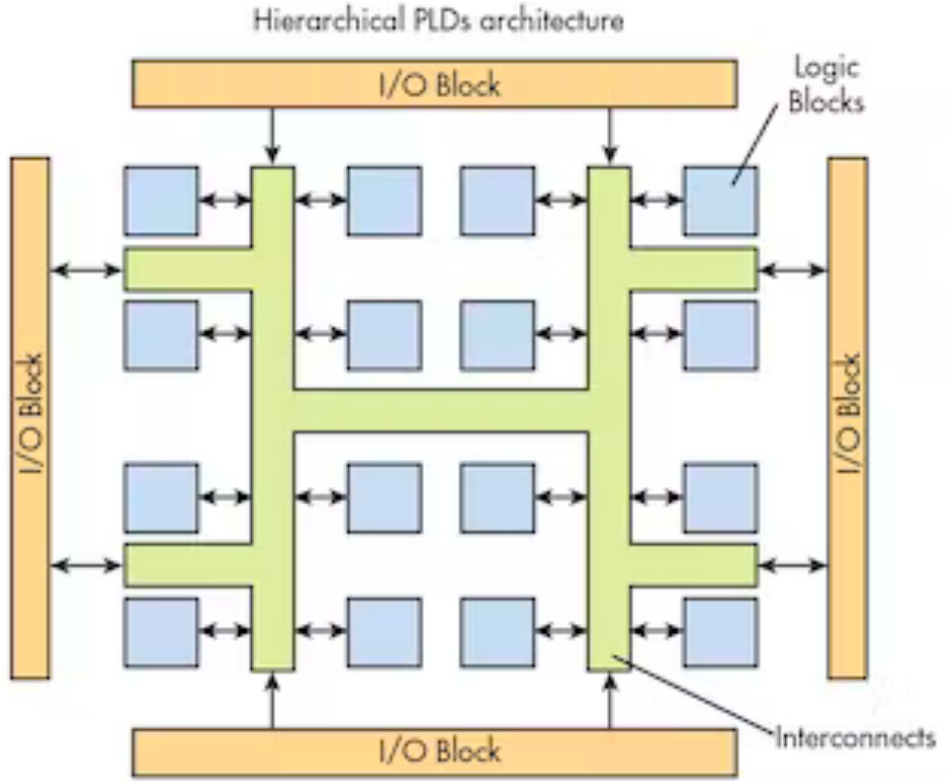
FPGA

- **FPGA（Field-Programmable Gate Array），现场可编程门阵列。**作为专用集成电路领域中的一种半定制电路出现。FPGA利用门电路直接运算，速度快，而用户可以自由定义这些门电路和存储器之间的布线，改变执行方案，以期得到最佳效果。
- FPGA可以采用OpenCL等更高效的编程语言，降低了硬件编程的难度，还可以集成重要的控制功能，整合系统模块，提高了应用的灵活性，与GPU相比，FPGA具备更强的平均计算能力和更低的功耗。
- FPGA适用于多指令，单数据流的分析，与GPU相反，因此常用于推理阶段。FPGA是用硬件实现软件算法，因此在实现复杂算法方面有一定的难度，缺点是价格比较高。

FPGA

- **FPGA（Field-Programmable Gate Array），即现场可编程门阵列。**



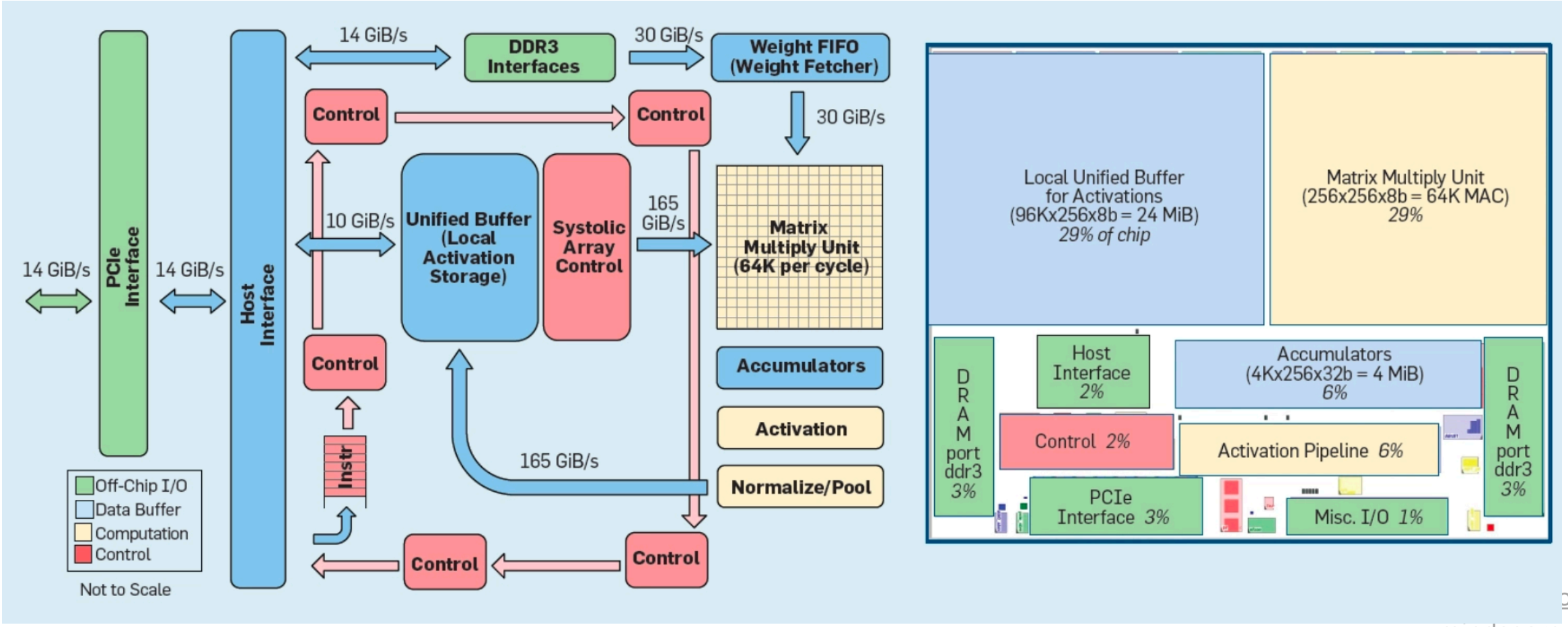


ASIC

- **ASIC（Application Specific Integrated Circuits），即专用集成电路**，是一种为专用目的设计，面向特定用户需求的定制芯片，在大规模量产的情况下具备性能更强、体积更小、功耗更低、成本更低、可靠性更高等优点。
- ASIC与GPU和FPGA不同，GPU和FPGA除了是一种技术路线之外，还是实实在在的确定产品，而ASIC就是一种技术路线或者方案，其呈现出的最终形态与功能也是多种多样。

ASIC

- ASIC（Application Specific Integrated Circuits），即专用集成电路。

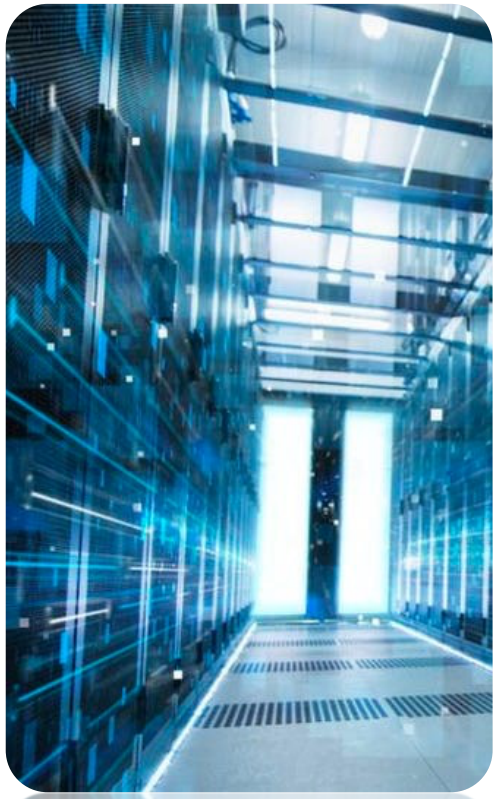


Application scenario

AI芯片应用场景

芯片场景

AI 计算中心



自动驾驶

安防应用

IOT AI应用

芯片场景

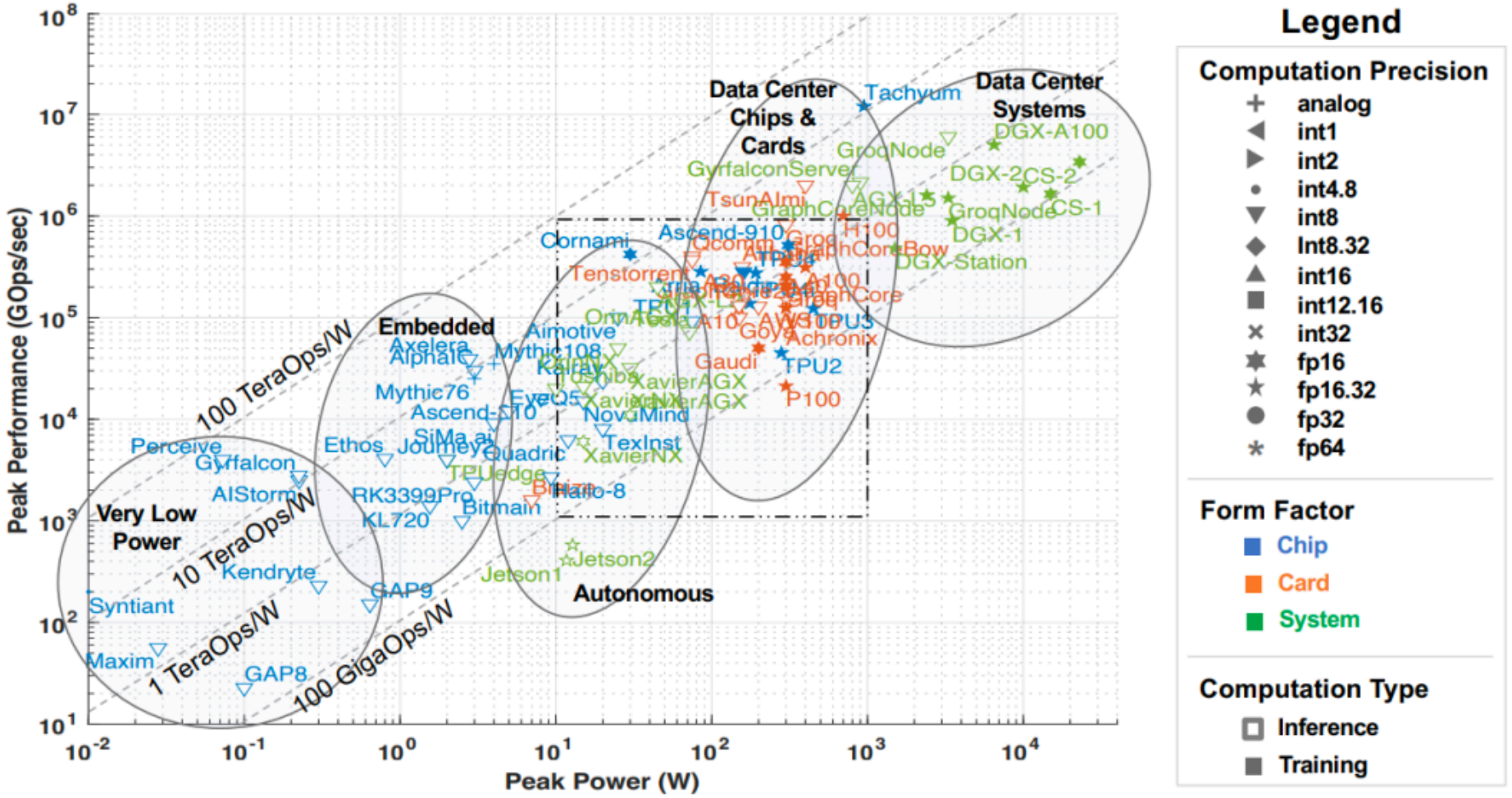


Fig. 2: Peak performance vs. power scatter plot of publicly announced AI accelerators and processors.

芯片场景

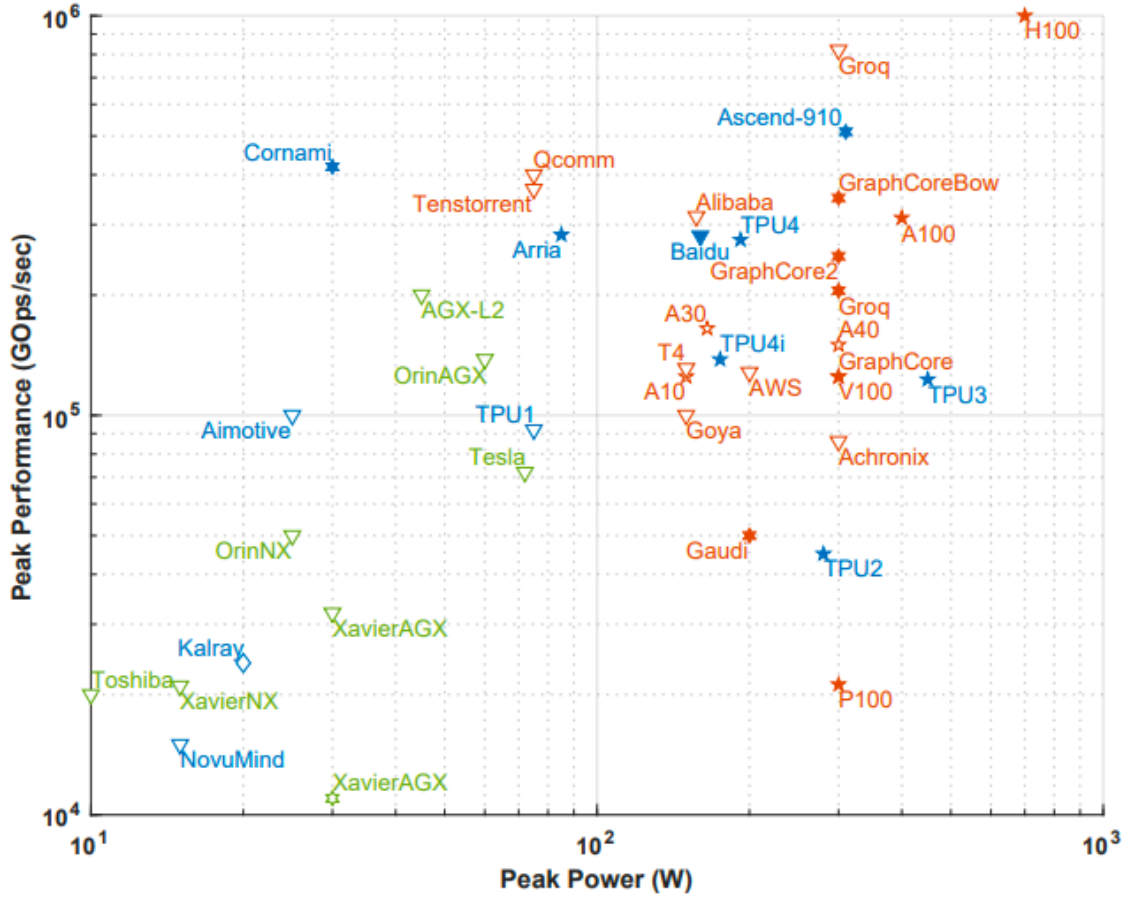


Fig. 3: Zoomed region of peak performance vs. power scatter plot.

芯片场景总结

1. Int8 是嵌入式、自主和数据中心推理应用的默认数值精度。对于大多数AI/ML应用程序，int8精度够用了。但是优部分加速器仍然使用 fp16/bf16进行训练或者推理。

2. 在数据中心领域，密度变得非常拥挤。在过去的几年里，NVIDIA、Huawei、Google等头部厂商都相续发布了对应AI计算中心的AI训练加速器。在计算中心中，为了突破互联带宽 PCIe v4 300W的功率限制，PCIe v5备受期待。

3. 各大厂商针对训练芯片发布了令人印象深刻的性能数据，而且这些公司还宣布了高度可扩展的互联技术，可以将数千张卡片连接在一起。这对于像 Cerebras、GraphCore、Groq、Tesla Dojo的数据流加速器尤其重要，这些加速器是显式/静态编程，或者是路由到计算硬件上的。互联技术使这些加速器能够适应像 ChatGPT 这样的千亿参数大的模型。

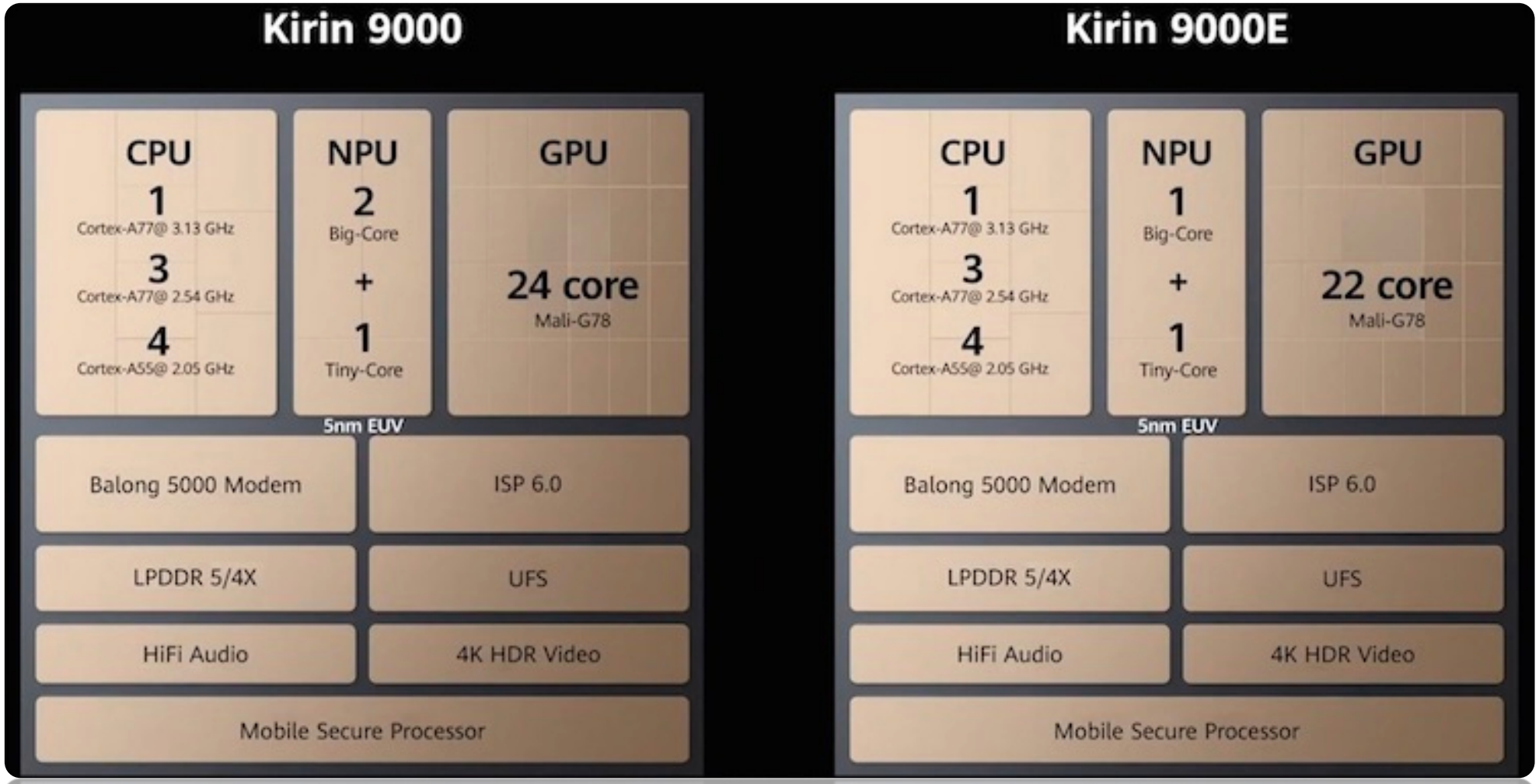
IOT AI应用

- 从2017年开始，苹果、华为海思、高通、联发科等主要芯片厂商相继发布支持AI加速功能的新一代芯片，AI芯片逐渐向中端产品渗透。AI 应用和 AI 部署也将会呈现更多的结合性态。

IOT AI应用

- 从2017年开始，苹果、华为海思、高通、联发科等主要芯片厂商相继发布支持AI加速功能的新一代芯片，AI芯片逐渐向中端产品渗透。AI 应用和 AI 部署也将会呈现更多的结合性态。

IOT AI应用



Talk Overview

1. AI 计算体系
   - 深度学习计算模式
   - 计算体系与矩阵运算
2. **AI 芯片基础**
   - 通用处理器 CPU
   - 从数据看 CPU 计算
   - 通用图形处理器 GPU
   - AI专用处理器 NPU/TPU
   - 计算体系架构的黄金10年

1. **华为昇腾 NPU**
   - 达芬奇架构
   - 昇腾AI处理器
2. **谷歌 TPU**
   - TPU 核心脉动阵列
   - TPU 系列架构
3. **特斯拉 DOJO**
   - DOJO 架构
4. **国内外其他AI芯片**
   - AI芯片的思考

Reference

1. https://caiwuchu.seu.edu.cn/
2. https://coqube.com/fpga-design-flow/