人工智能芯片设计导论

本页PDF

题型

- 填空
- 简答
- 一道大题

面向AI的处理器设计

世界三大尖端技术：空间技术、能源技术、人工智能

人工智能广泛应用的领域：人脸识别、智能医疗、智能语音、自动驾驶、智能机器人

CNN & IP

常见的CNN：GoogleNet、VGG16、AlexNet、DenseNet、MobileNet、ResNet等

CNN由很多层组成

- 卷积层：卷积运算的目的是提取输入的不同特征
- 池化层：对输入的特征图进行压缩，使特征图变小，简化计算；另一方面进行特征压缩，取其最大、最小或平均值，得到新的、维度较小的特征
- 线性整流层：引入非线性，引入非线性之后就可以逼近任意函数
- 全连接层：把所有局部特征结合变成全局特征，用来计算最后每一类的得分

CNN计算特点

- 操作相对固定
- 计算量大
- 需要的带宽大

CNN IP方案分类

- GPU类
- DSP类
- ASIC类
- ASIP类

GPU

特点

- 多核并行计算，核心数多，可支撑大量数据并行计算
- 拥有更高的访存速度
- 更高的浮点数运算能力
- GPU在深度学习领域（训练方面）特别适合

框架：Tensorflow、Caffe等

NVIDIA GPU 可以通过**PCI-e 接口**直接部署在服务器中

DSP - Cadence VP6

特定

- 指令丰富，专门定制了针对深度学习的指令
- 提供深度学习库
- VLIW
- SIMD
- 256MACs

ASIC - NVDLA

ASIP

寒武纪

- 专门针对某种特定的应用和算法定制的处理器，自定义指令集
- 具备**ASIC的高效性和DSP的灵活性**

云天励非NNP

- NU
  - CNN处理核心
  - 多个基本处理单元PE组成
  - 15级流水
- CU
  - 指令广播，任务调度
  - 8级流水

ASIP处理器架构设计

步骤

1. 算法需求分析
2. 软硬件切割
3. 架构定义
4. 指令集定义
5. 指令集模拟器开发（ISS）
6. ISS仿真 & 架构优化迭代
7. 确定微架构和指令集，进入开发阶段

需求分析

算法需求分析

- AI算法基本流程，使用哪些CNN模型
- CNN模型中需要哪些层、基本操作、特殊操作
- 操作是否有通用性和扩展性
- ASIP是否能独立完成算法，是否需要其他模块的配合

产品需求分析

- 产品应用场景

- ASIP处理能力
- 系统能提供多少带宽，需要多少带宽

架构定义

AI ASIP处理器架构设计重点

- 数据重用
- WEIGHT重用
- 计算并行度
- MAC利用率
  - IO是瓶颈
  - 数据重用、WEIGHT重用、降低DDR带宽、提高计算并行度都是**为了提高MAC利用率**
- DDR带宽
  - 有效的数据和WEIGHT重用能够降低DDR的带宽需求
  - 数据压缩
  - 模型压缩
- 硬件扩展性
- 初步流水时序
- 后端可实现性

指令集定义

两类指令

- 循环、跳转、算术逻辑运算等**基本指令**用于控制和NN Core调度
- **NN指令**用于CNN计算

指令集的规整以及扩展性

指令集的颗粒度

- 大颗粒度：效率高
- 小颗粒度：灵活

指令集模拟器开发（ISS）

ISS：指令集模拟器，使用高级语言编写的 ASIP 处理器模型

需要注意：

- 重要参数可灵活配置调整。例如：Memory大小、Buffer的大小、MAC个数、DDR带宽
- 丰富的 profiling 能力。例如MAC的利用率、Buffer/Mac是否饥饿
- 基本能够反映出硬件真实性能

有以下几类

- behavior model
  - 只模拟处理器的行为，无时序概念
  - 用于ASIP处理器架构设计阶段，用于快速建模，也可以作为RTL验证阶段的参考模型
- cycle accurate model
  - 有时序概念
  - 主要作为后续软件开发的参考模型，能够反映出软件在ASIP处理器上运行的真实情况

ISS仿真

1. 选择Benchmark
2. 利用Benchmark进行ISS架构仿真，并进行迭代
    - 检查指令集是否完备、是否过设计、是否便于编程；
    - 确定 Memory/Buffer size；
    - 统计不同 Benchmark 下的 MAC 利用率；
    - 判断常用操作是否足够高效、是否需要优化；
    - 确认频率、算力、带宽是否达到需求指标

ASIP处理器实现难点

RTL开发过程的注意事项

- 时序
    - 合理流水线切割，保证每一级流水时序平衡
    - 合理的 memory 切割，保证 memory 的读写时序
    - 良好的 coding style
- 功耗
    - memory：不读写的时候关闭时钟、有些 memroy 具有低功耗控制端口
    - clock gating：工具自动插入、手动插入控制整个模块的 clock
    - data gating：防止不使用的模块或者组合逻辑输入数据翻转
- 与后端密切迭代
    - 面向AI的ASIP处理器，由于计算资源较多，走线复杂，通常会遇到congestion的问题
    - Memory size 以及 Memory 的块数也会影响后端 floorplan、走线、DFT等
    - ASIP处理器 RTL 开发初期甚至微架构设计时就要和后端建立良好的沟通机制，充分考虑后端实现
- 处理器设计要考虑有效的 debug 机制
    - 侵入式
        - 需要支持基本调试指令
        - 通过JTAG访问memory以及内部寄存器
    - 非侵入式
        - 通过有效的 trace 机制，记录处理器正常运行期间的事件以及状态
- 处理器的设计中要设计合理的profile机制

验证过程中的难点

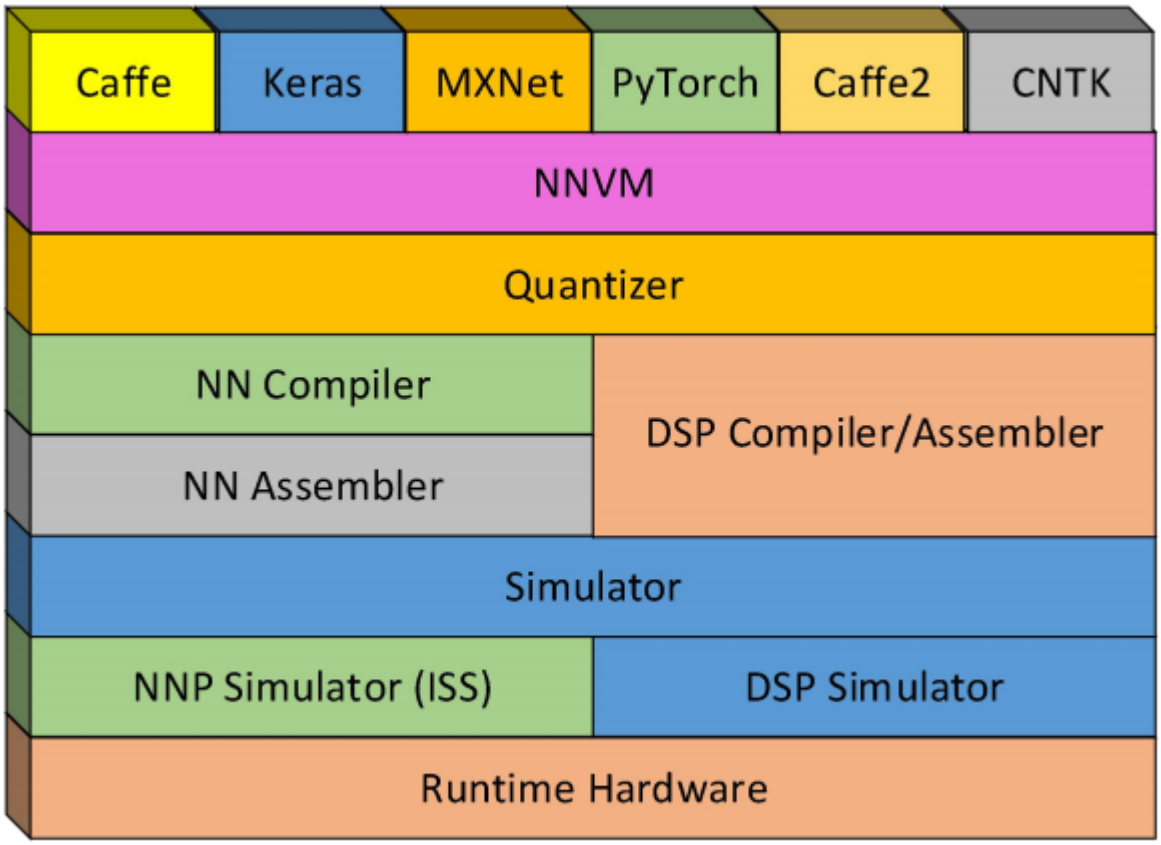
- 指令集灵活，每个 bit 域有很多种变化，加上处理器指令组合太多，验证难度大
- CNN计算中，层数多，每层级联计算多，定位问题困难

物理实现的难点

- memory size 过大，导致 memory timing 有问题
- memory 碎片化严重，导致 Floorplan、功耗、DFT等各种问题
- CNN的**计算特点是计算比较密集**，导致功耗和IR DROP问题
- AI ASIP处理器一般来说 MAC 数较多，这种情况下会导致连线异常复杂，导致 congestion 问题

解决方法 - memory size 过大，前端可以通过 memory 切割缓解部分 timing 问题 - 前端尽量避免有很多的小块memory设计 - 前端做好低功耗设计，控制clock、memory、计算单元的功耗 - 后端优化 Power Mesh 策略 - Floorplan 按照数据流来摆放

ASIP处理器配套工具链



中国AI芯片行业研究报告

人工智能芯片发展的**主要驱动力**

- 政府扶持
- 市场需求

基础理论、关键设备仍落后于国际一流水平

2021年十四五规划：我国新一代人工智能产业将着重前沿寄出**理论突破，专用芯片研发**，构建开源算法平台，并在学习推理与决策、图像图形等重点领域进行创新，聚焦高端芯片等关键领域

芯片最多的应用：计算机视觉、机器人、自然语言处理、机器学习、生物识别

可以分为

- 应用层
- 技术层
- 基础层
- 理论/算法/架构/材料

AI芯片的实现包括**软件和硬件**两方面

AI芯片发展依赖的**两个领域**：

- 模仿人脑建立的数学模型与算法
- 半导体集成电路即芯片

人工智能产业的主要驱动力：结合场景的应用落地

人工智能于芯片的发展分为三个阶段：

- 第一阶段由于芯片算力不足，神经网络算法未能落地
- 第二阶段芯片算力提升，但仍无法满足神经网络算法需求
- 第三阶段，GPU和新架构的AI芯片促进了人工智能的落地。

芯片解读

广义的AI芯片：专门用于处理人工智能应用中大量计算任务的模块，即面向人工智能领域的芯片均被称为AI芯片

狭义的AI芯片：针对人工智能算法做了特殊加速设计的芯片

AI芯片根据技术架构可以分为

- GPU
- FPGA
- ASIC
- 类脑芯片

AI芯片根据其在网络中的位置可以分为

- 云端AI芯片
- 边缘AI芯片
- 终端AI芯片

根据其在实践中的目标可以分为

- 训练芯片
- 推理芯片

云端主要部署训练和推理芯片，边缘主要部署推理芯片

技术架构种类	定制化程度	可编辑性	算力	价格	优点	缺点	应用场景
GPU	通用型	不可编辑	中	高	通用性较强且适合大规模并行运算；设计和制造工艺成熟	并行运算能力在推理端无法完全发挥	高级复杂算法和通用性人工智能平台
FPGA	半定制化	容易编辑	高	中	可通过编程灵活配置芯片架构适应算法迭代，平均性能较高；功耗较低；开发时间较短（6个月）	量产单价高；峰值计算能力较低；硬件编程困难	适用于各种具体的行业
ASIC	全定制化	难以编辑	高	低	通过算法固化实现极致的性能和能效、平均性很强；功耗很低；体积小；量产后成本最低	前期投入成本高；研发时间长（1年）；技术风险大	当客户处在某个特殊场景，可以为其独立设计一套专业智能算法软件
类脑芯片（不同的技术路线）	模拟人脑	不可编辑	高	-	最低功耗；通信效率高；认知能力强	目前仍处于探索阶段	适用于各种具体的行业

GPU

图形处理器

适用场景

- 运算密集

- 高度并行
- 控制简单
- 分多个阶段执行

开发环境

- CG
- CUDA
- ATIStream
- OpenCL

GPU体系的发展趋势

- 增加计算资源密度
- 提高存储体系性能和功能
- 增强通信能力和可靠性
- 降低功耗

GPU和CPU的对比

	CPU	GPU
浮点计算能力	1	~ 10
运算方式	串行	并行
带宽	内存带宽小	高显存带宽
延迟	通过大的缓存保证访问内存的低延迟。	直接访问显存因此延时较长。

ASIC与FPGA

二者对比

ASIC&FPGA总体对比

	FPGA	ASIC
运算速度	较低，FPGA结构上的通用性必然导致冗余；另外，不不同结构间的时延也不可忽略。	较高，结构上无特殊限制，设计时也可将特定模块靠近减少延迟
芯片规模	实现相同的功能时，需要更大的FPGA	实现相同的功能时，ASIC的规模更小
功耗	相同工艺条件下，功耗更大	相同工艺条件下，功耗更小
成本	几乎无开发工具和风险，主要成本都在单片上。	由于进入生产后硬件不可更改，开发工具和流片过程可能产生大量成本
运行过程	加载配置进入存储器需要时间	可立即运行
产品定位	适用于项目产品需要灵活变动等方面的产品及产品要求快速占领市场的情况	适用于设计规模较大，或应用成熟的产品如消费电子等
发展方向	大容量、低电压、低功耗、SoC	更大规模、IP复用技术、SoC

FPGA：可编程逻辑门阵列，是一种可重构芯片

- 具有开发周期短，上市速度快，可配置性等特点
- 应用在大型企业的**线上数据处理中心和军工单位**
- 特点
  - 可加速上市进程
  - 非提前支付的一次性开支
  - 简化的设计周期
  - 更具预测性的项目周期
  - 现场可重编功能

ASIC：是集成电路，从性能、能效、成本均极大的超越了标准芯片

- 分为全定制和半定制
- 应用偏向**消费电子**
- 特点
  - 完整的定制功能与更小的尺寸
  - 更低的器件成本
  - 高性能、低功耗

    - 可形成IP核复用

SoC

SoC 是片上系统，以嵌入式系统为核心，以 IP 复用技术为基础，集软硬件于一体的集成芯片

SoC芯片制造流程：设计、制造、封装、测试

优势：

- 降低耗电量
- 减少体积
- 丰富系统功能
- 提高速度
- 节省成本

产业发展趋势：

- 平台化设计
- 供应链之间合作加强
- 分工将更加明确

类脑芯片

DNN（深度神经网络） 与 SNN（脉冲神经网络） 对比

	DNN	SNN
训练方式	需大量数据	单个数据样本
学习方式	监督学习	无监督学习
输入类型	图像帧或数据阵列	脉冲
时延	高	极低（接近实时）
神经元模型复杂程度	低	高
功耗	由处理器与储存器存取决定	由每个事件功耗决定
分类精度	较高	较低
分类速度	低	高较
研究阶段	较成熟	探索及部分小规模试用阶段

类脑芯片的硬件实现方式

- 忆阻器
- 自旋电子器件
- 光子器件
- 电化学器件
- 二维材料

AI芯片发展

计算范式：

- 存内计算
- 模拟计算
- 量子计算

应用层面

算力向边缘测移动，专注于特殊场景的优化

应用领域包括

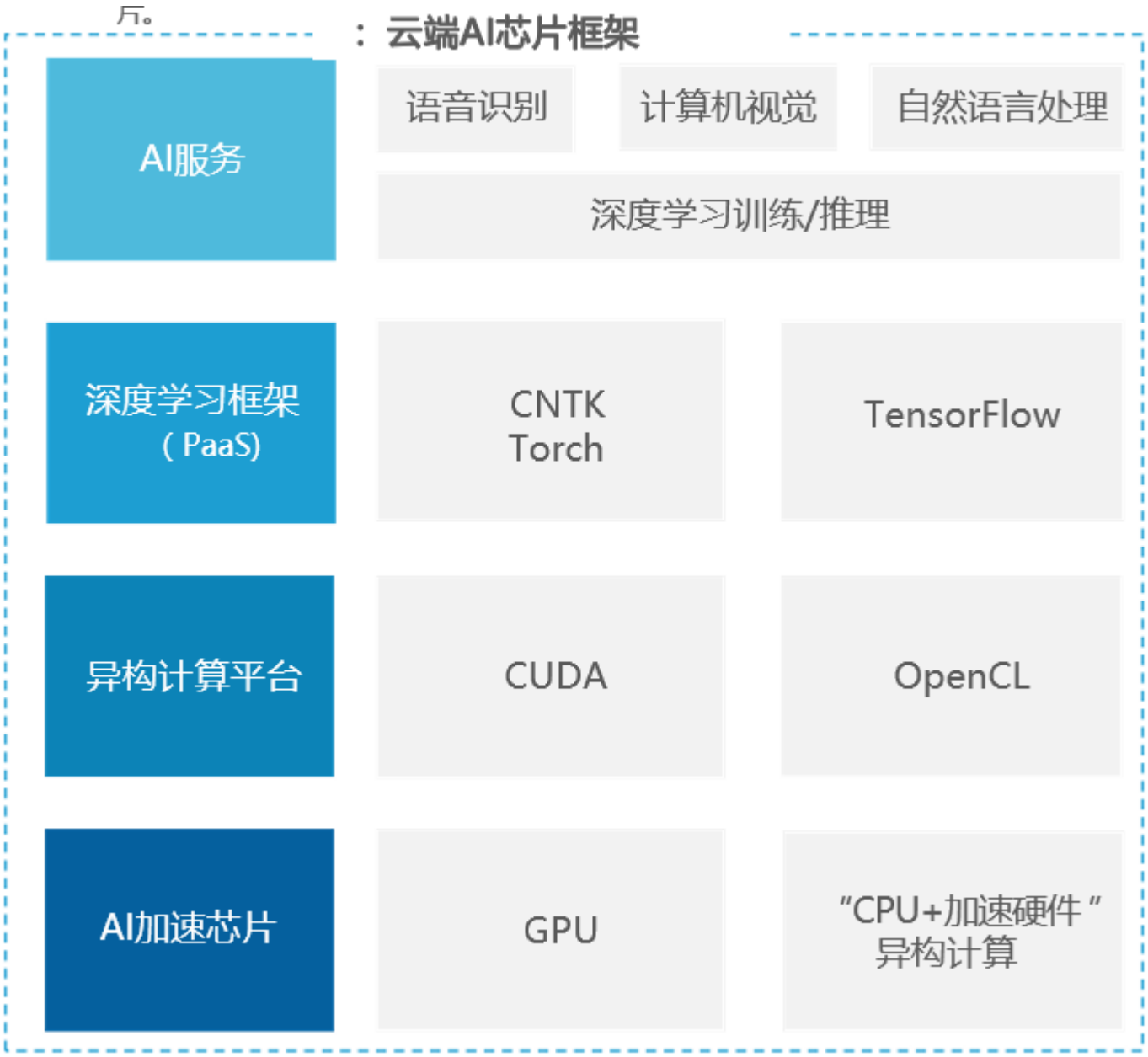
- 云端训练

- 云端推理
- 边缘计算
- 终端设备

云端训练可部署GPU/ASIC，其他领域可部署GPU/ASIC/FPGA

云端

云端AI芯片框架



云仍是AI的中心，更**高性能的计算芯片以及新的AI学习架构**是解决AI训练和推理工作负载问题的关键

中国云端AI芯片参与者：芯片设计、芯片代工、IP设计

边缘

数据向边缘下沉，随着行业落地有很大增量

边缘计算：实现数据的存储 、计算与应用

- **发展历程**：技术储备阶段、快速增长阶段、行业落地阶段
- **应用场景**：物联网边缘计算、工业边缘计算、智慧家庭/城市边缘计算、边缘云、多接入边缘计算、广域接入网络边缘计算

边缘计算的价值：CROSS

- C：Connection，连接海量设备
- R：Real-time，业务实时性
- O：Optimization，数据优化
- S：Smart，应用的智能化
- S：Security，安全与隐私保护

典型企业

酷芯微电子：依托**智能感知、智能计算、智能传输**三大核心技术，自主研发芯片架构及核心IP

昆仑芯科技产品优势：经过实践验证、领先的性能、开发环境友好

地平线的BPU处理器

- 高斯架构
- 伯努利架构
- 贝叶斯架构

黑芝麻智能：自动驾驶计算芯片和平台研发企业

SynSense时识科技

中国人工智能芯片行业挑战与机遇

半导体芯片产业链

- 上游支撑企业
- 半导体企业
- 下游应用

数据难以发挥有效价值的原因：

- 数据孤岛
- 数据安全
- 数据质量
- 管理体系

人工智能芯片生态

- 技术生态
- 商业生态

问题

:::tabs @tab:active 面向AI的处理器设计

世界三大尖端技术

人工智能广泛应用的领域

常见的CNN CNN由什么层组成 CNN计算特点 CNN IP方案分类

GPU特点 框架 可以通过什么接口部署在什么

ASIP是什么 特点

ASIP处理器架构设计步骤

AI ASIP处理器架构设计重点

MAC利用率什么是瓶颈

如何降低DDR带宽需求

指令集有哪两类 特性 颗粒度

ISS是什么 用途 需要注意什么 有什么类别 仿真的步骤

ASIP处理器实现难点

RTL开发过程的注意事项

验证过程中的难点

物理实现的难点和解决方法

ASIP处理器配套工具链

@tab 中国AI芯片行业研究报告

人工智能芯片发展的主要驱动力

芯片最多的应用

AI芯片发展依赖的领域

人工智能产业的主要驱动力

什么仍落后于国际一流水平

2021年十四五规划

广义和狭义的AI芯片

AI芯片的实现包括什么

AI芯片根据技术架构可以分为什么

AI芯片根据其在网络中的位置可以分为什么

根据其在实践中的目标可以分为什么

云端和边缘主要部署什么

GPU适用场景 开发环境 发展趋势 和CPU区别

FPGA与ASIC的区别

FPGA是什么 特点 应用

ASIC是什么 特点 应用

SoC是什么 优势是什么 流程 产业发展趋势

类脑芯片硬件实现方式 DNN与SNN对比

应用层的应用概况

芯片应用领域

解决AI训练和推理工作负载问题的关键

中国云端AI芯片参与者

云端AI芯片框架

边缘计算的概念 发展历程 应用场景

边缘计算的价值 分别是什么

数据难以发挥有效价值的原因