先进编译实验室
Advanced Compiler

编译论坛

深度学习硬件平台

嘉宾：韩梅

先进编译实验室
Advanced Compiler

主要内容

深度学习硬件

平台背景

四类

人工智能芯片

深度学习硬件

应用场景

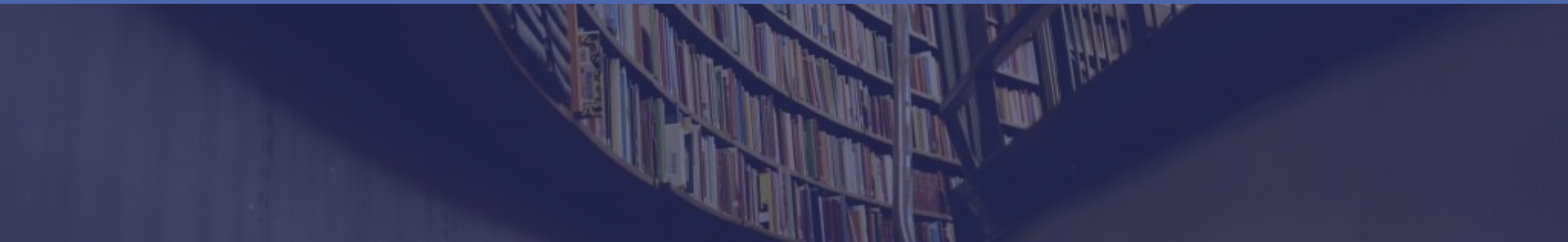
深度学习硬件平台背景

背景background

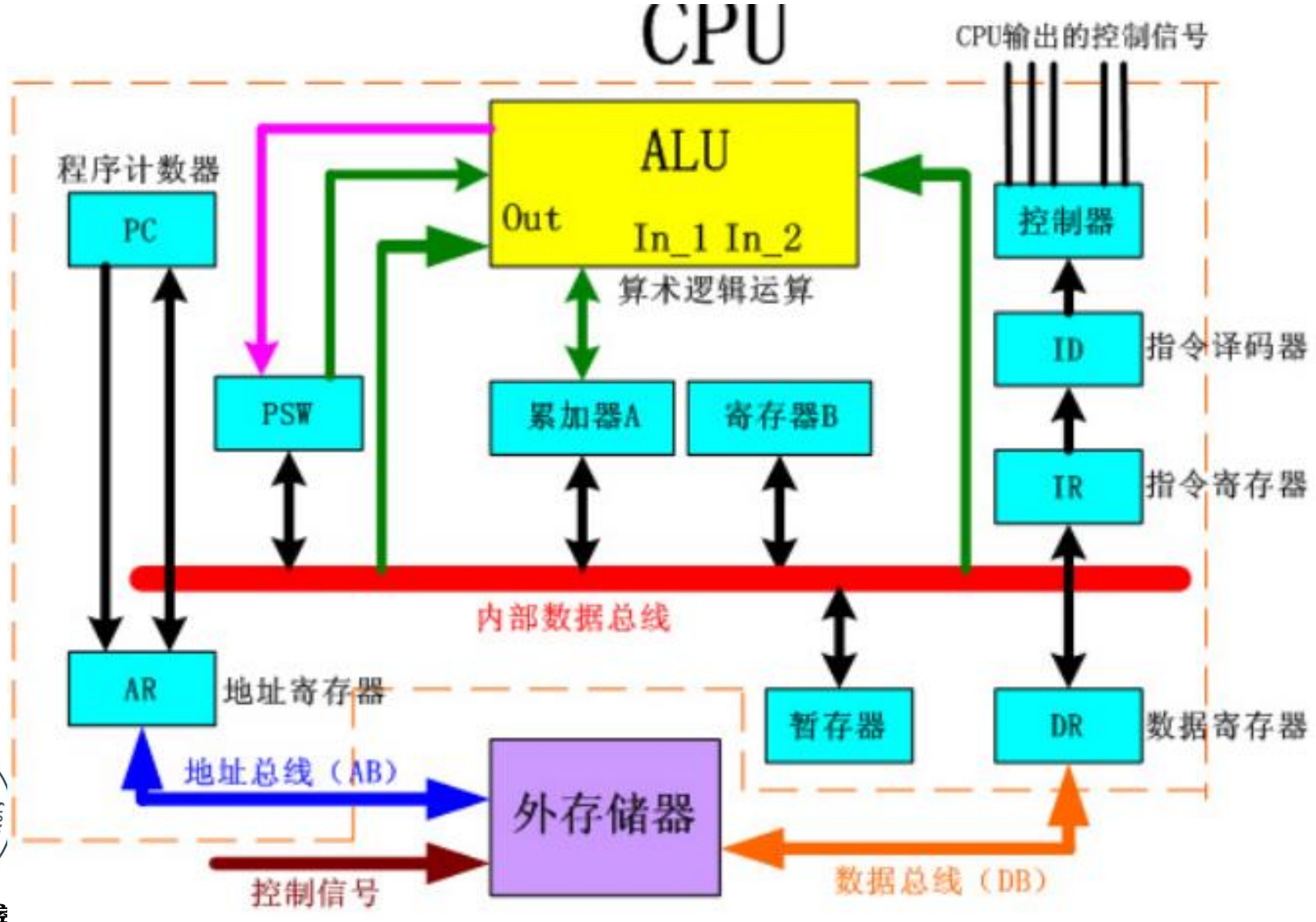
随着深度神经网络模型层数的增加，与之相对应的权重参数成倍地增长，从而对硬件的计算能力、内存带宽及数据存储等有较高的要求。所以必须找到更好的硬件计算加速方案，以满足不断增长的数据量和不断扩大的网络规模。

四类人工智能芯片

GPU　　FGPA　　ASIC　　类脑芯片



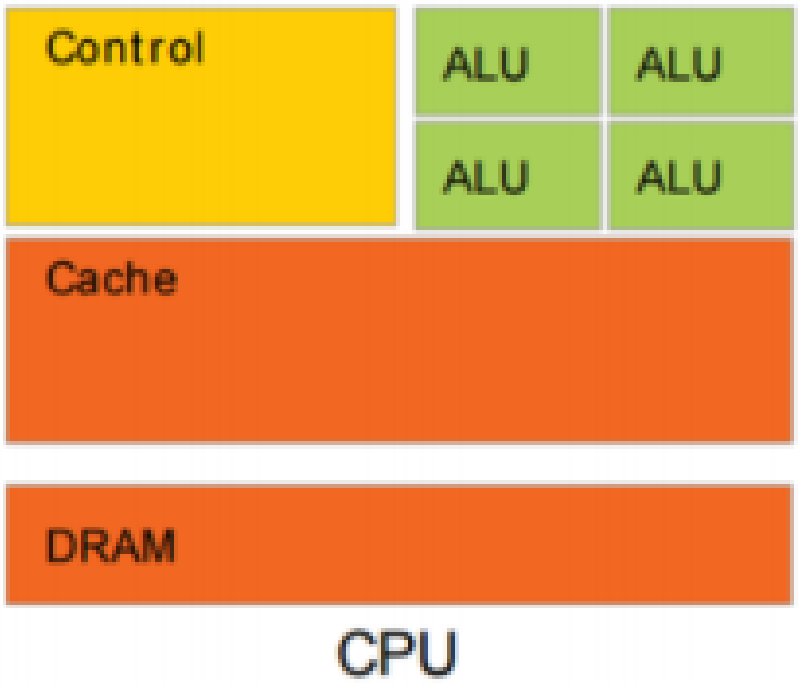
传统cpu局限性

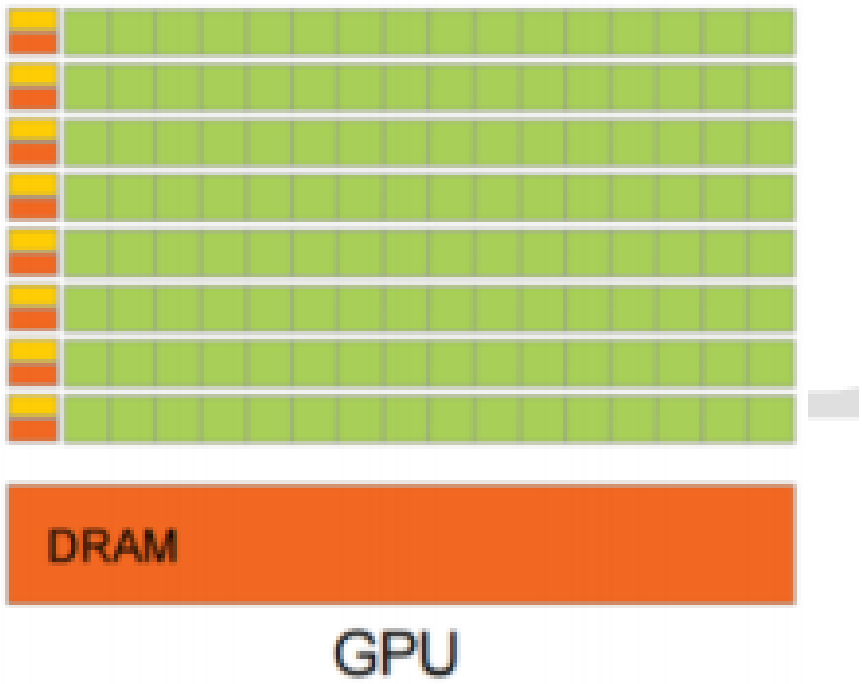


四大类人工智能芯片

根据其技术架构，可分为GPU、FPGA、ASIC及类脑芯片。针对数据训练阶段，被业内广泛接受的是 CPU + GPU 的异构模式。而针对数据推断阶段，则较多地依赖于 CPU + FPGA 或 ASIC。

GPU



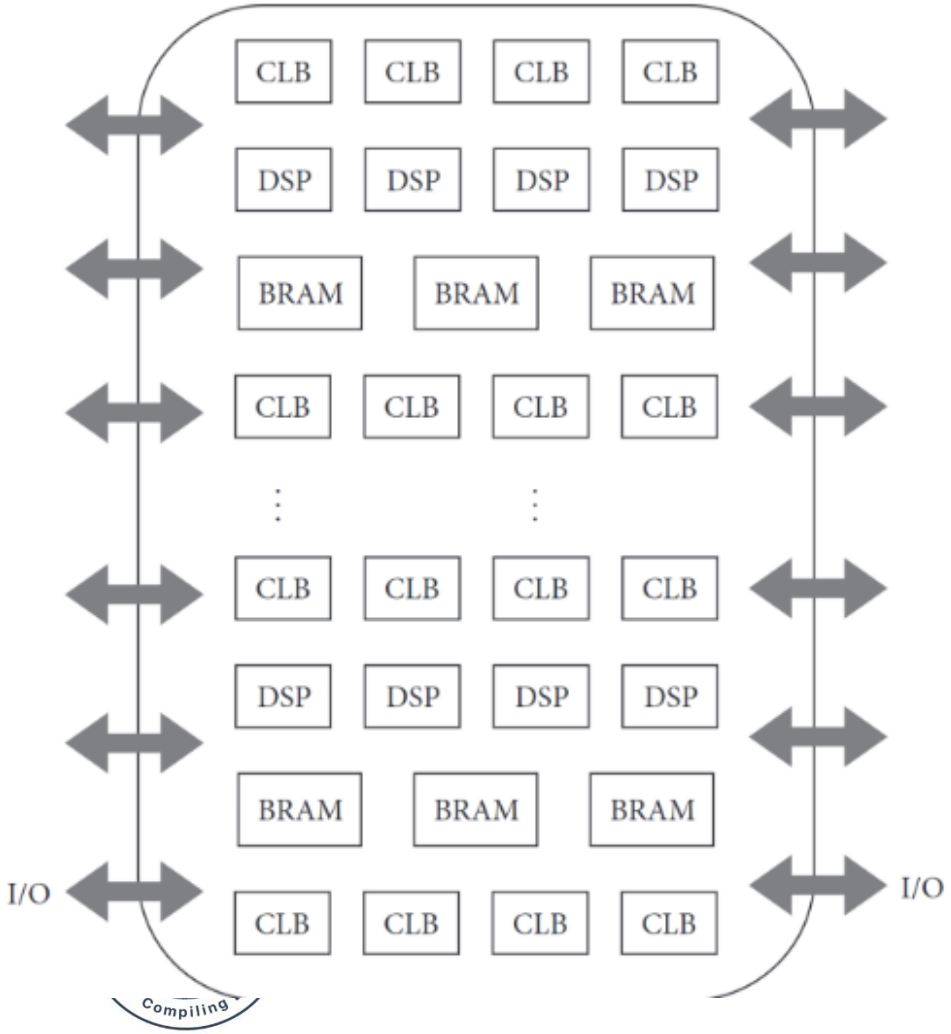


GPU

英伟达在 2006 年推出了统一计算设备构架 CUDA 及对应的 G80 平台，第一次让 GPU 具有可编程性，使得 GPU 的流式处理器除了处理图形也具备处理单精度浮点数的能力。自从 AlexNet在 2012 年的 ImageNet 比赛中取得优异成绩以来，

大量依赖 GPU 运算的深度神经网络软件框架（如TensorFlow 、PyTorch 、Caffe 等）的出现极大地降低了 GPU 的使用难度，因此它也成为人工智能硬件首选，在云端和终端各种场景均被率先应用，也是目前应用范围最广、灵活度最高的 AI 硬件。

FPGA



FPGA全称为可编程逻辑门阵列，是一种“可重构”芯片，主要包含可编程逻辑块、片上储存器及用于连接逻辑模块的可重构互连层次结构。

在计算单元上，FPGA采用了大量的可配置逻辑单元模块(CLB)，这些模块通过查找表（LUT)的方式实现各种功能。

在存储方面，出于灵活性的考虑，通常FPGA在片内提供了很多存储资源，可以配置成不同的形式来使用。

在控制上，FPGA则需要设计者通过配置CLB的方式来控制和使用片内的资源。

FPGA的优点

01

可重构

FPGA 芯片可以被重复编程

02

可定制

FPGA 可以根据应用需求灵活地对数据位宽进行配置

03

低功耗

FPGA平均功耗低于GPU

04

高性能

FPGA 芯片上具有大量的片上存储资源，可以提供强大的带宽和并行访存能力。

FPGA的局限性

基本单元的计算能力有限

硬件编程困难

FPGA 价格较为昂贵

先进编译实验室
Advanced Compiler

全定制化的ASIC

ASIC 是专用集成电路，是指应特定用户要求和特定电子系统的需要而设计、制造的集成电路。ASIC从性能、能效、成本均极大的超越了标准芯片，非常适合AI计算场景，是当前AI公司开发的目标产品之一。

例如近些年类似谷歌的 TPU、寒武纪的 NPU、地平线的 BPU、英特尔的 Nervana、微软的 DPU、百度的 XPU 等芯片，本质上都属于基于特定应用的人工智能算法的 ASIC 定制芯片。

周期长

需要大量设计时间以及验证和物理设计周期，因此需要相对多的上市时间

成本低

量产后 ASIC 的成本会远远低于FPGA 方案

不可更改

ASIC 一旦制造完成将不能更改

ASIC的三大特点

相比于同样工艺 FPGA 实现，ASIC 可以实现 5-10 倍的计算加速，且量产后 ASIC 的成本会大大降低。不同于可编程的 GPU 和 FPGA，ASIC 一旦制造完成将不能更改，因此具有开发成本高、周期长、门槛高等问题。

类脑芯片

类脑芯片不采用经典的冯·诺依曼架构，而是基于神经形态架构设计，算法是脉冲神经网络（SNN)。在基于冯诺依曼结构的计算芯片中，计算模块和存储模块分离处理从而引入了延时及功耗浪费。类脑芯片侧重于仿照人类大脑神经元模型及其信息处理的机制。

它的内存、CPU 和通信部件完全集成在一起。信息的处理完全在本地进行，而且由于本地处理的数据量并不大，传统计算机内存与 CPU 之间的瓶颈不复存在。其中典型的有 IBM 的 TrueNorth、英特尔的 Loihi、高通的 Zeroth、清华大学的天机芯等。

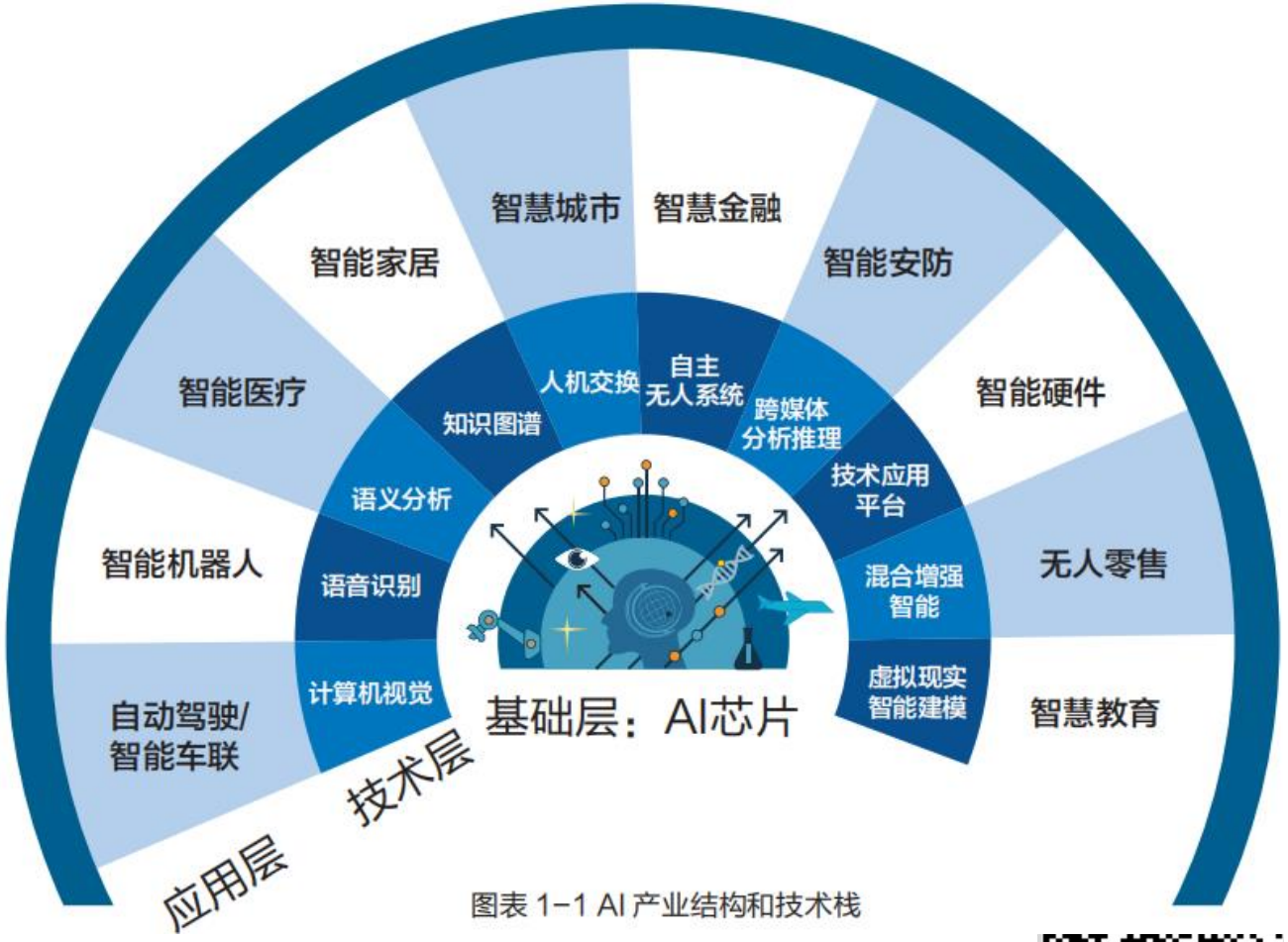
四类技术架构对比

技术架构	定制化程度	可编辑性	算力	价格	优点	缺点
GPU	通用性	不可编辑	中	高	1、通用性较强且适合大规模并行运算； 2、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
FPGA	半定制化	容易编辑	高	中	1、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； 2、功耗较低； 3、开发时间较短（6个月）	1、量产单价高； 2、峰值计算能力较低； 3、硬件编程困难
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	1、前期投入成本高； 2、研发时间长（1年）； 3、技术风险大
类脑芯片	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

深度学习硬件应用场景

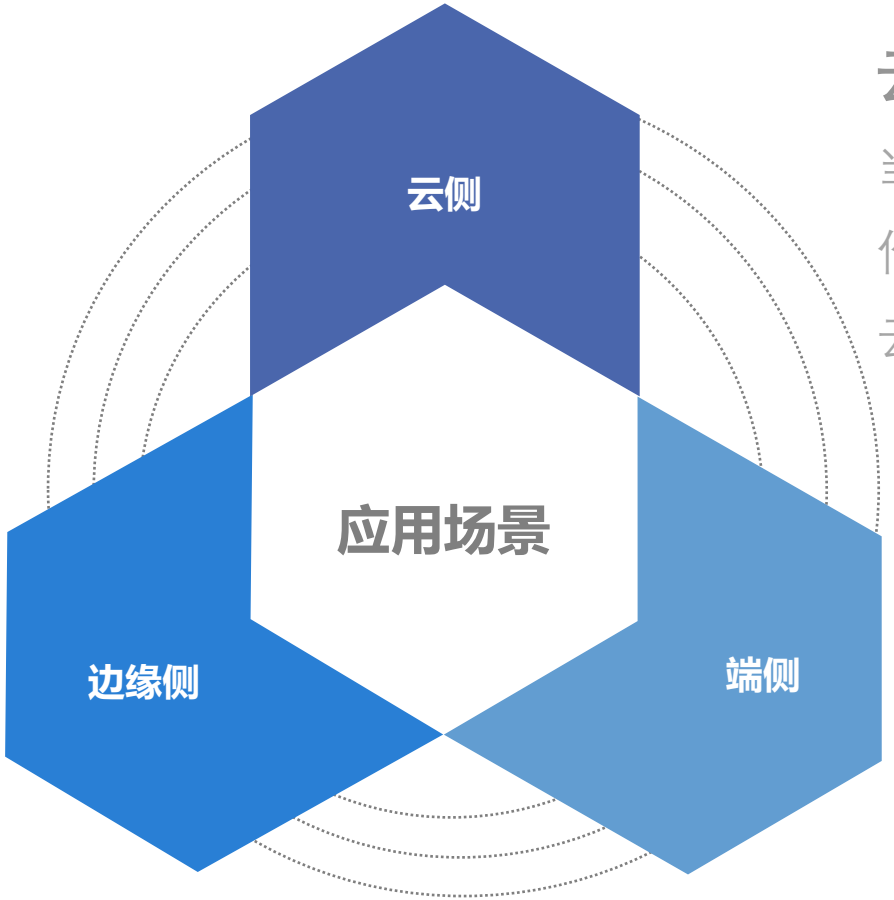
深度学习硬件应用场景

- 随着技术成熟化，AI芯片的应用场景除了在云端及大数据中心，也会随着算力逐渐向边缘侧移动。智能产品种类也日趋丰富。未来，AI计算将无处不在。



图表 1-1 AI 产业结构和技术栈

深度学习硬件应用场景



云端

当前，大多数AI训练和推理工作负载都发生在公共云和私有云中，云仍是AI的中心。

边缘侧

对于某些应用，由于各种原因（如延迟，带宽和隐私问题）必须在边缘节点上执行推断

终端设备

可以第一时间对收集的数据进行处理，极大加快了系统响应也减少了系统处理延迟

深度学习硬件应用场景

云端训练

- **可部署芯片**:
  GPU/GPU/ASIC
- **芯片特征**：高吞吐量、高精确率、可编程性、分布式、可扩展性、高内存与带宽
- **计算能力与功耗**:
  >30TOPS，>50W
- **应用**：云/HPC/数据中心

云端推理

- **可部署芯片**:
  GPU/GPU/ASIC/FPGA
- **芯片特征**：高吞吐量、高精确率、分布式、可扩展性、低延时
- **计算能力与功耗**:
  30TOPS，>50W
- **应用**：云/HPC/数据中心

边缘计算

- **可部署芯片**:
  GPU/GPU/ASIC/FPGA
- **芯片特征**：降低AI计算延迟、可单独部署或与其他设备组合（如5G基站）、可将多个终端用户进行虚拟化、较小的机架空间、扩展性及加速算法
- **计算能力与功耗**:
  5~30TOPS，4~15W
- **应用**：智能制造、智慧家居、智慧交通等、智慧金融等众多领域

终端设备

- **可部署芯片**:
  GPU/GPU/ASIC/FPGA
- **芯片特征**：低功耗、高能效、推理任务为主、较低的吞吐量、低延迟、成本敏感
- **计算能力与功耗**:
  <8TOPS，<5W
- **应用**：各类消费电子，产品形态多样；以及物联网领域

深度学习硬件应用场景

参考文献

1.2022中国人工智能芯片行业研究报告

2.aichip人工智能芯片研究报告2018

3.背景_深度学习相关研究综述_张军阳

4.深度神经网络 FPGA 设计进展、实现与展望_焦礼成

5.智能计算系统_陈云霁等

6.人工智能芯片技术白皮书（2018）

先进编译实验室
Advanced Compiler

分享完毕，感谢大家观看

THANKS

嘉宾：韩梅