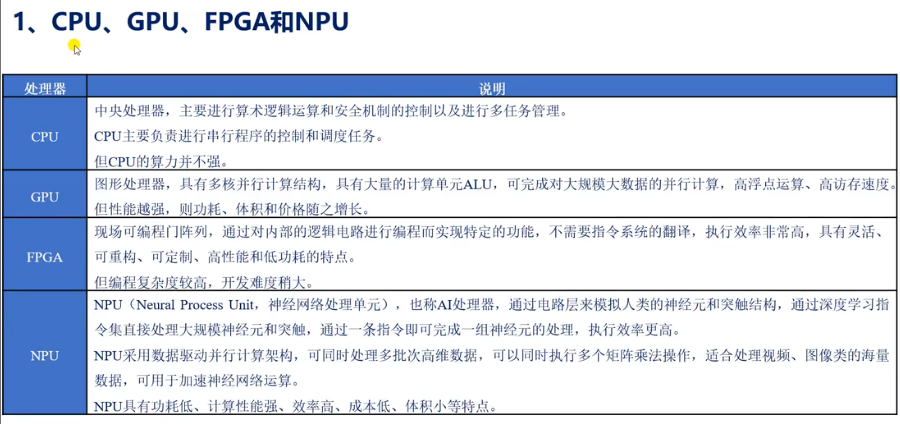
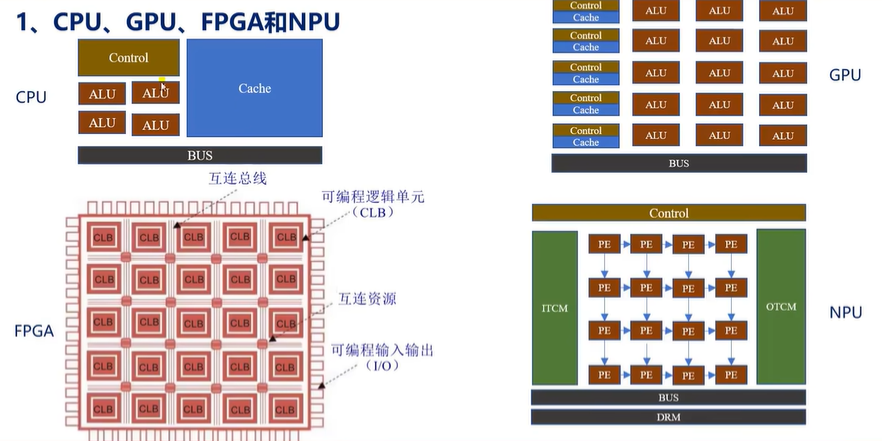


许多芯片自带NPU，用来支持AI运算。芯片自带NPU以及成为亮点之一。

AI模型部署，就是把AI模型放到AI芯片上去跑。



架构图：



1. CPU：深度学习这一块，要处理海量的数据，传统嵌入式处理器的CPU主要就是进行算数逻辑运算，和安全机制的控制，它擅长逻辑控制，主要进行多任务管理。但其算力并不强。无法进行大量矩阵数据的并行计算。因此CPU难以满足海量数据处理的需求了。就是因为，在CPU中神经网络运算时，效率低下。算力跟不上。

可以参考架构图。在CPU的架构图中主要就包括:

控制器（控制单元）Control:主要用于控制CPU的运行和指令的执行

缓冲存储器Cache ：可以存储CPU运行过程中需要频繁访问的数据和指令，缓存Cache大的话，CPU的运行速度和效率就会高。

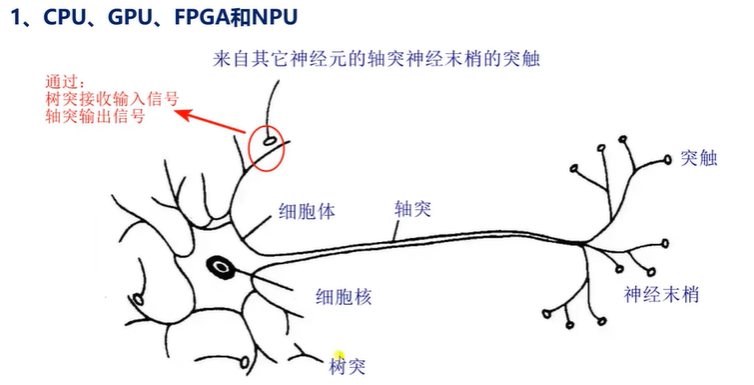
以及 用于计算的计算单元ALU：主要进行算数运算和逻辑运算（+-\*/&|！^）。ALU的运算速度和效率决定了CPU的处理能力和性能。对比CPU和GPU架构图可见：CPU的ALU比较少（仅20%,剩下的80%给了内存控制器和缓冲存储器以及指令译码器等结构）。

还有总线BUS:总线就是用于连接CPU和其它设备的。

当然还有内存控制器、指令译码器、寄存器组等没有列出来。

一般CPU擅长处理逻辑控制，主要负责进行串行程序的控制和调度任务。因为其算力不高，所以CPU在计算密集数据的处理时，计算效率不高。

1. 图形处理器GPU:其提供了多核并行计算的基础结构，能够完成对大规模大数据的计算，有很高的浮点运算能力，和高访存速度。能高效完成大量的卷积运算。训练大模型的时候，很多时候都是依靠GPU来完成训练的。（GPU大体积大功耗）
2. FPGA:可编程的硬件平台，可以通过对内部的逻辑电路进行编程，通过内部的门电路实现用户定制的算法。实现特定的功能。**不需要指令系统的翻译，执行效率非常高**。FPGA优点就是比较灵活，可重构，具有可定制、高性能和低功耗的特点。适合用在资源受到限制的设备，以及便携设备中。（FPGA在AI和大数据中还有极大的潜力，但FPGA的编程复杂度比较高，开发难度比较大，要掌握FPGA需要懂数字电路设计，有硬件的背景，要会使用硬件描述语言。）
3. NPU（Neural Process Unit，神经网络处理单元）:也称为AI处理器。因此NPU是一种专门进行神经网络计算的处理器。人工神经网络（或称为神经网络一般简写为NN），神经网络是一种模仿生物神经网络结构和功能的计算模型。



NPU就是专门用于进行神经网络计算的处理器。其实就是想用电路模拟人类的神经元和突触结构。NPU通过电路层来模拟人类的神经元和突触结构通过 深度学习指令集实现处理大规模的神经元和突触。 一般NPU通过一条指令就可以完成一条神经元的处理。而CPU和GPU则需要数千条指令才能完成一条神经元的处理。

那么，NPU与CPU或GPU相比，NPU的执行效率更高。通过架构图，我们发现，NPU与前三者的架构稍有不同。ITCM和OTCM为存储单元。PE就是NPU的计算单元（处理单元）。NPU其实是有很多个处理单元组成的（中间的PE阵列）。PE阵列就是NPU中的大量的计算单元了。PE之间是互相有联系的。说明PE之间是有通信的，PE阵列通过同时并行运算而达到高效运算的目的。也就是说NPU采用数据驱动并行计算的架构。可以同时处理多批次的高维数据。这些处理单元可以同时执行多个矩阵算法操作。所以NPU非常适合深度学习中的矩阵运算。因此NPU非常适合视频、图像类的海量数据的处理。

例如：一张图片在计算机中就是像素矩阵，神经网络的核心就是进行卷积运算。而卷积运算就是将被卷积的矩阵展开后与 卷积核矩阵进行相乘以后得到卷积阵列。 所以卷积运算实质上是进行矩阵乘法。在神经网络的训练和推理过程中，主要就是涉及到矩阵的+-\*等浮点运算的。前面说过，NPU可以同时处理多批次的高维数据，可以同时执行多个矩阵乘法操作。所以NPU就特别擅长处理视频、图形类的海量多媒体数据。可以用于加速神经网络运算。

（卷积运算中主要进行的就是矩阵的乘法）

NPU是具有数据流的架构，NPU以单指令多数据模式运行，即每个PE同步运行相同的程序，充分利用了数据的复用性，减少数据在计算单元与存储单元之间的通信，可大大节省运算的延时和功耗。NPU功耗低、计算力强、效率高、成本低、体积小。可以将深度学习模型做到最小化。

可以说：NPU是专门为嵌入式人工智能而设计的计算平台。NPU非常适合用于硬件资源有限制的嵌入式设备、便携移动设备中。



了解：

APU：Accelerated Processing Unit APU是AMD的一个处理器品牌、

BPU：Brain Processing Unit。地平线机器人、

CPU：Central Processing Unit, 中央处理器；、

DPU：Deep learning Processing Unit, 深度学习处理器、

EPU：Emotion Processing Unit。Emoshape 并不是这两年才推出EPU的，号称是全球首款情绪合成（emotion synthesis）引擎，可以让机器人具有情绪。、

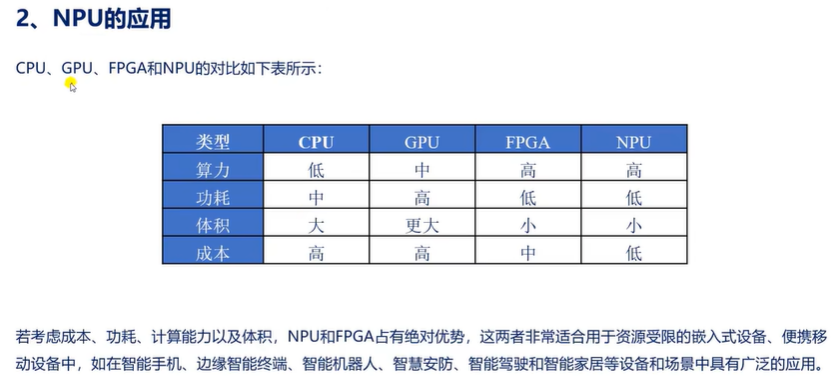
FPU：Floating Point Unit。浮点单元，不多做解释了。现在高性能的CPU、DSP、GPU内都集成了FPU做浮点运算、

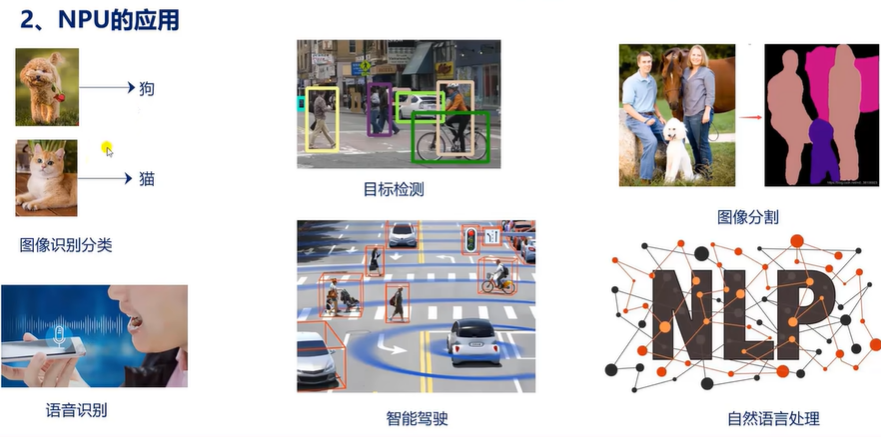
GPU：Graphics Processing Unit, 图像处理器、

NPU：Neural network Processing Unit, 神经网络处理器、

TPU：Tensor Processing Unit, 张量处理器、

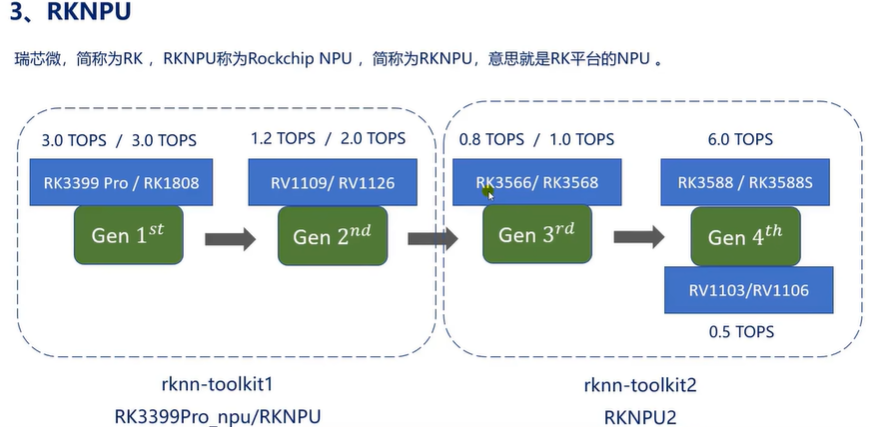
VPU：Video Processing Unit，专门设计用于处理视频相关任务的处理器单元





NPU在图像识别分类、目标检测、图像分割、智能驾驶、和自然语言处理等方面有广泛的应用。

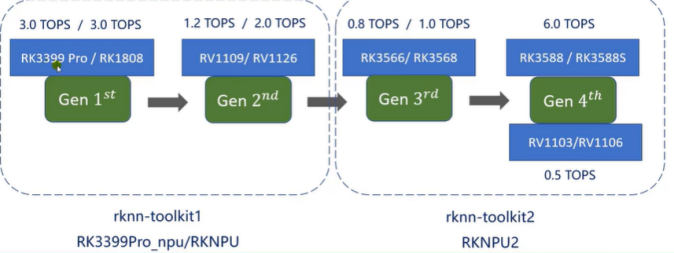
自然语言处理：如文本分类、机器翻译、情感分析等。像chartgpt,简称GPT，是由openAI开发的一种基于自然语言处理技术的聊天机器人。可以接收图像的文本输入，并提供文本输出。例如给GPT上传一张图片，就可以很快的生成网站的html代码。GPT就是一种深度神经网络模型。



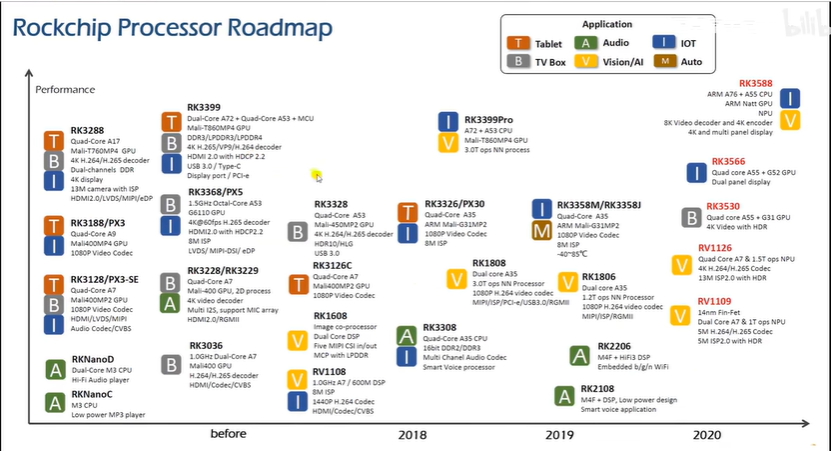
我们所说的NPU就是指瑞芯微的RKNPU，瑞芯微简称RK，为满足日渐增长的AI算力需求，RK在处理器中集成了NPU，这个NPU称为Rockchip NPU。简称RKNPU.意思就是RK平台的NPU。这一节课程里，主要就是以瑞芯微平台的芯片为例。来讲解模型的部署。

本课程对于RK的许多芯片都是通用的。例如RK3568、RK3588等平台，也可以看本课程。本课程里主要是以瑞芯微为例来讲解模型的部署。就涉及到RKNPU。

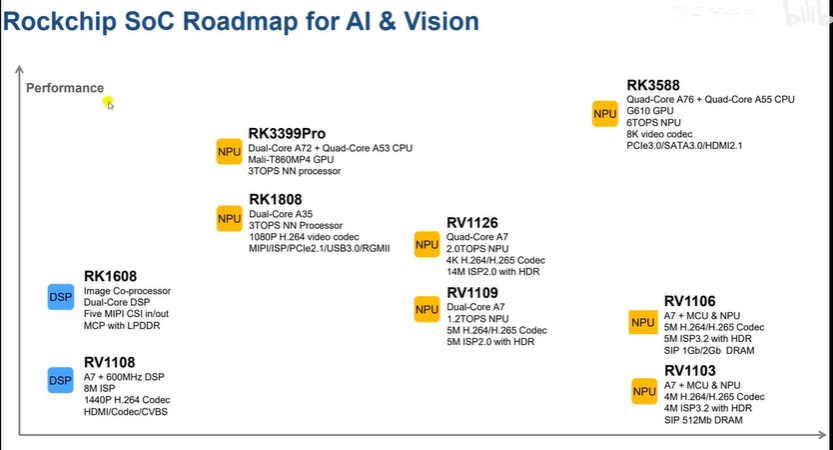
NPU的例程：

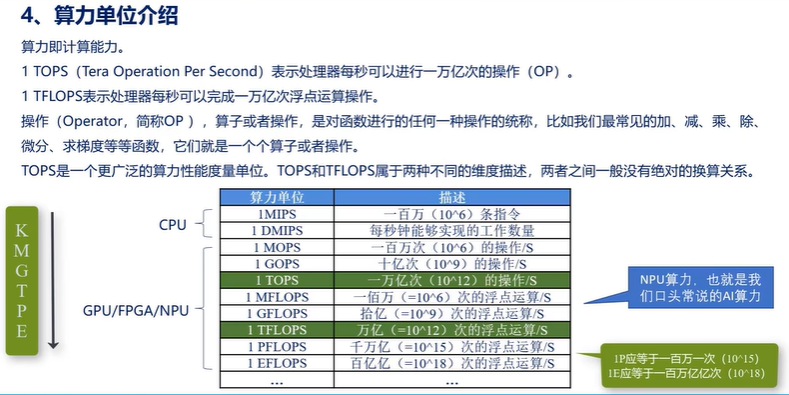


RK的产品路线图：



AI芯片的性能图：





算力就是指计算能力，也就是用于衡量计算系统或硬件系统的处理能力，在深度学习中算力一般是用来描述：GPU、FPGA、NPU等硬件加速器的性能的。是计算性能的度量。

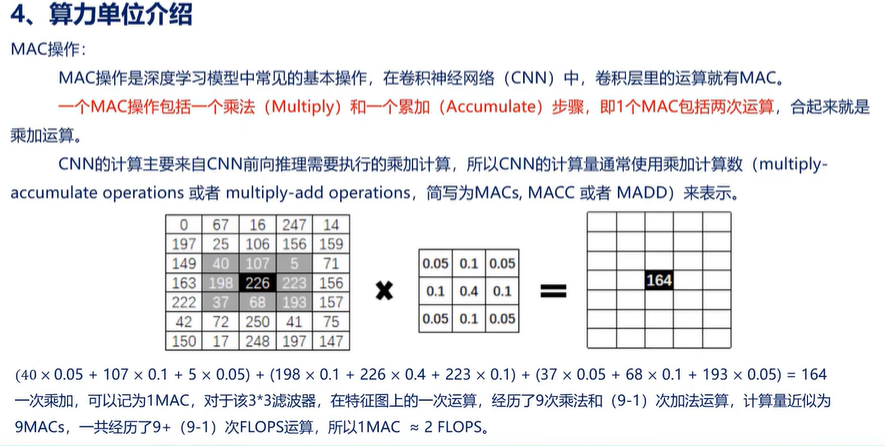
1TOPS:就是指1秒可以进行1万亿次的操作（op）。

1TFLOPS:表示处理器每秒可以完成1万亿次的浮点运算操作。

MIPS:表示1秒钟可以完成一百万条指令。

DMIPS:表示每秒钟能够实现的工作数量。

MIPS和DMIPS主要用于评价CPU的性能指标。



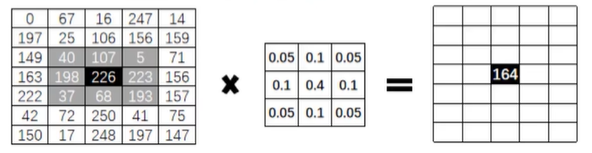
MAC操作：

是深度学习模型中常见的基本操作。如图上所示。

一个MAC包括一个乘法和一个累加步骤。即，一个MAC包括两次运算。（一个MAC就是一次乘加运算）

在CNN网络模型中（也就是卷积神经网络模型中），主要进行的就是卷积运算，那么在CNN中，乘加运算就是无处不在的，CNN的计算主要来自CNN的前向推理，需要执行乘加运算。

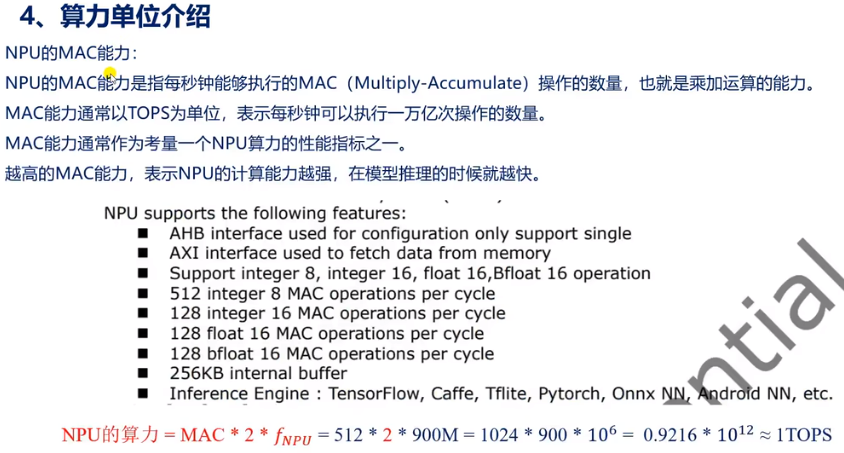
我们在NPU上运行模型（推理模型），实质就是执行前向运算的过程。例如：一个输入的图片与3\*3大小的卷积核进行卷积运算，一次运算的过程就如图所示：左侧表示图片，中间表示3\*3大小的卷积核。



一次运算的结果如下所示：

IMG_256执行了9次乘法，8次加法。近似经历了9次MAC，一共经历了9+(9-1)次FLOPS运算，所以1MAC近似等于 2FLOPS。

**再来看：NPU的MAC能力：**



MAC以TOPS为单位。

