第3章

嵌入式AI系统的硬件解决方案

CHAPTER 3

本章学习目标

- 通用类芯片——GPU
- 半定制化芯片——FPGA
- 全定制化芯片——ASIC
- 类脑芯片

视频讲解

为了使嵌入式 AI 系统满足轻量化、低成本、低功耗等的需求与挑战,各大芯片厂家均在开展 AI 专用设备及 AI 芯片的研发。本章详细地介绍四种 AI 芯片(通用类芯片、半定制化芯片、全定制芯片及类脑芯片),详细介绍了各自的工作原理、特点及优缺点,最后进行了总结与展望。通过本章的学习,可以根据嵌入式 AI 系统的不同应用特点,选择更适合的 AI 芯片。本章内容的框图如图 3-1 所示。

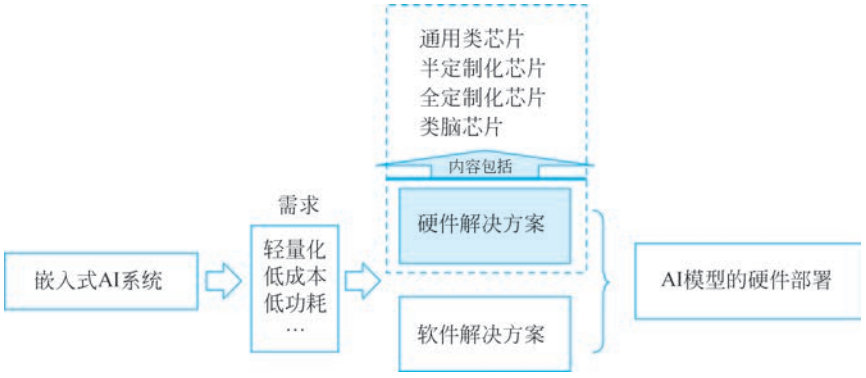

图 3-1 本章内容框图

常用的 DNN 算法一般包括 CNN 卷积网络(用于图像识别领域)和 RNN 网络(用于语音识别、自然语言处理领域),它们的本质都是矩阵或向量的乘法和加法运算。例如,一个针对输入为图像的目标检测算法需要一万亿次的加法与乘法运算。如图 3-2 所示,AI 芯片从技术架构的角度可以分为以下四个类型,其中 GPU、FPGA、ASIC 都属于冯·诺依曼架构[①],其类别、特点及代表公司总结如表 3-1 所示。

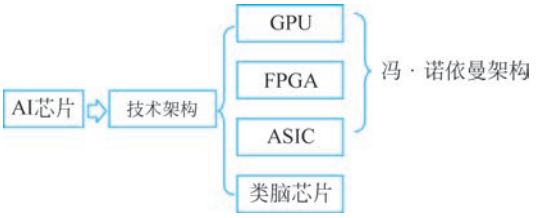

图 3-2 AI 芯片技术架构

表 3-1 AI 芯片的四个类别

类别	GPU	FPGA	ASIC	类脑芯片
特点	性能高 功耗高 通用性好	可编程、灵活 功耗与通用性介于 GPU 与 ASIC 之间	定制化设计 性能稳定 功耗控制性能好	功耗低 响应速度快 尚不成熟
代表公司	NVIDIA AMD	Xilinx Intel	谷歌(TPU) 寒武纪 地平线	IBM

① 冯·诺依曼结构也称普林斯顿结构,是一种将程序指令存储器和数据存储器合并在一起的存储器结构。数学家冯·诺依曼提出了计算机制造的三个基本原则,即采用二进制逻辑、程序存储执行以及计算机由五个部分组成(运算器、控制器、存储器、输入设备、输出设备),这套理论被称为冯·诺依曼体系结构。

3.1　通用类芯片——GPU

CPU 与 GPU 的结构对比示意图如图 3-3 所示。传统 CPU 的计算指令是遵循串行执行方式的，在执行 DNN 推理时，难以发挥出 CPU 的全部潜力，所以它并不适用于 DNN 推理。而 GPU 具有并行结构，在处理图形数据和复杂算法方面拥有比 CPU 更高的效率。对比 CPU 和 GPU 在结构上的差异，CPU 大部分面积为控制器和寄存器，而 GPU 拥有更多的用于数据处理的逻辑运算单元(Arithmetic Logic Unit，ALU)，这样的结构更适合对密集型数据进行并行处理。由此可以看出，具有大量重复运算的 DNN 推理在 GPU 上的运行速度与单核 CPU 比较可以提升几十倍乃至上千倍。NVIDIA、AMD 等公司不断地推进 GPU 大规模并行架构的研发，因此 GPU 已成为加速可并行应用程序的关键。

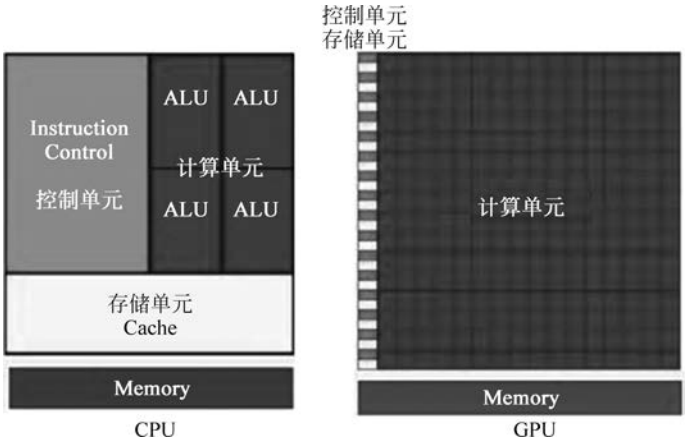

图 3-3　CPU 与 GPU 的结构对比示意图

从图 3-3 中可以很明显地看出，GPU 的构成相对简单，有数量众多的计算单元和超长的流水线，特别适合处理大量的类型统一的数据。但 GPU 无法单独工作，必须由 CPU 进行控制调用才能工作。CPU 可单独工作，处理复杂的逻辑运算和不同的数据类型，但当需要处理大量的类型统一的数据时，则需调用 GPU 进行并行计算。CPU 根据功能划分，将需要大量并行计算的任务分配给 GPU。GPU 从 CPU 获得指令后，把大规模、无结构化的数据分解成许多独立部分，分配给各个流处理集群。每个流处理集群再次把数据分解，分配给调度器，调度器将任务放入自身所控制的计算核心区中完成最终的数据处理任务。但是 GPU 也有一定的局限性，深度学习算法分为训练和推理两部分，GPU 平台在算法的训练过程中表现出高效的特点，但在推理过程中对单项输入进行处理时，并行计算的优势就无法完全发挥出来。

GPU 性能较强但功耗较高。以 NVIDIA 开发的 GPU 为例，Xavier 最高算力为 30 TOPS[①]，功耗为 30W，NVIDIA 最新发布的 GPUA 100 的性能大幅增强，支持全新的 TF32

① 在功耗方面，用 TOPS/W 评价处理器运算能力，TOP 是 Tera Operations Per Second 的缩写，TOPS/W 用于度量在 1W 功耗的情况下，处理器能进行多少万亿(10^{12})次操作；GOPS/W 度量处理器在 1W 功耗的情况下进行多少十亿次(10^{9})操作；1MOPS 代表处理器每秒钟可进行一百万次(10^{6})操作。TOPS 同 GOPS 与 MOPS 可以换算，都代表每秒钟能处理的次数，是不同的单位。

运算,浮点性能 156 TFLOPS。TFLOPS 是 Floating-point Operations per Second 的缩写,指每秒所执行的浮点运算次数。同时 INT8 浮点性能为 624 TOPS,FP16 性能为 312 TFLOPS,但功耗也达到了 400W。

全球知名的 GPU 生产厂商包括 NVIDIA、AMD、ARM、Qualcomm 的 Adreno 等。

3.2 半定制化芯片——FPGA

现场可编程门阵列(Field-Programmable Gate Array,FPGA)是在可编程阵列逻辑(Programming Array Logic,PAL)、通用阵列逻辑器件(Generic Array Logic,GAL)、复杂可编程逻辑器件(Complex Programmable Logic Device,CPLD)等可编程器件的基础上进一步发展的产物。FPGA 是一种可以重构电路的芯片,是一种硬件可重构的体系结构,作为专用集成电路(Application Specific Integrated Circuit,ASIC)领域中的一种半定制电路而出现的,既解决了定制电路的不足,又克服了原有可编程器件门电路数量有限的缺点。通过编程,用户可以随时改变它的应用场景,它可以模拟 CPU、GPU 等硬件的各种并行运算。

1985 年,Xilinx 公司推出了全球第一款 FPGA 产品 XC2064,采用 2μm 工艺,包含 64 个逻辑模块和 85 000 个晶体管,门电路数量不超过 1000 个。2016 年,Xilinx 发布的 VIRTEX UltraScale 为 16nm 制程,系统逻辑单元最高达 378 万个。FPGA 制程迭代在提高算力的同时降低了功耗,减小了芯片面积,推动了 FPGA 的性能提升。Xilinx 和 Intel 相继发布 ACAP 和 Agilex 平台型产品,根据 Xilinx 披露的数据,新的平台型产品速度超过当前最高速 FPGA 的 20 倍,比目前最快的 CPU 快 100 倍,该平台面向数据中心、有线网络、5G 无线和汽车驾驶辅助应用。

FPGA 的内部基本结构如图 3-4 所示。用户可以通过写入 FPGA 配置文件来定义这些门电路以及存储器之间的连线。这种写入不是一次性的,例如,用户可以把 FPGA 配置成一个微控制器,在使用完毕后,再通过编辑配置文件把同一个 FPGA 配置成一个音频编解码器。因此,它既解决了定制电路灵活性不足的问题,又克服了原有可编程器件门电路数有限的缺点。

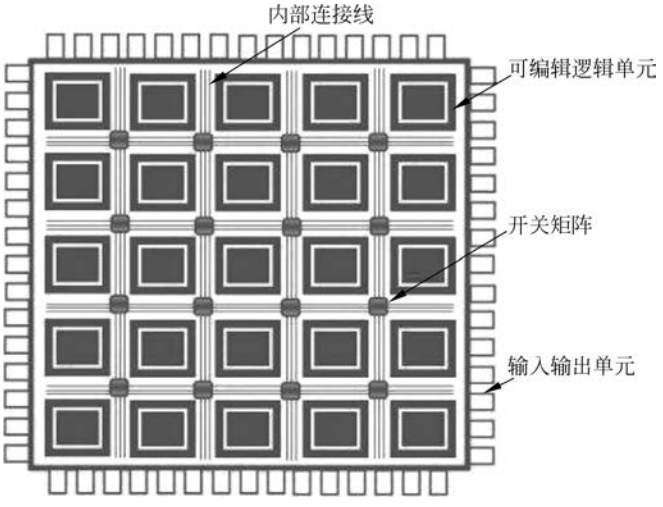

图 3-4 FPGA 的内部基本结构

FPGA 的主要优点介绍如下。

FPGA 具有更高的性能。FPGA 同时拥有流水线并行和数据并行，而 GPU 几乎只有数据并行。FPGA 可以同时进行数据并行和任务并行的计算，在处理特定应用时有更加明显的效率提升。对于某个特定运算，通用 CPU 可能需要多个时钟周期，而 FPGA 则通过编程重组电路，直接生成专用电路，仅消耗少量甚至一次时钟周期就可完成运算。所以说，在功耗和延迟方面，FPGA 在体系结构上是具有天生优势的。

FPGA 具有更低的功耗和延迟。在冯·诺依曼结构中，由于执行单元（如 CPU 核）可能执行任意指令，就需要有指令存储器、译码器、各种指令的运算器、分支跳转处理逻辑。由于指令流的控制逻辑复杂，不可能有太多条独立的指令流，因此 GPU 使用单指令流多数据流（Single Instruction Multiple Data，SIMD）来让多个执行单元以相同的步调处理不同的数据，同时 CPU 也支持 SIMD 指令。而 FPGA 每个逻辑单元的功能在重编程（即烧写）时就已经确定，不再需要指令。因此，体系结构的区别使 FPGA 比 GPU 的延迟低很多。

FPGA 具有灵活性。很多使用通用类芯片 GPU 或全定制化 ASIC 芯片难以实现的底层硬件控制操作技术，利用 FPGA 都可以很方便地实现，这个特性为算法的功能实现和优化留出了更大空间。FPGA 是作为 ASIC 领域中的一种半定制电路而出现的，它既解决了定制电路的不足，又克服了原有可编程器件门电路数有限的缺点。与 ASIC 芯片相比，FPGA 的一项重要特点是可编程特性，即用户可通过程序指定 FPGA 实现某一特定数字电路，而且 FPGA 芯片是小批量系统提高系统集成度、可靠性的最佳选择之一。同时 FPGA 的一次性成本（光刻掩模制作成本）远低于 ASIC，在芯片需求还未成规模、深度学习算法暂未稳定，需要不断迭代改进的情况下，利用 FPGA 芯片具备可编程的特性来实现半定制的 AI 芯片是最佳选择之一。

综上所述，FPGA 具有高性能、低能耗、灵活性的特点，归纳如表 3-2 所示。

表 3-2　FPGA 的特点

特　点	描　述
高性能	除了 GPU，FPGA 也擅长并行计算，基于 FPGA 开发的处理器可以实现更高的并行计算。而且 FPGA 带有丰富的片上存储资源，可以大大减少访问片外存储的延迟，提高计算性能，访问 DRAM 存储大约是访问寄存器存储延迟的几百倍以上
低能耗	相比于 CPU 和 GPU，FPGA 的能耗优势主要有两个原因：①相比于 CPU、GPU，FPGA 架构有一定的优化，CPU、GPU 需要频繁地访问 DRAM，而这个能量消耗较大，FPGA 可以减少这方面的能耗。②FPGA 的主频低，CPU 和 GPU 的主频一般在 1～3GHz 之间，而 FPGA 的主频一般在 500MHz 以下。因此，FPGA 的能耗要低于 CPU、GPU
灵活性	FPGA 可硬件编程，并且可以进行静态重复编程和动态系统重配置。用户可像编程修改软件一样修改系统的硬件功能，大大增强了系统设计的灵活性和通用性。使得 FPGA 可以灵活地部署在需要修改硬件设置场景中

全球知名的 FPGA 生产厂商包括 Altera（Intel 收购）、Xilinx、Actel、Lattice、Atmel 等。国内的厂商包括深圳紫光同创、上海安路科技、广东高云半导体、上海复旦微电子、京微齐力等。

3.3 全定制化芯片——ASIC

ASIC,即专用集成电路,是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。用 CPLD 和 FPGA 来进行 ASIC 设计是目前最为流行的方式之一。

CPU、GPU、FPGA 都属于通用类的芯片,GPU 与 CPU 相比并行处理的能力较好。ASIC 是依照产品需求不同而定制化的特殊规格的集成电路,根据使用者的要求和特定电子系统的需要而设计并制造。它作为集成电路技术与特定用户的整机或系统技术紧密结合的产物,与通用集成电路相比具有体积更小、重量更轻、功耗更低、可靠性提高、性能提高、保密性增强、成本降低等优点。ASIC 芯片技术发展迅速,目前 ASIC 芯片间的转发性能通常可达到 1Gb/s 甚至更高,为交换矩阵提供了极好的物质基础。

ASIC 芯片的缺点是开发周期较长。基于 ASIC 人工智能芯片更像是电路设计,需要反复优化,还要经历较长的流片周期,故开发周期较长。相较于 FPGA,ASIC 人工智能芯片需要经历较长的开发周期,并且需要价格昂贵的流片投入,但量产后,ASIC 人工智能芯片的成本和价格会低于 FPGA 芯片。ASIC 芯片性能功耗比较高。从性能功耗比来看,ASIC 作为全定制化芯片,其性能要比基于半定制化芯片 FPGA 开发出的各种半定制 AI 芯片更具有优势。而且 ASIC 也不是完全不具备可配置能力,只是没有 FPGA 那么灵活,只要在设计的时候把电路做成某些可调参数即可。

张量处理器(Tensor Processing Unit,TPU)就是谷歌专门为了加速 DNN 的运算能力而研发的芯片,属于典型的 ASIC 芯片。TPU 与同期的 CPU 和 GPU 相比,性能可以提升 15～30 倍,效率可提升 30～80 倍。谷歌的 TPU、寒武纪的 MLU、地平线的 BPU 都属于 ASIC 芯片。TPU 与同期的 CPU 和 GPU 相比,缩小了控制部分,减少了芯片的面积,从而降低了功耗。

3.4 类脑芯片

生物体的神经网络由若干人工神经元结点互联而成,神经元之间通过突触两两连接,突触记录了神经元之间的联系。由于深度学习的基本操作是神经元和突触的处理,传统的处理器是为了进行通用计算而发展起来的,其基本操作为算术操作(加减乘除)和逻辑操作(与或非),所以往往需要数百甚至上千条指令才能完成一个神经元的处理,导致深度学习的处理效率不高。而且神经网络中存储和处理是一体化的,都是通过突触权重来体现。神经网络处理器(Neural network Processing Unit,NPU)的原理是用电路模拟人类的神经元和突触结构。而传统的冯·诺依曼结构中存储和处理是分离的,分别由存储器和运算器来实现,二者之间存在巨大的差异。所以使用基于冯·诺依曼结构的经典计算机(如 X86 处理器和 GPU)来运行 DNN 推理时,就不可避免地受到存储和处理分离式结构的制约而影响效率。

NPU 芯片的典型代表有国内的寒武纪芯片和 IBM 的 TrueNorth。以中国的寒武纪为

例，DianNaoYu 指令直接应对大规模神经元和突触的处理，一条指令即可完成一组神经元的处理，并专门为神经元和突触数据在芯片上的传输提供了一系列的支持。IBM 研究人员将存储单元作为突触、计算单元作为神经元、传输单元作为轴突搭建了神经芯片的原型。由于神经突触要求权重可变且要有记忆功能，IBM 采用与 CMOS 工艺兼容的相变非挥发存储器（PCM）的技术实验性地实现了新型突触，加快了商业化进程。类脑芯片是人工智能最终的发展模式，但是距离产业化还很遥远。

3.5　对四大类型 AI 芯片的总结与展望

3.5.1　对 AI 芯片的总结

综上对四大类型 AI 芯片的介绍，再次归纳可以得到以下结论。

（1）CPU 有强大的调度、管理、协调能力；应用范围广；开发方便且灵活；但在大量数据的处理上没有 GPU 专业，相对运算量低，功耗较高。

（2）GPU 是单指令、多数据处理结构，有数量众多的计算单元和超长的流水线，GPU 善于处理图像领域的运算加速。但 GPU 无法单独工作，必须由 CPU 进行控制调用才能工作。CPU 可单独工作，处理复杂的逻辑运算和不同的数据类型，但当需要处理大量的类型统一的数据时，则需调用 GPU 进行并行计算。

（3）FPGA 和 GPU 相反，FPGA 适用于多指令、单数据流的分析，因此常用于推理阶段。将 FPGA 和 GPU 对比发现，一是 FPGA 缺少内存和控制所带来的存储和读取部分，速度更快；二是因为 FPGA 缺少读取的作用，所以功耗低，FPGA 的劣势是运算量并不大。

（4）ASIC 芯片是全定制化芯片，是为实现特定要求而设计制定的芯片。除了不能扩展以外，在功耗、可靠性、体积方面都有优势，尤其在高性能、低功耗的移动端和嵌入式端，但其灵活性差，开发周期较长。

3.5.2　对 AI 芯片的展望

物联网及人工智能时代对于智能硬件的需求对嵌入式系统的软硬件都带来了挑战，主要来源于算力、存储及功耗三个方面。

如图 3-5 所示，实现 10FPS（DNN 模型每秒的推理频率为 10）时图像分类所需要的运算量，通常需要 $10\times10^9\sim150\times10^9$ 次的乘加运算。图 3-6 展示了几种 DNN 模型参数存储量，可以看出参数存储量在 $5\times10^6\sim140\times10^6$ 之间，使用单精度浮点数进行存储时对应的存储量为 20～560MB。传统的低成本嵌入式系统的 RAM 存储空间往往不超过 16MB。一般用于图像及语音处理的实时 DNN 网络在处理器上的运算能力会超过 100GOPS 甚至 1TOPS，同时每次 DNN 推理都需要获取数百万个网络参数，并完成大量的运算操作，这些庞大的参数提取及运算操作所消耗的能量是在能量匮乏的嵌入式设备中进行 DNN 推理时遇到的主要困难。图 3-7 显示了在 30FPS（DNN 模型每秒的推理频率为 30）运行 DNN 网络模型需要全系统的等效能源效率。可以看出，随着更小规模的 DNN 网络结构不断地被

提出,其需要的能源效率变得越来越小,但对于现有的嵌入式计算资源,实时执行 DNN 推理仍然存在着瓶颈,仍需要以优化 GPU、NPU 和 ASIC 的形式进行硬件方面的创新。

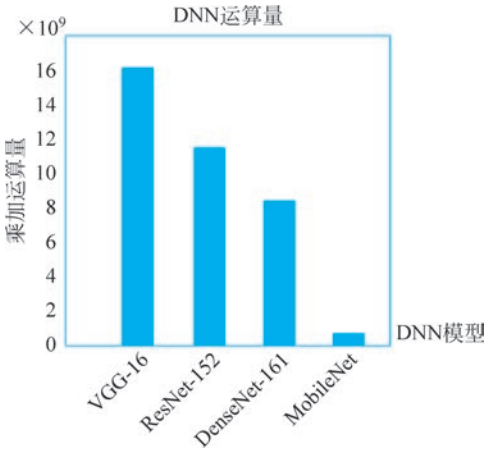

图 3-5　不同 DNN 网络模型实时运行时运算量对比

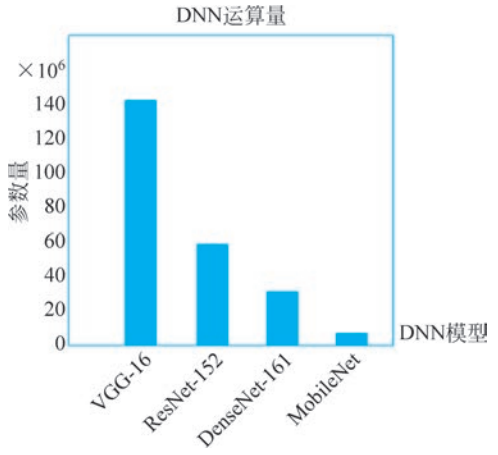

图 3-6　不同 DNN 网络模型参数量对比

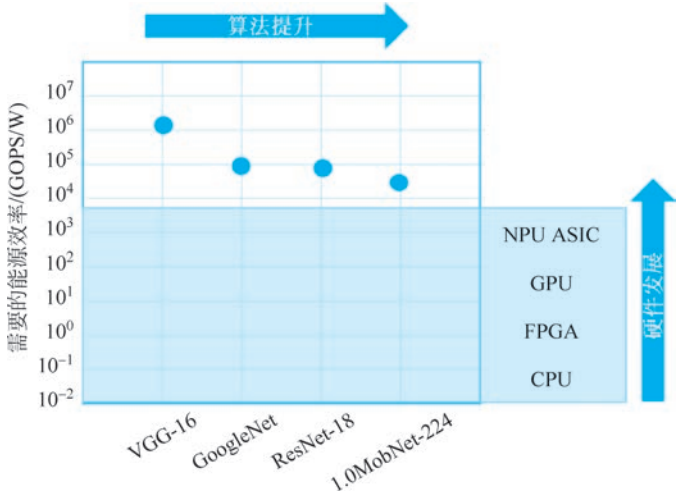

图 3-7　不同 DNN 网络模型实时运行时所需要的等效能源

综上所述,无论在运算量、存储及功耗方面,在嵌入式 AI 芯片上实现实时的 DNN 推理还存在着瓶颈问题,需要通过软件优化及硬件优化之间紧密地相互作用来解决。本书在第 4～6 章介绍算法层面的优化方法。

3.6　本章小结

本章介绍了 AI 芯片的主要分类,即通用类芯片、半定制化芯片、全定制化芯片及类脑芯片,并描述了每类芯片的设计原理、特点、优势及缺点,且进行了对比与分析,为选择 AI 系统的硬件解决方案提供了理论支持。

3.7　习题

1. 常用的 AI 芯片可以分成哪些类？
2. 简述 CPU 和 GPU 的区别与各自的特点。
3. 简述 FPGA 的工作原理，并描述 FPGA 与 CPU、GPU 的区别。
4. 简述 ASIC 的概念，并描述它与 FPGA 的区别。