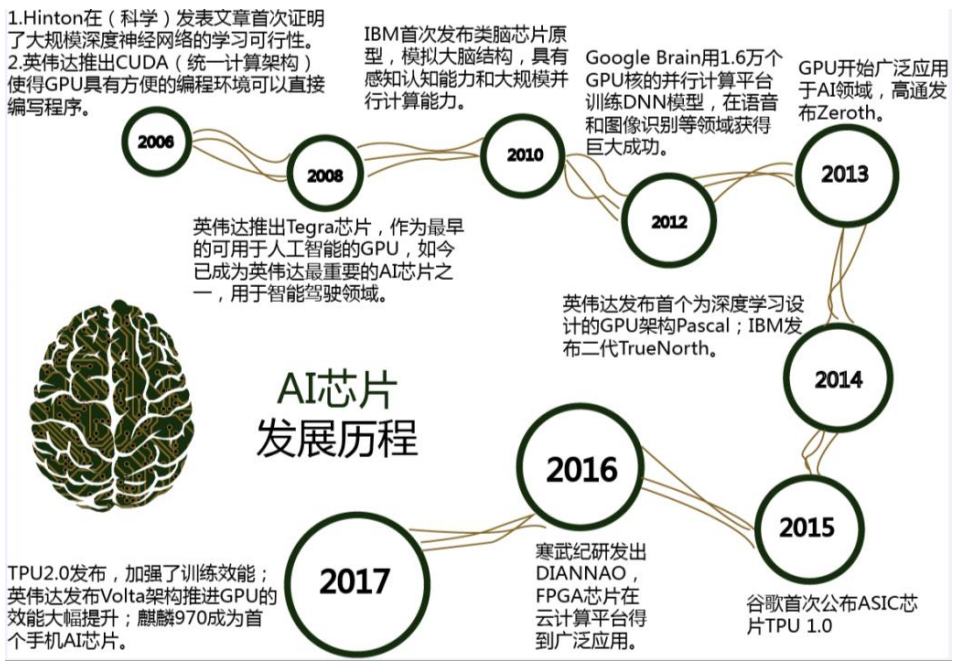
1.1 深度神经网络的主流硬件平台

深度神经网络的模型层数多、参数量大且计算复杂，对硬件的计算能力、内存带宽及数据存储等有较高的要求。因此，计算能力强、可并行加速、数据吞吐高的高性能硬件平台对于模型训练和产业应用来说显得尤为重要。

随着硬件技术和深度神经网络的发展，目前形成了以“CPU+GPU”的异构模式服务器为主的深度神经网络的研究平台，如英伟达的 DGX-2，其具有16 块 Tesla V100 GPU，可以提供最高达 2 PFLOPs 的计算能力。面对复杂的实际应用需求和不断加深的神经网络结构，多样化的深度神经网络硬件平台也不断发展起来，形成了以通用性芯片（CPU、GPU）、半制定化芯片（FPGA）、全制定化芯片（ASIC）、集成电路芯片（SoC）和类脑芯片等为主的硬件平台市场。计算性能、灵活性、易用性、成本和功耗等成为评价深度神经网络硬件平台的因素和标准。

作为人工智能核心的底层硬件 AI 芯片，经历了多次的起伏和波折，总体看来， AI 芯片的发展前后经历了四次大的变化，其发展历程如图所示。



1.2.1 GPU（Graphics Processing Unit，图形处理单元）：

GPU作为最早从事并行加速计算的处理器，相比 CPU 速度快，同时比其他加速器芯片编程灵活简单。传统的 CPU 之所以不适合人工智能算法的执行，主要原因在于在传统的冯·诺依曼结构中，CPU 每执行一条指令都需要从存储器中读取数据，根据指令对数据进行相应的操作。从这个特点可以看出，CPU 的主要职责并不只是数据运算，还需要执行存储读取、指令分析、分支跳转等命令。深度学习算法通常需要进行海量的数据处理，用 CPU 执行算法时，CPU 将花费大量的时间在数据/指令的读取分析上，而 CPU的频率、内存的带宽等条件又不可能无限制提高，因此限制了处理器的性能。而 GPU 的控制相对简单，大部分的晶体管可以组成各类专用电路、多条流水线，使得 GPU 的计算速度远高于 CPU；同时 GPU 拥有了更加强大的浮点运算能力，可以缓解深度学习算法的训练难题，释放人工智能的潜能。CPU大部分面积为控制器和寄存器，而 GPU拥有更多的 ALU(ARITHMETIC LOGIC UNIT，逻辑运算单元)用于数据处理，这样的结构适合对密集型数据进行并行处理。程序在 GPU系统上的运行速度相较于单核 CPU往往提升几十倍乃至上千倍。随着英伟达、AMD 等公司不断推进其对 GPU 大规模并行架构的支持，面向通用计算的 GPU(即 GPGPU GENERAL PURPOSE GPU，通用计算图形处理器)已成为加速可并行应用程序的重要手段。

GPU的发展历程可分为 3 个阶段，发展历程示意图如图 5 所示：

第一代 GPU(1999 年以前 ) ， 部分功能从 CPU 分离，实现硬件加速，以GE(GEOMETRY ENGINE)为代表，只能起到 3D 图像处理的加速作用，不具有软件编程特性。

第二代 GPU(1999-2005 年)，实现进一步的硬件加速和有限的编程性。1999 年，英伟达发布了“专为执行复杂的数学和几何计算的”GeForce256 图像处理芯片，将更多的晶体管用作执行单元，而不是像 CPU 那样用作复杂的控制单元和缓存，将 T&L(TRANSFORMAND LIGHTING)等功能从 CPU分离出来，实现了快速变换，这成为 GPU真正出现的标志。之后几年，GPU技术快速发展，运算速度迅速超过 CPU。2001 年英伟达和 ATI 分别推出的

GEFORCE3 和 RADEON 8500，图形硬件的流水线被定义为流处理器，出现了顶点级可编

程性，同时像素级也具有有限的编程性，但 GPU的整体编程性仍然比较有限。

第三代 GPU(2006年以后)，GPU实现方便的编程环境创建，可以直接编写程序。2006

年英伟达与ATI分别推出了CUDA(Compute United Device Architecture，计算统一设备架构)

编程环境和 CTM(CLOSE TO THE METAL)编程环境，使得 GPU打破图形语言的局限成为真正的并行数据处理超级加速器。

2008 年，苹果公司提出一个通用的并行计算编程平台 OPENCL（OPEN COMPUTING

LANGUAGE，开放运算语言），与 CUDA 绑定在英伟达的显卡上不同，OPENCL 和具体的 计算设备无关。

GPU起初专门用于处理图形任务，主要由控制器、寄存器和逻辑单元构成。GPU 包含几千个流处理器，可将运算并行化执行，大幅缩短模型的运算时间，由于其强大的计算能力，目前主要被用于处理大规模的计算任务。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。在深度神经网络中，大多数计算都是矩阵的线性运算，它涉及大量数据运算，但控制逻辑简单。对于这些庞大的计算任务，GPU的并行处理器表现出极大的优势。自从 AlexNet在 2012 年的 ImageNet 比赛中取得优异成绩以来，GPU 被广泛应用于深层神经网络的训练和推理。大量依赖 GPU 运算的深度神经网络软件框架（如：TensorFlow 、 PyTorch 、 Caffe 、 Theano 和Paddle-Paddle 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。

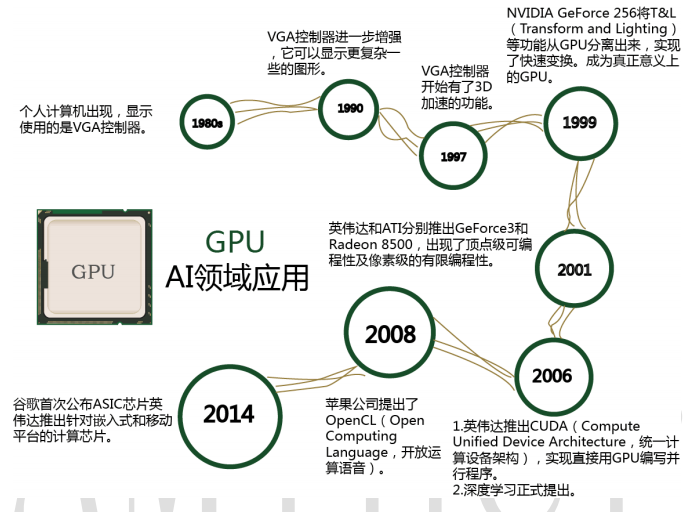


图 2 GPU 芯片的发展阶段

目前，GPU 已经发展到较为成熟的阶段。谷歌、FACEBOOK、微软、TWITTER 和百

度等公司都在使用 GPU 分析图片、视频和音频文件，以改进搜索和图像标签等应用功能。

此外，很多汽车生产商也在使用 GPU 芯片发展无人驾驶。不仅如此，GPU 也被应用于

VR/AR 相关的产业。

但是 GPU也有一定的局限性。深度学习算法分为训练和推断两部分，GPU平台在算法

训练上非常高效。但在推断中对于单项输入进行处理的时候，并行计算的优势不能完全发

挥出来。

1.2.2 FPGA

FPGA 是在 PAL、GAL、CPLD 等可编程器件基础上进一步发展的产物。用户可以通过

烧入 FPGA 配置文件来定义这些门电路以及存储器之间的连线。这种烧入不是一次性的，

比如用户可以把 FPGA 配置成一个微控制器 MCU，使用完毕后可以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的不足，又克服了原有可编程器件门电路数有限的缺点。 FPGA 可同时进行数据并行和任务并行计算，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 可以通过编程重组电路，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。 此外，由于 FPGA的灵活性，很多使用通用处理器或 ASIC 难以实现的底层硬件控制操作技术，利用 FPGA 可以很方便的实现。这个特性为算法的功能实现和优化留出了更大空间。同时 FPGA 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 芯片具备可重构的特性来实现半定制的人工智能芯片是最佳选择之一。功耗方面，从体系结构而言，FPGA 也具有天生的优势。传统的冯氏结构中，执行单元（如 CPU 核）执行任意指令，都需要有指令存储器、译码器、各种指令的运算器及分支跳转处理逻辑参与运行，而 FPGA 每个逻辑单元的功能在重编程（即烧入）时就已经确定，不需要指令，无需共享内存，从而可以极大的降低单位执行的功耗，提高整体的能耗比。

由于 FPGA 具备灵活快速的特点，因此在众多领域都有替代 ASIC 的趋势。FPGA 在人

工智能领域的应用如图 6 所示

FPGA（Field Programmable GateArray)全称“现场可编程门阵列”， 其基本原理是在 FPGA芯片内集成大量的基本门电路以及存储器，用户可以通过更新FPGA 配置文件来定义这些门电路以及存储器之间的连线。 与 GPU 不同，FPGA 同时拥有硬件流水线并行和数据并行处理能力，适用于以硬件流 水线方式处理一条数据，且整数运算性能更高，因此常用于深度学习算法中的推断阶段。不过 FPGA 通过硬件的配置实现软件算法，因此在实现复杂算法方面有一定的难度。将 FPGA 和 CPU 对比可以发现两个特点，一是 FPGA 没有内存和控制所带来的存储和读取部分，速度更快，二是 FPGA 没有读取指令操作，所以功耗更低。劣势是价格比较高、编程复杂、整体运算能力不是很高。目前国内的 AI 芯片公司如深鉴科技就提供基于 FPGA 的解决方案。

FPGA允许无限次的编程，并利用小型查找表来实现组合逻辑。FPGA 可以定制化硬件流水线，可以同时处理多个应用或在不同时刻处理不同应用，具有可编程、高性能、低能耗、高稳定、可并行和安全性的特点，在通信、航空航天、汽车电子、工业控制、测试测量等领域取得了很大应用市场。人工智能产品中往往是针对一些特定应用场景而定制的，定制化芯片的适用性明显比通用芯片的适用性高。FPGA 成本低并且具有较强的可重构性，可进行无限编程。因此，在芯片需求量不大或者算法不稳定的时候，往往使用 FPGA 去实现半定制的人工智能芯片，这样可以大大降低从算法到芯片电路的成本。随着人工智能技术的发展，FPGA 在加速数据处理、神经网络推理、并行计算等方面表现突出，并在人脸识别、自然语言处理、网络安全等领域取得了很好的应用。

1.2.3 ASIC

目前以深度学习为代表的人工智能计算需求，主要采用 GPU、FPGA 等已有的适合并

行计算的通用芯片来实现加速。在产业应用没有大规模兴起之时，使用这类已有的通用芯

片可以避免专门研发定制芯片（ASIC）的高投入和高风险。但是，由于这类通用芯片设计

初衷并非专门针对深度学习，因而天然存在性能、功耗等方面的局限性。随着人工智能应

用规模的扩大，这类问题日益突显。

GPU 作为图像处理器，设计初衷是为了应对图像处理中的大规模并行计算。因此，在

应用于深度学习算法时，有三个方面的局限性：第一，应用过程中无法充分发挥并行计算

优势。深度学习包含训练和推断两个计算环节，GPU 在深度学习算法训练上非常高效，但

对于单一输入进行推断的场合，并行度的优势不能完全发挥。第二，无法灵活配置硬件结

构。GPU 采用 SIMT 计算模式，硬件结构相对固定。目前深度学习算法还未完全稳定，若

深度学习算法发生大的变化，GPU 无法像 FPGA 一样可以灵活的配制硬件结构。第三，运

行深度学习算法能效低于 FPGA。

尽管 FPGA 倍受看好，甚至新一代百度大脑也是基于 FPGA 平台研发，但其毕竟不是

专门为了适用深度学习算法而研发，实际应用中也存在诸多局限：第一，基本单元的计算

能力有限。为了实现可重构特性，FPGA 内部有大量极细粒度的基本单元，但是每个单元

的计算能力(主要依靠 LUT 查找表)都远远低于 CPU和 GPU中的 ALU模块；第二、计算资

源占比相对较低。为实现可重构特性，FPGA 内部大量资源被用于可配置的片上路由与连

线，；第三，速度和功耗相对专用定制芯片(ASIC)仍然存在不小差距；第四，FPGA 价格较

为昂贵，在规模放量的情况下单块 FPGA 的成本要远高于专用定制芯片。

因此，随着人工智能算法和应用技术的日益发展，以及人工智能专用芯片ASIC产业环

境的逐渐成熟，全定制化人工智能ASIC也逐步体现出自身的优势，从事此类芯片研发与应

用的国内外比较有代表性的公司如表 1 所示，后续产业篇会做相应的详细介绍。

ASIC 芯片非常适合人工智能的应用场景。首先，ASIC 的性能提升非常明显。例如英

伟达首款专门为深度学习从零开始设计的芯片 Tesla P100 数据处理速度是其 2014 年推出

GPU 系列的 12 倍。谷歌为机器学习定制的芯片 TPU 将硬件性能提升至相当于当前芯片按

摩尔定律发展 7 年后的水平。正如 CPU 改变了当年庞大的计算机一样，人工智能 ASIC 芯

片也将大幅改变如今 AI 硬件设备的面貌。如大名鼎鼎的 AlphaGo 使用了约 170 个图形处理

器（GPU）和 1200 个中央处理器（CPU），这些设备需要占用一个机房，还要配备大功率

的空调，以及多名专家进行系统维护。而如果全部使用专用芯片，极大可能只需要一个普

通收纳盒大小的空间，，且功耗也会大幅降低。

第二，下游需求促进人工智能芯片专用化。从服务器，计算机到无人驾驶汽车、无人

机再到智能家居的各类家电，至少数十倍于智能手机体量的设备需要引入感知交互能力和

人工智能计算能力。而出于对实时性的要求以及训练数据隐私等考虑，这些应用不可能完

全依赖云端，必须要有本地的软硬件基础平台支撑，这将带来海量的人工智能芯片需求。

目前人工智能专用芯片的发展方向包括：主要基于 FPGA 的半定制、针对深度学习算

法的全定制和类脑计算芯片 3 个方向。

在芯片需求还未形成规模、深度学习算法暂未稳定，AI 芯片本身需要不断迭代改进的

情况下，利用具备可重构特性的 FPGA 芯片来实现半定制的人工智能芯片是最佳选择之一。

这类芯片中的杰出代表是国内初创公司深鉴科技，该公司设计了“深度学习处理单元”

（Deep Processing Unit，DPU）的芯片，希望以 ASIC 级别的功耗达到优于 GPU 的性能，

其第一批产品就是基于 FPGA 平台开发研制出来的。这种半定制芯片虽然依托于 FPGA 平

台，但是抽象出了指令集与编译器，可以快速开发、快速迭代，与专用的 FPGA 加速器产

品相比，也具有非常明显的优势。

深度学习算法稳定后，AI 芯片可采用 ASIC 设计方法进行全定制，使性能、功耗和面

积等指标面向深度学习算法做到最优。

ASICc（Application-Specific Integrated Circuit）专用集成电路，是专用定制芯片，即为实现特定要求而定制的芯片。定制的特性有助于提高 ASIC 的性能功耗比，缺点是电路设计需要定制，相对开发周期长，功能难以扩展。但在功耗、可靠性、集成度等方面都有优势，尤其在要求高性能、低功耗的移动应用端体现明显。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、 亚马逊的 Inderentia、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。谷歌的 TPU比 CPU和 GPU的方案快 30 至 80 倍，与 CPU 和 GPU 相比，TPU 把控制电路进行了简化，因此减少了芯片的面积，降低了功耗。与通用集成电路相比，由于 ASIC 是专为特定目的而设计。它具有体积更小、功耗更低、性能提高、保密性增强等优点，ASIC 可以实现 5-10 倍的计算加速，具有很高的商业价值，特别适合移动终端的消费电子领域的产业应用，且量产后 ASIC 的成本会大大降低。

1.2.4 SoC

SoC (System on Chip) 是系统级芯片, 一般是将中央处理器、储存器、控制器、软件系统等集成 在单一芯片上，通常是面向特殊用途的指定产品， 如手机 SoC、电视 SoC、汽车 SoC 等。系统级芯片 能降低开发和生产成本，相比于 ASIC 芯片的开发 周期短，因此更加适合量产商用。目前，高通、AMD、 ARM、英特尔、英伟达、阿里巴巴等都在致力于 SoC 硬件的研发，产品中集成了人工智能加速引 擎，从而满足市场对人工智能应用的需求。英特尔 旗下子公司 Movidius 在 2017 年推出了全球第一个 配备专用神经网络计算引擎SoC（Myriad X） , 芯片上集成了专为高速、低功耗的神经网络而设计 的硬件模块，主要用于加速设备端的深度神经网络 推理计算。赛灵思推出的可编程片上系统(Zynq 系 列) 是基于 ARM 处理器的 SoC，具有高性能、低 功耗、多核和开发灵活的优势。华为推出的升腾 310

是面向计算场景的人工智能 SoC 芯片。

1.2.5 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，以 IBM Truenorth

为代表。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。目前，Truenorth 用三星 28nm 功耗工艺技术，由 54 亿个晶体管组成

的芯片构成的片上网络有 4096 个神经突触核心，实时作业功耗仅为 70mW。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）

的技术实验性的实现了新型突触，加快了商业化进程。

在国内，清华大学类脑计算中心于 2015 年 11 月成功的研制了国内首款超大规模的神

经形态类脑计算天机芯片。该芯片同时支持脉冲神经网络和人工神经网络（深度神经网络），

1314

可进行大规模神经元网络的模拟。中心还开发了面向类脑芯片的工具链，降低应用的开发

难度并提升效率。第二代 28nm 天机芯片也已问世，在性能功耗比上要优于 Truenorth。

当前，类脑 AI 芯片的设计目的不再仅仅局限于加速深度学习算法，而是在芯片基本结

构甚至器件层面上改变设计，希望能够开发出新的类脑计算机体系结构，比如采用忆阻器

和 ReRAM 等新器件来提高存储密度。这类芯片技术尚未完全成熟，离大规模应用还有一

定的差距，但是长期来看类脑芯片有可能会带来计算机体系结构的革命

类脑芯片：类脑芯片是模拟生物神经网络的计算机制。类脑芯片从结构层面去逼近大脑，其研究工作还可进一步分为两个层次，一是神经网络层面，与之相应的是神经拟态架构和处理器，如 IBM 的 TrueNorth 芯片，这种芯片把定制化的数字处理内核当作神经元，把内存作为突触。其逻辑结构与传统冯·诺依曼结构不同：它的内存、CPU 和通信部件完全集成在一起，因此信息的处理在本地进行，克服了传统计算机内存与 CPU 之间的速度瓶颈问题。同时神经元之间可以方便快捷地相互沟通，只要接收到其他神经元发过来的脉冲(动作电位)，这些神经元就会同时做动作。二是神经元与神经突触层面，与之相应的是元器件层面的创新。如 IBM 苏黎世研究中心宣布制造出世界上首个人造纳米尺度的随机相变神经元，可实现高速无监督学习。

类脑芯片仿照人类大脑的信息处理方式，打破了存储和计算分离的架构，实现数据并行传送、分布式处理的低功耗芯片。 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，利用扁平化的设计结构，从而在降低能 耗的前提下高效地完成计算任务。在人工智能火热 的时代，各国政府、大学、公司纷纷投入到类脑芯

片的研究当中，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。

通过以上分析，我们可以总结出以下几个特点。

⚫ CPU通用性最强，但延迟严重，散热高，效率最低。

⚫ GPU 通用性强、速度快、效率高，特别适合用在深度学习训练方面，但是性能功耗比

较低。

⚫ FPGA 具有低能耗、高性能以及可编程等特性，相对于 CPU与 GPU 有明显的性能或者

能耗优势，但对使用者要求高。

⚫ ASIC 可以更有针对性地进行硬件层次的优化，从而获得更好的性能、功耗比。但是

ASIC 芯片的设计和制造需要大量的资金、较长的研发周期和工程周期，而且深度学习

算法仍在快速发展，若深度学习算法发生大的变化，FPGA 能很快改变架构，适应最

新的变化，ASIC 类芯片一旦定制则难于进行修改。

当前阶段，GPU 配合 CPU 仍然是 AI 芯片的主流，而后随着视觉、语音、深度学习的

算法在 FPGA以及 ASIC芯片上的不断优化，此两者也将逐步占有更多的市场份额，从而与

GPU达成长期共存的局面。从长远看，人工智能类脑神经芯片是发展的路径和方向

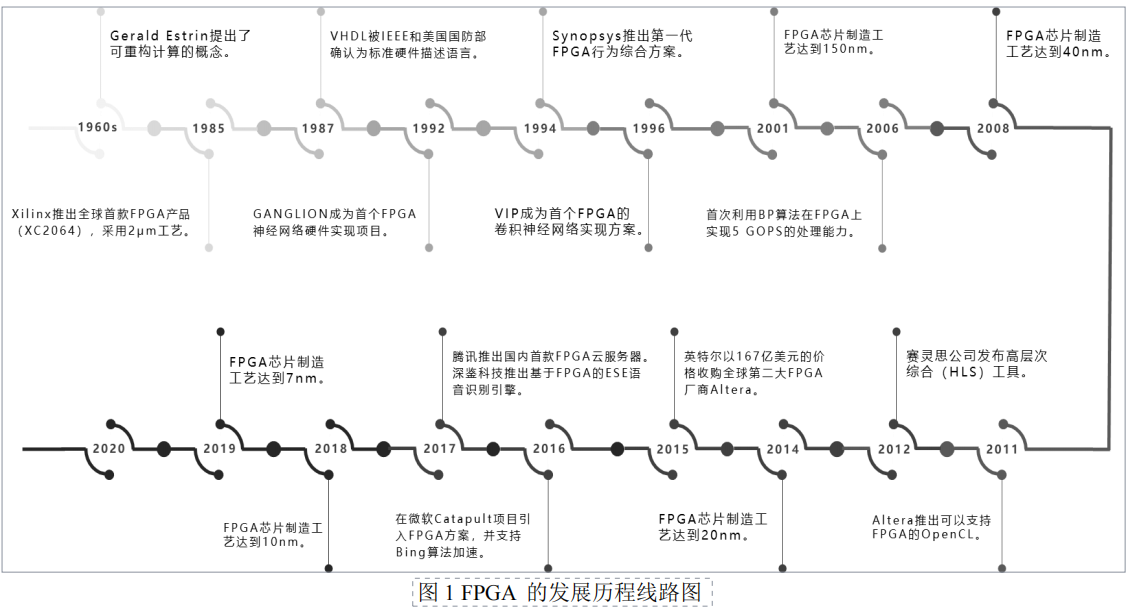
目前，深度神经网络芯片正在不断研究开发中，每种芯片均是针对一定的问题而设计的，因此，不同的芯片有其独特的优势和不足。通过上述对不同芯片的描述，我们可以了解到相比 GPU，FPGA 具有更强的计算能力和较低的功耗，相比 ASIC 和 SoC，FPGA 具有更低的设计成本和灵活的可编程性，相比类脑芯片，FPGA 的开发设计更简单。

**2** FPGA 的基本介绍

近年来，随着人工智能的快速发展，FPGA 由 于其独有的硬件特点成为深度神经网络产业应用 的宠儿。经过近六十年的发展，FPGA 的制作工艺、 封装密度、硬件结构以及开发方式发生了巨大的变 化。本章节将围绕 FPGA 的发展历程以及开发方式 展开论述。

**2.1** 深度神经网络的主流硬件平台

FPGA 是基于可编程逻辑器件发展的一种半定制电路，它可以使用硬件描述语言（Verilog 或 VHDL）或 C/C++/OpenCL 编程，利用小型查找表 来实现组合逻辑，并对 FPGA 上的门电路以及存储 器之间的连线进行调整，从而实现程序功能。早在 20 世纪 60 年代，Gerald Estrin 就提出了可重构计 算的概念。直到 1985 年，Xilinx 推出全球第一款 FPGA 产品 XC2064，该产品采用 2um 制作工艺， 包含了 64 个逻辑单元、85K 个晶体管和数量不超过 1K 个的门。1992 年，GANGLION 成为神经网 络首次在 FPGA 上实现运行的项目[81]，1996 年卷 积神经网络首次在 Altera 的 EPF81500 上实现运行 [82]，随着神经网络的迅速发展，FPGA 做了一系 列针对其需求的开发设计，如 Xilinx 推出的 Versal AI Core 系列和 xDNN 处理引擎为深度神经网络推 断加速带来突破性的改善。另外，为了促进深度神 经网络的发展，不少公司设计提出神经网络编译及 框架，如 ALAMO 编译器和 Lattice 公司设计的 sensAI 编译器，FP-DNN 框架和 FPGAConvNet 框 架。经过 30 多年的发展，FPGA 的制作工艺、逻辑单元和晶体管的封装密度均得到飞速发展，其发展历程线路如图 1 所示。



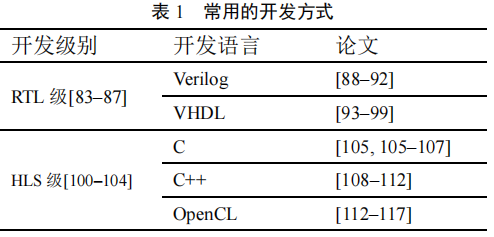
**2.2** FPGA 的开发方式及流程

结合 FPGA 自身硬件架构特点，目前其开发主 要有两种方式：寄存器传输级（RTL）描述和高层 次综合（HLS）描述。寄存器传输级描述，又称 RTL 级描述，是指用寄存器这一级别的描述方式来描述 过 1K 个的门。1992 年，GANGLION 成为神经网络首次在 FPGA 上实现运行的项目[81]，1996 年卷 积神经网络首次在 Altera 的 EPF81500 上实现运行 [82]，随着神经网络的迅速发展，FPGA 做了一系 列针对其需求的开发设计，如 Xilinx 推出的 Versal AI Core 系列和 xDNN 处理引擎为深度神经网络推 断加速带来突破性的改善。另外，为了促进深度神 经网络的发展，不少公司设计提出神经网络编译及 框架，如 ALAMO 编译器和 Lattice 公司设计的 sensAI 编译器，FP-DNN 框架和 FPGAConvNet 框

架。经过 30 多年的发展，FPGA 的制作工艺、逻 辑单元和晶体管的封装密度均得到飞速发展，其发 展历程线路如图 1 所示。

**2.2** FPGA 的开发方式及流程

电路的数据流。开发人员利用硬件描述语言 （Verilog 和 VHDL）或者 IP 核对硬件结构进行描 述。RTL 级开发的主要优势是高稳定性、高资源利 用率、高性能等。其劣势也很突出，主要有开发难 度大、开发效率低、周期长、成本高等。基于此， 利用高级语言实现算法的开发方式应运而生，即 HLS 级，开发人员只需要利用高级语言（C、C++） 实现算法，而算法程序到 FPGA 硬件结构的映射由 编译器自动完成。广义上讲，Xilinx 公司推出了高层次综合 HLS 工具和 Altera 公司主推的 OpenCL SDK 都属于 HLS 级。HLS 级开发的主要优势是门 槛低、开发效率高、周期短，其缺点主要是资源利 用率低、性能低、不透明等。我们对现有论文中的开发方式做了统计，如表 1 所示。



FPGA 的设计流程通常包括设计需求理解、方案评估、芯片理解、详细方案设计、仿真、综合、 布局布线、时序优化、芯片编程与调试。神经网络加速设计开始之前，需要先确定神经网络的网络结构和各层级参数，避免因修改而造成的资源浪费和工程延期。在神经网络算法效果符合预期后，需要对神经网络算法进行拆解和分析。通过分析算法所需的计算带宽，存储带宽，存储容量，关键计算瓶颈，计算数据流等，结合 FPGA芯片的资源特性进 行加速效果推算，把适合 FPGA 做加速的算法交给 FPGA 来加速，其中需要特别注意的是，FPGA 加速计算推演的数据精度需要和软件验证的数据精度保持一致，否则可能在 FPGA 加速后会有精度损失而达不到预期的算法效果。