AI 芯片研究综述：以 GPU 为例

冯一鸣 21307346

引言

近年来，随着深度学习和强化学习等人工智能技术的迅速发展和广泛应用，在语音识别、图像识别、计算机视觉等领域取得了重大突破。然而，由于实际问题的复杂性和学习模型的大规模性，给人工智能应用的推广带来了极大的挑战。由此产生的存储空间和功耗限制，以及巨大的计算需求，使相关的模型难以在个人计算机上进行训练。因此这些模型通常部署在云端，即便如此，算力不足也会导致成本过高，训练周期长等问题。

随着人工智能技术的高速发展，传统芯片已不能满足 AI 产业对芯片性能及算力方面的要求。因此，如何构建出高效的 AI 芯片，将芯片技术与 AI 技术有效地结合起来已成为当前的热点话题。[1]

什么是 AI 芯片

从广义上讲，只要能够运行或加速人工智能算法的芯片都叫作 AI 芯片。但通常意义上的 AI 芯片是一种专门设计用于执行人工智能计算任务的集成电路，人工智能中涉及的运算，如卷积运算，非常复杂，而且涉及到大量重复的计算，由于传统的 CPU 架构缺乏并行性和内存带宽，性能较低，因此出现了 AI 芯片。这些芯片对人工智能算法做了特殊加速设计，所以，AI 芯片也被称为 AI 加速器或 AI 计算芯片，即专门用于处理人工智能应用中的大量计算任务的模块。

为什么要发展 AI 芯片

算法、计算能力和大数据是推动人工智能崛起的三大要素。这三者相辅相成，必须平衡发展，任何一个环节的短板都会妨碍人工智能领域取得进步。计算能力是人工智能的基础，如果计算能力跟不上算法和大数据的发展，人工智能也不会取得突破。近年来，由于大数据产业的发展，数据量呈现爆发式增长，传统的计算架构无法支撑深度学习大量并行计算需求。因此，AI 芯片成为了提供这种计算能力的关键。

AI 芯片的主要特点

为了满足 AI 计算任务的特殊性和高性能要求，AI 芯片必须具备一定的特性。首先，它需要具备高度并行的处理能力，因为大多数 AI 计算任务涉及大规模的并行计算。其次，AI 芯片必须支持高速浮点计算，以处理复杂的数学运算，这对于深度学习等任务至关重要。另外，AI 芯片需要提供大存储器带宽，以实现低内存访问延迟，从而确保计算的高效性。

当前，AI 技术在各个领域都处于快速发展和迭代的阶段。考虑到芯片的研发成本和生产周期，为特定的应用、算法或场景进行定制化设计难以适应技术发展。因此，AI 芯片设计的一个指导原则是针对通用领域而不是特定应用进行设计。这种通用设计能力非常重要，因为它使 AI 芯片能够在多种应用中广泛使用，并且可以通过重新配置以适应新的 AI 算法，从而保持其灵活性和适应性。

另外，在 AI 计算中，卷积运算占据了重要地位，特别是在神经网络的训练

过程中。卷积运算非常复杂且计算成本极高。因此，针对卷积运算的硬件加速成为了一个主要的研究方向。这些针对卷积运算的 AI 芯片具有高运算速度、低功耗、高能效的特点，并经过专门优化以适应应用的需求。这些芯片通常包括一个或多个高性能处理器，以及专门的计算单元和内存结构以支持矩阵乘法和张量计算等 AI 运算。

AI 芯片的分类

根据目前 AI 芯片的研究重点，AI 芯片根据其设计和应用可以分为以下几类：通用处理器、图形处理单元(GPU)、现场可编程门阵列(FPGA)和专用集成电路(ASIC)。[2]

通用处理器(CPU)：通用处理器可以执行各种任务，因此也可以执行通用的 AI 计算，但是针对大规模数据的计算，其性能和能效通常不足。

通用计算 GPU：GPU 最初是为图形渲染而设计的，由于其具有高度并行的结构特点，逐渐应用于 AI 计算。是目前应用最广泛、最成熟的通用 AI 处理器。

现场可编程门阵列：FPGA 是可编程的，具有高灵活性的特点，它利用门电路直接运算、速度较快，但其性能却相对较差。

专用集成电路：ASIC 是为特定用户、应用或任务设计的定制芯片，它通常具备体积小、性能高、功耗低、可靠性高等优点，但是它的灵活性非常低。

总的来说，通用 AI 处理器具备良好的通用性，可以适应多样的程序执行场景，而专用 AI 处理器则具备更好的性能，可以在某些具体应用上达到远超通用 AI 处理器的性能。

本文将以 GPU 为例，阐述 AI 处理器的原理和结构。

GPU 的原理

工作流程

图形处理单元(Graphics Processing Unit, GPU)将要显示的信息进行转换和显示，并向显示器提供扫描信号，控制计算机的显示。它是连接显示器和计算机主板的重要部件。与传统 CPU 相比，GPU 具有更高的并行架构，在处理图形数据和复杂算法时效率更高。

GPU 通常不能单独工作，它需要 CPU 予以支持，目前用 GPU 来运行模型实质上是通过 CUDA 编程来完成。CUDA 编程是针对 GPU 的基于 CUDA c 编程模型的并行计算平台，CUDA 可以通过将程序分解成小任务并在具有大量可用内核的 GPU 上同时运行来加速程序[3]。开发者通过 CUDA 在 CPU 上开启计算和调度，并将数据拷贝到 GPU 里进行处理。GPU 会按照某段自行编写的特定程序，凭借其内部独特的并行机制对数据进行加工，最后将所得的结果发送回 CPU 上，从而完成整个流程的计算。[4]

指挥 GPU 处理数据的特定程序叫核函数，它是一种只在 GPU 上执行的程序。在核函数运行前，开发者需要指定该函数会同时运行多少个线程块以及每个线程块包括多少个线程。GPU 会对这些线程做同步计算，这使得 GPU 特别适用于大规模并行计算的场景，而 CPU 则专门用于处理逻辑复杂任务。

硬件结构

比较 GPU 和 CPU 的结构差异，CPU 的大部分空间是控制器和寄存器，而

GPU 有更多的 ALU(算术逻辑单元)用于并行处理密集数据。结构比较如图 1 所示[2]。

图 1

在上图 GPU 的逻辑结构中，每一个 ALU 都执行一个线程，且每行的多个线程由一个控制器控制，因此这些线程执行的是相同的指令（数据可以不同），所以 GPU 的实现方式是单指令多数据，在向量和矩阵运算上具有极高的性能。

可以看到，执行模块占据 GPU 内部硬件结构面积的绝大部分，并依靠调度模块和存储模块提供支持。与之相比，CPU 在硬件结构上会更突出调度和存储，以保证复杂操作的执行可行性。

接下来深入 GPU 内部，观察 GPU 的内部实现，如下图 2 所示。

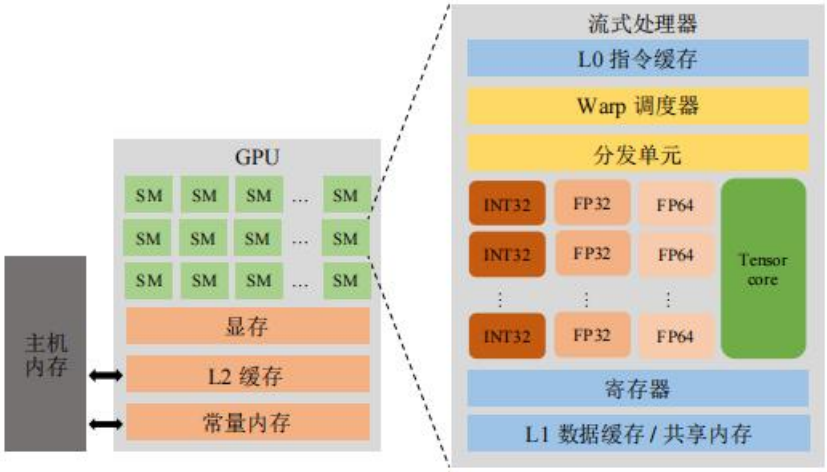

图 2

图 2 中的流多处理器(SM)，也叫 GPU 线程块，对应于图 1GPU 的受控于同一个控制器的一行线程单元；而其中的运算单元，如 INT32 和 FP32 等，对应于图 1 的一个 ALU，也就是线程。CPU 调用 GPU 时，指定 SM 的数量和每个 SM 中使用的线程数，然后将数据和指令发送到 GPU 内存中，再由 GPU 根据指令和数据类型调用相应的执行单元，以 SM 为单位进行处理。

与单核 CPU 相比，GPU 上的程序在特定情况下运行速度往往会提高数十倍甚至数千倍。尤其是对于图像数据的处理，由于图像上的每一个像素点都需要进行处理，产生了相当大的数据量，而图像数据通常又需要进行相同的运算，加之 GPU 采用单指令多数据处理方式，十分适合这种运算的方式，所以使用 GPU，图像处理领域是计算加速中最明显的。GPU 采用了大量的计算单元和超长的管

道，主要解决类似图像处理的领域的运算加速问题。它通用性强，速度快，效率高，特别适合深度学习算法中的卷积运算。但是其设计的初衷是为了应对图像处理中的大规模并行计算。因此，当应用于深度学习算法时，存在三个局限性:第一，不能充分利用并行计算的优势。深度学习包括训练和推理。GPU 在训练过程中效率很高，但在推理过程中，效率一般。二是硬件结构不能灵活配置。GPU 的硬件结构相对固定，不能像 FPGA 那样灵活配置。第三，运行算法的能效低于 FPGA。[2]

GPU 的发展趋势

GPU 在其通用计算得到开发后，目前已经占领了深度学习领域计算设备的半壁江山。在 GPU 和 CPU 混合加速实现方面也取得了一些成果，它介于完全规则的计算(例如，密集矩阵乘法)和不规则的计算(树、链表和图计算)之间。随着人工智能应用的进一步落地和推广，可以预见相关技术的迭代将会不断加快，AI 处理器的市场也会不断扩大。

然而，还有一个更广泛的问题，即 GPU 何时优于 CPU，以及 GPU 优化带来的生产力损失是否被性能提升所抵消。需要进一步的研究和突破。无论如何，如何更好地加速 GPU 的合作将是未来的研究趋势。

另外，在未来，以 GPU 为代表的的通用 AI 处理器将会往特定领域架构的方向发展[5]。例如，GPU 新推出的成品在性能上与前代相比的提升，除了体现在增加了 SM 的数量、升级了 Tensor core 并扩充了其内核和缓存以外，很大程度上还得益于新推出的对低精度数据格式计算的支持以及对特殊处理单元的运用。以后者为例，英伟达决定在最新的 GPU 中开发专门用于自然语言处理的计算模块，是因为该领域对 GPU 的运用与其它领域相比更为广泛，因此可以选择在硬件设计上有所偏重，而这种偏重不影响 GPU 在其它领域的使用。从这个角度上看，未来 AI 处理器的设计方式将会受人工智能应用发展趋势的重大影响，通用 AI 处理器若能寻求在成熟领域中发挥灵活性的机会，往往能获得更大的收益和商机。作为参照，我们也可以看到目前的英伟达已经在往提高 GPU 灵活性的方面进行战略布局。

参考文献：

1. 赵玥,肖梦燕,罗军等.人工智能芯片及测评体系分析[J].电子与封装,2023,23(05):31-37.DOI:10.16257/j.cnki.1681-1070.2023.0055.

2. B. Li, J. Gu and W. Jiang, "Artificial Intelligence (AI) Chip Technology Review," 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 2019, pp. 114-117, doi: 10.1109/MLBDBI48998.2019.00028.

https://ieeexplore.ieee.org/document/8945797

3. Y. Hu, Y. Liu and Z. Liu, "A Survey on Convolutional Neural Network Accelerators: GPU, FPGA and ASIC," 2022 14th International Conference on Computer Research and Development (ICCRD), Shenzhen, China, 2022, pp. 100-107, doi: 10.1109/ICCRD54409.2022.9730377.

https://ieeexplore.ieee.org/abstract/document/9730377

4. R. S. Dehal, C. Munjal, A. A. Ansari and A. S. Kushwaha, "GPU Computing Revolution: CUDA," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2018, pp. 197-201, doi: 10.1109/ICACCCN.2018.8748495.

https://ieeexplore.ieee.org/document/8748495

5. 尹首一,郭珩,魏少军.人工智能芯片发展的现状及趋势[J].科技导报,2018,36(17):45-51.