AI 芯片的过去和未来，看这篇文章就够了

李鲁AI 报道10 月 31 日

本文授权转载自微信公众号：**硅谷密探**

相信你一定还记得击败了李世石和柯洁的谷歌“阿尔法狗”（Alpha Go），那你知道驱动 Alpha Go 的是什么吗？

如果你觉得 Alpha Go 和人相似，只不过是把人脑换成了芯片，那么你就大错特错了。击败李世石的 Alpha Go 装有 48 个谷歌的 AI 芯片，而这 48 个芯片不是安装在 Alpha Go 身体里，而是在云端。所以，真正驱动 Alpha Go 的装置，看上去是这样的...

图片来自网络，版权属于作者

因此李世石和柯洁不是输给了“机器人”，而是输给了装有 AI 芯片的云工作站。

然而近几年，AI 技术的应用场景开始向移动设备转移，比如汽车上的自动驾驶、手机上的人脸识别等。产业的需求促成了技术的进步，而 AI 芯片作为产业的根基，必须达到更强的性能、更高的效率、更小的体积，才能完成 AI 技术从云端到终端的转移。

目前，AI 芯片的研发方向主要分两种：一是基于传统冯·诺依曼架构的 FPGA（现场可编程门阵列）和 ASIC（专用集成电路）芯片，二是模仿人脑神经元结构设计的类脑芯片。其中 FPGA 和 ASIC 芯片不管是研发还是应用，都已经形成一定规模；而类脑芯片虽然还处于研发初期，但具备很大潜力，可能在未来成为行业内的主流。

这两条发展路线的主要区别在于，前者沿用冯·诺依曼架构，后者采用类脑架构。你看到的每一台电脑，采用的都是冯·诺依曼架构。它的核心思路就是处理器和存储器要分开，所以才有了 CPU（中央处理器）和内存。而类脑架构，顾名思义，模仿人脑神经元结构，因此 CPU、内存和通信部件都集成在一起。

接下来小探将为读者分别介绍两种架构的简要发展史、技术特点和代表性产品。

从 GPU 到 FPGA 和 ASIC 芯片

2007 年以前，受限于当时算法和数据等因素，AI 对芯片还没有特别强烈的需求，通用的 CPU 芯片即可提供足够的计算能力。比如现在在读这篇文章的你，手机或电脑里就有 CPU 芯片。

之后由于高清视频和游戏产业的快速发展，GPU （图形处理器）芯片取得迅速的发展。因为 GPU 有更多的逻辑运算单元用于处理数据，属于高并行结构，在处理图形数据和复杂算法方面比 CPU 更有优势，又因为 AI 深度学习的模型参数多、数据规模大、计算量大，此后一段时间内 GPU 代替了 CPU，成为当时 AI 芯片的主流。

GPU 比 CPU 有更多的逻辑运算单元（ALU）

图片来自网络，版权属于作者

然而 GPU 毕竟只是图形处理器，不是专门用于 AI 深度学习的芯片，自然存在不足，比如在执行 AI 应用时，其并行结构的性能无法充分发挥，导致能耗高。

与此同时，AI 技术的应用日益增长，在教育、医疗、无人驾驶等领域都能看到 AI 的身影。然而 GPU 芯片过高的能耗无法满足产业的需求，因此取而代之的是 FPGA 芯片，和 ASIC 芯片。

那么这两种芯片的技术特点分别是什么呢？又有什么代表性的产品呢？

“万能芯片” FPGA

FPGA（FIELD-PROGRAMMABLE GATE ARRAY），即 “现场可编程门阵列”，是在 PAL、GAL、CPLD 等可编程器件的基础上进一步发展的产物。

FPGA 可以被理解为“万能芯片”。用户通过烧入 FPGA 配置文件，来定义这些门电路以及存储器之间的连线，用硬件描述语言（HDL）对 FPGA 的硬件电路进行设计。每完成一次烧录，FPGA 内部的硬件电路就有了确定的连接方式，具有了一定的功能，输入的数据只需要依次经过各个门电路，就可以得到输出结果。

用大白话说，“万能芯片” 就是你需要它有哪些功能、它就能有哪些功能的芯片。

尽管叫“万能芯片”，FPGA 也不是没有缺陷。正因为 FPGA 的结构具有较高灵活性，量产中单块芯片的成本也比 ASIC 芯片高，并且在性能上，FPGA 芯片的速度和能耗相比 ASIC 芯片也做出了妥协。

也就是说，“万能芯片” 虽然是个 “多面手”，但它的性能比不上 ASIC 芯片，价格也比 ASIC 芯片更高。

但是在芯片需求还未成规模、深度学习算法需要不断迭代改进的情况下，具备可重构特性的 FPGA 芯片适应性更强。因此用 FPGA 来实现半定制人工智能芯片，毫无疑问是保险的选择。

目前，FPGA 芯片市场被美国厂商 Xilinx 和 Altera 瓜分。据国外媒体 Marketwatch 的统计，前者占全球市场份额 50%、后者占 35%左右，两家厂商霸占了 85% 的市场份额，专利达到 6000 多项，毫无疑问是行业里的两座大山。

Xilinx 的 FPGA 芯片从低端到高端，分为四个系列，分别是 Spartan、Artix、Kintex、Vertex，芯片工艺也从 45 到 16 纳米不等。芯片工艺水平越高，芯片越小。其中 Spartan 和 Artix 主要针对民用市场，应用包括无人驾驶、智能家居等；Kintex 和 Vertex 主要针对军用市场，应用包括国防、航空航天等。

Xilinx 的 Spartan 系列 FPGA 芯片

图片来自网络，版权属于作者

我们再说说 Xilinx 的老对手 Altera。Altera 的主流 FPGA 芯片分为两大类，一种侧重低成本应用，容量中等，性能可以满足一般的应用需求，如 Cyclone 和 MAX 系列；还有一种侧重于高性能应用，容量大，性能能满足各类高端应用，如 Startix 和 Arria 系列。Altera 的 FPGA 芯片主要应用在消费电子、无线通信、军事航空等领域。

专用集成电路 ASIC

在 AI 产业应用大规模兴起之前，使用 FPGA 这类适合并行计算的通用芯片来实现加速，可以避免研发 ASIC 这种定制芯片的高投入和风险。

但就像我们刚才说到的，由于通用芯片的设计初衷并非专门针对深度学习，因此 FPGA 难免存在性能、功耗等方面的瓶颈。随着人工智能应用规模的扩大，这类问题将日益突出。换句话说，我们对人工智能所有的美好设想，都需要芯片追上人工智能迅速发展的步伐。如果芯片跟不上，就会成为人工智能发展的瓶颈。

所以，随着近几年人工智能算法和应用领域的快速发展，以及研发上的成果和工艺上的逐渐成熟，ASIC 芯片正在成为人工智能计算芯片发展的主流。

ASIC 芯片是针对特定需求而定制的专用芯片。虽然牺牲了通用性，但 ASIC 无论是在性能、功耗还是体积上，都比 FPGA 和 GPU 芯片有优势，特别是在需要芯片同时具备高性能、低功耗、小体积的移动端设备上，比如我们手上的手机。

但是，因为其通用性低，ASIC 芯片的高研发成本也可能会带来高风险。**然而如果考虑市场因素，ASIC 芯片其实是行业的发展大趋势。**

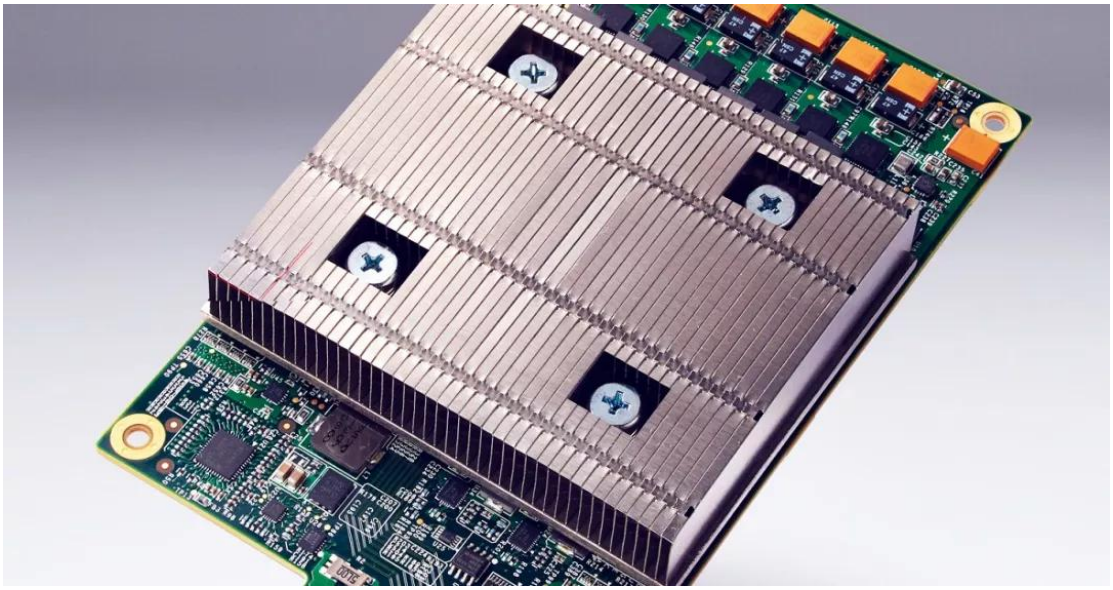
为什么这么说呢？因为从服务器、计算机到无人驾驶汽车、无人机，再到智能家居的各类家电，海量的设备需要引入人工智能计算能力和感知交互能力。出于对实时性的要求，以及训练数据隐私等考虑，这些能力不可能完全依赖云端，必须要有本地的软硬件基础平台支撑。而 ASIC 芯片高性能、低功耗、小体积的特点恰好能满足这些需求。

ASIC 芯片市场百家争鸣

2016 年，英伟达发布了专门用于加速 AI 计算的 Tesla P100 芯片，并且在 2017 年升级为 Tesla V100。在训练超大型神经网络模型时，Tesla V100 可以为深度学习相关的模型训练和推断应用提供高达 125 万亿次每秒的张量计算（张量计算是 AI 深度学习中最经常用到的计算）。然而在最高性能模式下，Tesla V100 的功耗达到了 300W，**虽然性能强劲，但也毫无疑问是颗“核弹”，因为太费电了**。

英伟达 Tesla V100 芯片
图片来自网络，版权属于作者

同样在 2016 年，谷歌发布了加速深度学习的 TPU（Tensor Processing Unit）芯片，并且之后升级为 TPU 2.0 和 TPU 3.0。与英伟达的芯片不同，谷歌的 TPU 芯片设置在云端，就像文章在 Alpha Go 的例子中说的一样，并且“只租不卖“，服务按小时收费。不过谷歌 TPU 的性能也十分强大，算力达到 180 万亿次每秒，并且功耗只有 200w。



谷歌 TPU 芯片

图片来自网络，版权属于作者

关于各自 AI 芯片的性能，谷歌 CEO Sundar Pichai 和英伟达 CEO 黄仁勋之前还在网上产生过争论。别看两位大佬为自家产品撑腰，争得不可开交，实际上不少网友指出，这两款产品没必要“硬做比较”，因为一个是在云端，一个是在终端。

除了大公司，初创企业也在激烈竞争 ASIC 芯片市场。那么初创企业在行业中该如何生存呢？对此，AI 芯片初创企业 Novumind 的中国区 CEO 周斌告诉小探：创新是初创企业的核心竞争力。

2017 年，NovuMind 推出了第一款自主设计的 AI 芯片：NovuTensor。这款芯片使用原生张量处理器（Native Tensor Processor）作为内核构架，这种内核架构由 NovuMind 自主研发，并在短短一年内获得美国专利。除此之外，NovuTensor 芯片采用不同的异构计算模式来应对不同 AI 应用领域的三维张量计算。2018 年下半年，Novumind 刚推出

了新一代 NovuTensor 芯片，这款芯片在做到 15 万亿次计算每秒的同时，全芯片功耗控制在 15W 左右，效率极高。

Novumind 的 NovuTensor 芯片

尽管 NovuTensor 芯片的纸面算力不如英伟达的芯片，**但是其计算延迟和功耗却低得多，因此适合边缘端 AI 计算，也就是服务于物联网**。

虽然大家都在追求高算力，但实际上不是所有芯片都需要高算力的。比如用在手机、智能眼镜上的芯片，虽然也对算力有一定要求，但更需要的是低能耗，否则你的手机、智能眼镜等产品，用几下就没电了，也是很麻烦的一件事情。并且据 EE Times 的报道，在运行 ResNet-18、ResNet-34、ResNet70、VGG16 等业界标准神经网络推理时，

NovuTensor 芯片的吞吐量和延迟都要优于英伟达的另一款高端芯片 Xavier。

结合 Novumind 现阶段的成功，我们不难看出：在云端市场目前被英伟达、谷歌等巨头公司霸占，终端应用芯片群雄逐鹿的情形下，专注技术创新，在关键指标上大幅领先所有竞争对手，或许是 AI 芯片初创企业的生存之道。

类脑芯片

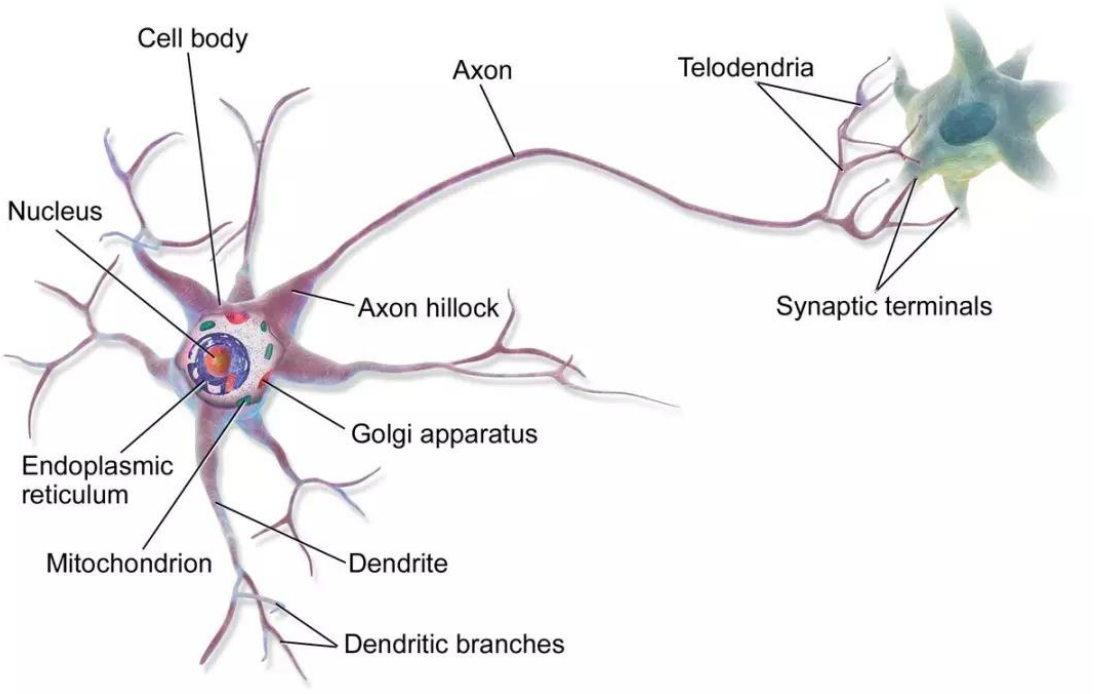
如文章开头所说，目前所有电脑，包括以上谈到的所有芯片，都基于冯·诺依曼架构。

然而这种架构并非十全十美。**将 CPU 与内存分开的设计，反而会导致所谓的冯·诺伊曼瓶颈（von Neumann bottleneck）**：CPU 与内存之间的资料传输率，与内存的容量和 CPU 的工作效率相比都非常小，因此当 CPU 需要在巨大的资料上执行一些简单指令时，资料传输率就成了整体效率非常严重的限制。

既然要研制人工智能芯片，那么有的专家就回归问题本身，开始模仿人脑的结构。

人脑内有上千亿个神经元，而且每个神经元都通过成千上万个突触与其他神经元相连，形成超级庞大的神经元回路，以分布式和并发式的方式

传导信号，相当于超大规模的并行计算，因此算力极强。人脑的另一个特点是，不是大脑的每个部分都一直在工作，从而整体能耗很低。



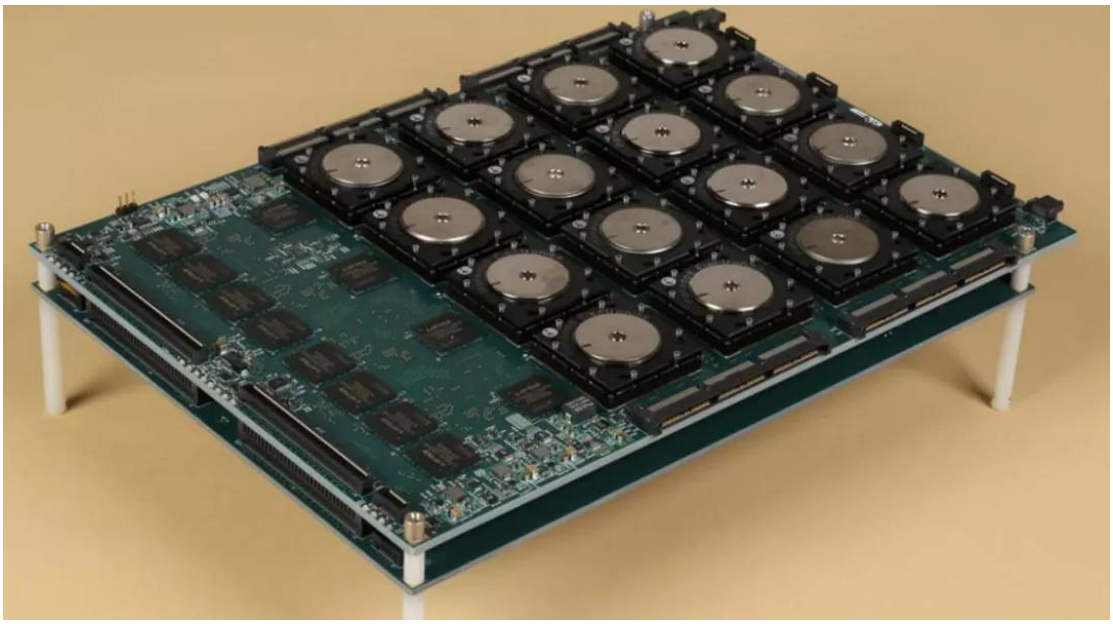
神经元结构

图片来源：维基百科

这种类脑芯片跟传统的冯·诺依曼架构不同，它的内存、CPU 和通信部件是完全集成在一起，把数字处理器当作神经元，把内存作为突触。除此之外，在类脑芯片上，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了。同时，神经元只要接收到其他神经元发过来的脉冲，这些神经元就会同时做动作，因此神经元之间可以方便快捷地相互沟通。

在类脑芯片的研发上，IBM 是行业内的先行者。2014 年 IBM 发布了 TrueNorth 类脑芯片，这款芯片在直径只有几厘米的方寸的空间里，集

成了 4096 个内核、100 万个“神经元”和 2.56 亿个“突触”，**能耗只有不到 70 毫瓦**，可谓是高集成、低功耗的完美演绎。



装有 16 个 TrueNorth 芯片的 DARPA SyNAPSE 主板
图片来自网络，版权属于作者

那么这款芯片的实战表现如何呢？IBM 研究小组曾经利用做过 DARPA 的 NeoVision2 Tower 数据集做过演示。它能以 30 帧每秒速度，实时识别出街景视频中的人、自行车、公交车、卡车等，准确率达到了 80%。相比之下，一台笔记本编程完成同样的任务用时要慢 100 倍，能耗却是 IBM 芯片的 1 万倍。

然而目前类脑芯片研制的挑战之一，是在硬件层面上模仿人脑中的神经突触，换而言之就是设计完美的人造突触。

在现有的类脑芯片中，通常用施加电压的方式来模拟神经元中的信息传输。但存在的问题是，由于大多数由非晶材料制成的人造突触中，离子通过的路径有无限种可能，难以预测离子究竟走哪一条路，造成不同神经元电流输出的差异。

针对这个问题，今年麻省理工的研究团队制造了一种类脑芯片，其中的人造突触由硅锗制成，每个突触约 25 纳米。对每个突触施加电压时，所有突触都表现出几乎相同的离子流，突触之间的差异约为 4%。与无定形材料制成的突触相比，其性能更为一致。

即便如此，类脑芯片距离人脑也还有相当大的距离，**毕竟人脑里的神经元个数有上千亿个，而现在最先进的类脑芯片中的神经元也只有几百万个，连人脑的万分之一都不到**。因此这类芯片的研究，离成为市场上可以大规模广泛使用的成熟技术，还有很长的路要走，但是长期来看类脑芯片有可能会带来计算体系的革命。