人工智能之终端芯片研究报告

一、人工智能与深度学习

2016 年，AlphaGo 与李世石九段的围棋对决无疑掀起了全世界对人工智能领域的新一轮关注。在与李世石对战的 5 个月之前，AlphaGo 因击败欧洲围棋冠军樊麾二段，围棋等级分上升至 3168 分，而当时排名世界第二的李世石是 3532 分。按照这个等级分数对弈，AlphaGo 每盘的胜算只有约 11%，而结果是 3 个月之后它在与李世石对战中以 4 比 1 大胜。AlphaGo 的学习能力之快，让人惶恐。

1.人工智能：让机器像人一样思考

自 AlphaGo 之后，“人工智能”成为 2016 年的热词，但早在 1956 年，几个计算机科学家就在达特茅斯会议上首次提出了此概念。他们梦想着用当时刚刚出现的计算机来构造复杂

的、拥有与人类智慧同样本质特性的机器，也就是我们今日所说的“强人工智能”。这个无所不能的机器，它有着我们所有的感知、所有的理性，甚至可以像我们一样思考。

人们在电影里也总是看到这样的机器：友好的，像星球大战中的 C-3PO；邪恶的，如终结者。强人工智能目前还只存在于电影和科幻小说中，原因不难理解，我们还没法实现它们，至少目前还不行。

我们目前能实现的，一般被称为“弱人工智能”。弱人工智能是能够与人一样，甚至比人更好地执行特定任务的技术。例如，Pinterest 上的图像分类，或者 Facebook 的人脸识别。这些人工智能技术实现的方法就是“机器学习”。

2.机器学习：使人工智能真实发生

人工智能的核心就是通过不断地机器学习，而让自己变得更加智能。机器学习最基本的做法，是使用算法来解析数据、从中学习，然后对真实世界中的事件做出决策和预测。与传统的为解决特定任务、硬编码的软件程序不同，机器学习是用大量的数据来“训练”，通过各种算法从数据中学习如何完成任务。

机器学习最成功的应用领域是计算机视觉，虽然也还是需要大量的手工编码来完成工作。以识别停止标志牌为例：人们需要手工编写形状检测程序来判断检测对象是不是有八条边；写分类器来识别字母“S-T-O-P”。使用以上这些手工编写的分类器与边缘检测滤波器，人们

总算可以开发算法来识别标志牌从哪里开始、到哪里结束，从而感知图像，判断图像是不是一个停止标志牌。

这个结果还算不错，但并不是那种能让人为之一振的成功。特别是遇到雾霾天，标志牌变得不是那么清晰可见，又或者被树遮挡一部分，算法就难以成功了。这就是为什么很长一段时间，计算机视觉的性能一直无法接近到人的能力。它太僵化，太容易受环境条件的干扰。

3.人工神经网络：赋予机器学习以深度

人工神经网络是早期机器学习中的一个重要的算法，历经数十年风风雨雨。神经网络的原理是受我们大脑的生理结构——互相交叉相连的神经元启发。但与大脑中一个神经元可以连接一定距离内的任意神经元不同，人工神经网络具有离散的层，每一次只连接符合数据传播方向的其它层。

例如，我们可以把一幅图像切分成图像块，输入到神经网络的第一层。在第一层的每一个神经元都把数据传递到第二层。第二层的神经元也是完成类似的工作，把数据传递到第三层，以此类推，直到最后一层，然后生成结果。

每一个神经元都为它的输入分配权重，这个权重的正确与否与其执行的任务直接相关。最终的输出由这些权重加总来决定。

我们仍以停止标志牌为例 将一个停止标志牌图像的所有元素都打碎 ,然后用神经元进行“检查”：八边形的外形、救火车般的红颜色、鲜明突出的字母、交通标志的典型尺寸和静止不动运动特性等等。神经网络的任务就是给出结论，它到底是不是一个停止标志牌。神经网络会根据所有权重，给出一个经过深思熟虑的猜测——“概率向量”。

这个例子里，系统可能会给出这样的结果：86%可能是一个停止标志牌；7%的可能是一个限速标志牌；5%的可能是一个风筝挂在树上等等。然后网络结构告知神经网络，它的结论是否正确。

即使是这个例子，也算是比较超前了。直到前不久，神经网络也还是为人工智能圈所淡忘。其实在人工智能出现的早期，神经网络就已经存在了，但神经网络对于“智能”的贡献微乎其微。主要问题是，即使是最基本的神经网络，也需要大量的运算，而这种运算需求难以得到满足。

4.深度学习：剔除神经网络之误差

深度学习由人工神经网络衍生而来 ,是一种需要训练的具有大型神经网络的多隐层层次结构，其每层相当于一个可以解决问题不同方面的机器学习。利用这种深层非线性的网络结构，深度学习可以实现复杂函数的逼近，将表征输入数据分布式表示，继而展现强大的从少数样本集中学习数据集本质特征的能力，并使概率向量更加收敛。

简单来说，深度学习神经网络对数据的处理方式和学习方式与人类大脑的神经元更加相似，比传统的神经网络更准确。

我们回过头来看这个停止标志识别的例子 :深度学习神经网络从成百上千甚至几百万张停止标志图像中提取表征数据，通过重复训练将神经元输入的权重调制得更加精确，无论是否有雾，晴天还是雨天，每次都能得到正确的结果。只有这个时候，我们才可以说神经网络成功地自学习到一个停止标志的样子。

Google 的 AlphaGo 也是先学会了如何下围棋，然后通过不断地与自己下棋，训练自己的神经网络，这种训练使得 AlphaGo 成功在三个月后击败了等级分数更高的李世石。

二、深度学习的实现

深度学习仿若机器学习最顶端的钻石，赋予人工智能更璀璨的未来。其摧枯拉朽般地实现了各种我们曾经想都不敢想的任务，使得几乎所有的机器辅助功能都变为可能。更好的电影推荐、智能穿戴，甚至无人驾驶汽车、预防性医疗保健，都近在眼前，或者即将实现。人工智能就在现在，就在明天。你的 C-3PO 我拿走了，你有你的终结者就好。

但是正如前面提到的，人工神经网络，即深度学习的前身，已经存在了近三十年，但直到最近的 5 到 10 年才再次兴起，这又是因为什么?

1.突破局限的学习算法

20 世纪 90 年代，包括支撑向量机（SVM）与最大熵方法（LR）在内的众多浅层机器学习算法相继提出，使得基于反向传播算法（BP）的人工神经网络因难以弥补的劣势渐渐淡出人们的视线。直到 2006 年，加拿大多伦多大学教授、机器学习领域的泰斗 Geoffrey Hinton 和他的学生在《科学》上发表了一篇文章，解决了反向传播算法存在的过拟合与难训练的问题，从而开启了深度学习在学术界和工业界的浪潮。

深度学习的实质，是通过构建具有很多隐层的机器学习模型和海量的训练数据，来学习更有用的特征，从而最终提升分类或预测的准确性。因此，“深度模型”是手段，“特征学习”是目的。**区别于传统的浅层学习，深度学习的不同在于：**

- 强调了模型结构的深度，通常有 5 层、6 层，甚至 10 多层的隐层节点；

- 明确突出了特征学习的重要性，也就是说，通过逐层特征变换，将样本在原空间的特征表示变换到一个新特征空间，从而使分类或预测更加容易。

这种算法的差别提升了对训练数据量和并行计算能力的需求，而在当时，移动设备尚未普及，这使得非结构化数据的采集并不是那么容易。

2.骤然爆发的数据洪流

深度学习模型需要通过大量的数据训练才能获得理想的效果。以语音识别问题为例，仅在其声学建模部分，算法就面临着十亿到千亿级别的训练样本数据。训练样本的稀缺使得人工智

能即使在经历了算法的突破后依然没能成为人工智能应用领域的主流算法。直到 2012 年，分布于世界各地的互相联系的设备、机器和系统促进了非结构化数据数量的巨大增长，并终于在可靠性方面发生了质的飞跃，大数据时代到来。

大数据到底有多大？一天之中，互联网产生的全部内容可以刻满 1.68 亿张 DVD；发出的邮件有 2940 亿封之多，相当于美国两年的纸质信件数量；发出的社区帖子达 200 万个，相当于《时代》杂志 770 年的文字量；卖出的手机为 37.8 万台，高于全球每天出生的婴儿数量 37.1 万倍。然而，即使是人们每天创造的全部信息，包括语音通话、电子邮件和信息在内的各种通信，以及上传的全部图片、视频与音乐，其信息量也无法匹及每一天所创造出的关于人们自身活动的数字信息量。

我们现在还处于所谓“物联网”的最初级阶段，随着技术的成熟，我们的通讯设备、交通工具和可穿戴科技将能互相连接与沟通，信息量的增加也将以几何倍数持续下去。

3.难以满足的硬件需求

骤然爆发的数据洪流满足了深度学习算法对于训练数据量的要求，但是算法的实现还需要相应处理器极高的运算速度作为支撑。当前流行的包括 X86 和 ARM 在内的传统 CPU 处理器架构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，但对于并不需要太多的程序指令，却需要海量数据运算的深度学习的计算需求，这种结构就显得非常笨拙。尤其是在当前功耗限制下无法通过提升 CPU 主频来加快指令执行速度，这种矛盾愈发不可调和，深度学习研究人员迫切需要一种替代硬件来满足海量数据的运算需求。

或许终有一日将会诞生全新的、为人工智能而专门设计的处理器架构，但在那之前的几十年，人工智能仍然要向前走，便只能改进现有处理器，使之成为能够最大程度适应大吞吐量运算的计算架构。**目前来看，围绕现有处理器的主流改进方式有两个：**

- **图形处理器通用化：**

将图形处理器 GPU 用作矢量处理器。在这种架构中，GPU 擅长浮点运算的特点将得到充分利用，使其成为可以进行并行处理的通用计算芯片 GPGPU。英伟达公司从 2006 年下半年已经开始陆续推出相关的硬件产品以及软件开发工具，目前是人工智能硬件市场的主导。

- **多核处理器异构化：**

将 GPU 或 FPGA 等其他处理器内核集成到 CPU 上。在这种架构中，CPU 内核所不擅长的浮点运算以及信号处理等工作，将由集成在同一块芯片上的其它可编程内核执行，而 GPU 与 FPGA 都以擅长浮点运算著称。AMD 与 Intel 公司分别致力于基于 GPU 与 FPGA 的异构处理器，希望借此切入人工智能市场。

三、现有市场——通用芯片 GPU

在深度学习的领域里，最重要的是数据和运算。谁的数据更多，谁的运算更快，谁就会占据优势。因此，在处理器的选择上，可以用于通用基础计算且运算速率更快的 GPU 迅速成为

人工智能计算的主流芯片。可以说，在过去的几年，尤其是 2015 年以来，人工智能大爆发就是由于英伟达公司的 GPU 得到广泛应用，使得并行计算变得更快、更便宜、更有效。

1.GPU 是什么？

图形处理器 GPU 最初是用在个人电脑、工作站、游戏机和一些移动设备上运行绘图运算工作的微处理器，可以快速地处理图像上的每一个像素点。后来科学家发现，其海量数据并行运算的能力与深度学习需求不谋而合，因此，被最先引入深度学习。2011 年吴恩达教授率先将其应用于谷歌大脑中便取得惊人效果，结果表明，12 颗英伟达的 GPU 可以提供相当于 2000 颗 CPU 的深度学习性能，之后纽约大学、多伦多大学以及瑞士人工智能实验室的研究人员纷纷在 GPU 上加速其深度神经网络。

2.GPU 和 CPU 的设计区别

那么 GPU 的快速运算能力是如何获得的？这就要追溯到芯片最初的设计目标了。中央处理器 CPU 需要很强的处理不同类型数据的计算能力以及处理分支与跳转的逻辑判断能力，这些都使得 CPU 的内部结构异常复杂；而图形处理器 GPU 最初面对的是类型高度统一的、相互无依赖的大规模数据和不需要被打断的纯净的计算环境，所以 GPU 只需要进行高速运算而不需要逻辑判断。**目标运算环境的区别决定了 GPU 与 CPU 不同的设计架构：**

CPU 基于低延时的设计

• 大量缓存空间 Cache，方便快速提取数据。CPU 将大量访问过的数据存放在 Cache 中，当需要再次访问这些数据时 ,就不用从数据量巨大的内存中提取了 ,而是直接从缓存中提取。

• 强大的算术运算单元 ALU，可以在很短的时钟周期内完成算数计算。当今的 CPU 可以达到 64bit 双精度，执行双精度浮点源计算加法和乘法只需要 1 ~ 3 个时钟周期，时钟周期频率达到 1.532 ~ 3gigahertz。

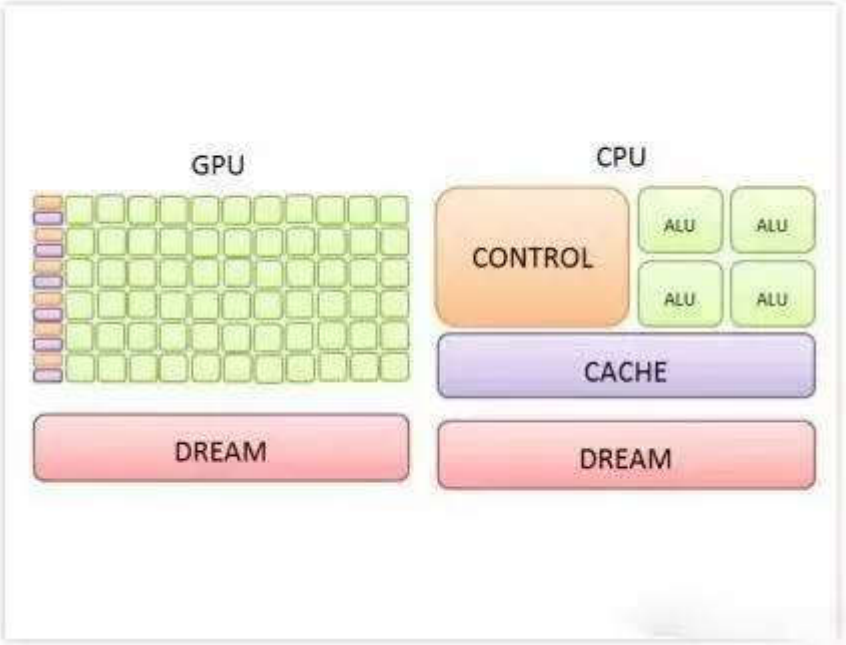
• 复杂的逻辑控制单元，当程序含有多个分支时，它通过提供分支预测来降低延时。

• 包括对比电路单元与转发电路单元在内的诸多优化电路 ,当一些指令依赖前面的指令结果时 ,它决定这些指令在 pipeline 中的位置并且尽可能快的转发一个指令的结果给后续指令。

GPU 基于大吞吐量的设计

• 压缩缓存空间 Cache，从而最大化激发内存吞吐量，可以处理超长的流水线。缓存的目的不是保存之后需要访问的数据 ,而是担任数据转发的角色 ,为线程提高服务。如果有很多线程需要访问同一个数据，缓存会合并这些访问，再去 DRAM 中访问数据，获取的数据将通过缓存转发给对应的线程。这种方法虽然减小了缓存 ,但由于需要访问内存 ,因而自然会带来延时效应。

• 高效的算数运算单元和简化的逻辑控制单元，把串行访问拆分成多个简单的并行访问，并同时运算。例如，在 CPU 上约有 20%的晶体管是用作计算的，而 GPU 上有 80%的晶体管用作计算。



3.GPU 和 CPU 的性能差异

CPU 与 GPU 在各自领域都可以高效地完成任务，但当同样应用于通用基础计算领域时，设计架构的差异直接导致了两种芯片性能的差异。

CPU 拥有专为顺序逻辑处理而优化的几个核心组成的串行架构，这决定了其更擅长逻辑控制、串行运算与通用类型数据运算；而 GPU 拥有一个由数以千计的更小、更高效的核心组成的大规模并行计算架构，大部分晶体管主要用于构建控制电路和 Cache，而控制电路也相对简单，且对 Cache 的需求小，只有小部分晶体管来完成实际的运算工作。所以大部分

晶体管可以组成各类专用电路、多条流水线，使得 GPU 的计算速度有了突破性的飞跃，拥有了更强大的处理浮点运算的能力。这决定了其更擅长处理多重任务，尤其是没有技术含量的重复性工作。

当前最顶级的 CPU 只有 4 核或者 6 核，模拟出 8 个或者 12 个处理线程来进行运算，但是普通级别的 GPU 就包含了成百上千个处理单元，高端的甚至更多，这对于多媒体计算中大量的重复处理过程有着天生的优势。

举个常见的例子，一个向量相加的程序，可以让 CPU 跑一个循环，每个循环对一个分量做加法，也可以让 GPU 同时开大量线程，每个并行的线程对应一个分量的相加。CPU 跑循环的时候每条指令所需时间一般低于 GPU，但 GPU 因为可以同时开启大量的线程并行地跑，具有 SIMD 的优势。

4.GPU 行业的佼佼者：Nvidia

目前全球 GPU 行业的市场份额有超过 70%被英伟达公司占据，而应用在人工智能领域的可进行通用计算的 GPU 市场则基本被英伟达公司垄断。

2016 年三季度英伟达营收为 20.04 亿美元，较上年同期的 13.05 亿美元增长 54%；净利润为 5.42 亿美元，较上年同期的 2.46 亿美元增长 120%，营收的超预期增长推动其盘后股价大幅上涨约 16%。以面向的市场平台来划分，游戏业务营收 12.4 亿美元，同比增长 63%，

是创造利润的核心部门；数据中心业务营收 2.4 亿美元，同比增长 193%，成为增长最快的部门；自动驾驶业务营收 1.27 亿美元，同比增长 61%，正在逐步打开市场。

Revenue by Market Platform					
($ in millions)	Q3 FY17	Q2 FY17	Q3 FY16	Q/Q	Y/Y
Gaming	$1,244	$781	$761	Up 59 %	Up 63 %
Professional Visualization	207	214	190	Down 3 %	Up 9 %
Datacenter	240	151	82	Up 59 %	Up 193 %
Automotive	127	119	79	Up 7 %	Up 61 %
OEM and IP	186	163	193	Up 14 %	Down 4 %
Total	$2,004	$1,428	$1,305	Up 40 %	Up 54 %

这样的业绩创下了英伟达的历史最好季度收入，但这并非是其股票暴涨的理由，事实上，在过去的六年里，英伟达的业绩基本一直呈现上升趋势。从 2012 年财年至 2016 财年，英伟达的营业收入实现了从 40 亿美元到 50 亿美元的跨越，而其净利润也从 2012 财年的 5.8 亿美元逐步上升到了 2016 财年的 6.14 亿美元。但在此期间，英伟达的股价并未出现翻番式的增长。

真正促成英伟达股价飙升的是人工智能的新市场。在刚刚过去的 2016 年，英伟达的股价上涨了 228%，过去的 5 年内累计上涨 500%。500 亿美元的市值将会持续给英伟达带来 40 倍的市场收入，这几乎是业内拥有最高收益的公司。

5.Nvidia 的市场定位：人工智能计算公司

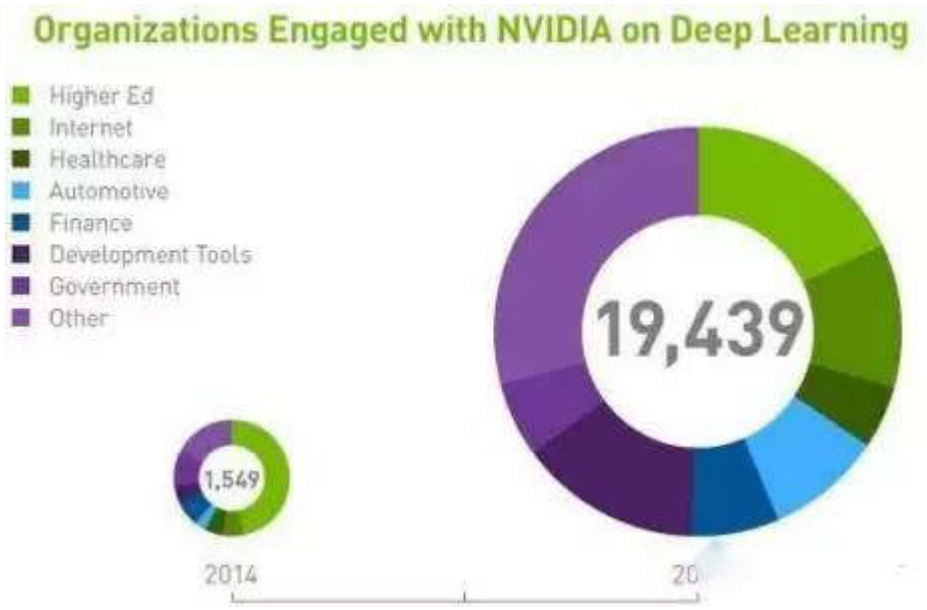
自 1999 年发布第一款 GPU 以来，GPU 就成为了英伟达最为核心的产品，占到了英伟达总营业收入的八成，而英伟达也以显卡厂商的身份进入人们的视线。这些芯片最初是以板卡的

形式出售给游戏玩家的，游戏玩家需要自己动手将芯片装到 PC 主板上，从而拥有更快的 3D 图形处理速度。他们的产品命名也很有讲究，用"GeForce"这样具有超能力的字眼来开辟市场。

今日的英伟达，已经不再是一家单纯的显卡技术厂商，他现在很赶时髦地称自己为“人工智能计算公司”。据英伟达官网数据显示，2016 年，有近两万家机构将英伟达产品用于深度学习加速计算，相比 2014 年翻了 13 倍。医疗、生命科学、教育、能源、金融、汽车、制造业以及娱乐业等诸多行业均将得益于海量数据的分析。

谷歌、微软、Facebook 和亚马逊等技术巨头大量购买英伟达的芯片来扩充自己数据中心的处理能力；Massachusetts General Hospital 等医疗研究机构用英伟达的芯片来标记 CT 扫描图片上的病变点；特斯拉将在所有的汽车上安装英伟达的芯片来实现无人驾驶； June 等家电公司用英伟达的芯片制造人工智能驱动的家用电器。在人工智能到来之前，英伟达从来都没有处于一个如此巨大的市场的中心，这也充分表明了一个事实，那就是英伟达在 GPU 的计算处理技术上无人能及。

同时，英伟达还在投资不同领域里新兴的、需要借助深度学习来构建业务的公司，使这些公司能够更好地借助其提供的人工智能平台起步，这类似于以前一些初创公司通过微软 Windows 来构建服务以及最近通过 iTunes 来发布应用。



6.Nvidia 的核心产品：Pascal 家族

英伟达的传统强项是桌面和移动终端的 GPU，但是坚定地向着人工智能大步迈进的英伟达显然已经不满足于仅仅在单一领域做提高 GPU 性能的事了。相比于传统的计算密集型 GPU 产品来说，英伟达努力的方向是使得 GPU 芯片不仅仅只针对训练算法这一项起到作用，更是能处理人工智能服务的推理工作负载，从而加速整个人工智能的开发流程。目前该公司的核心产品包括基于 Pascal 架构的 TeslaP4 与 Tesla P40 深度学习芯片，这两款芯片均已于 2016 年第四季度开始投入量产。

Tesla P4 为资料中心带来最高的能源效率

其小尺寸及最小 50 瓦特的低功率设计可安装于任何服务器内，让生产作业负载推论的能源效率达 CPU 的 40 倍。在进行视频推论作业负载时，单一服务器裡安装单颗 Tesla P4 即可取代 13 台仅采用 CPU 的服务器，而包含服务器及用电量的总持有成本则能节省达 8 倍。

Tesla P40 为深度学习作业负载带来最大的处理量

一台搭载 8 颗 Tesla P40 加速器的服务器拥有每秒 47 兆次运算的推论性能及 INT8 指令，可取代 140 台以上的 CPU 服务器的性能。若以每台 CPU 服务器约 5,000 美元计算，可节省 65 万美元以上的服务器采购成本。

基于上述两种人工智能芯片，英伟达为资料中心提供唯一的端对端深度学习平台，并能够将训练时间从数天大幅缩短至数小时，从而实现资料的立即解析与服务的及时回应。

7.Nvidia 的应用布局：自动驾驶

不仅仅是底层架构，英伟达在应用层面上也有非常明确的布局，其中最看重也最有领先优势的就是自动驾驶。早在 2014 年 1 月，英伟达就发布了为移动平台设计的第一代 Tegra 系列处理器，适用于智能手机、平板电脑和自动驾驶汽车，四个月后，DRIVE PX 自动驾驶计算平台发布，可实现包括高速公路自动驾驶与高清制图在内的自动巡航功能。同年 10 月，搭载了 Tegra K1 处理器并应用了 DRIVEPX 计算平台的特斯拉新款 Model S 开始量产，英伟达成为第一个享受到自动驾驶红利的厂商。

2016 年英伟达在自动驾驶领域并没有什么重大突破，基本只是从技术升级及厂商合作两个方面入手，除了特斯拉这个老朋友外，百度、沃尔沃也跟英伟达达成了合作，他们都将生产

搭载 DRIVE PX 2 的智能驾驶汽车。恰逢此时，AI 概念变得更加火热，智能驾驶也逐渐成熟，这些客观因素让英伟达收割了更多的红利，也让公司站在了聚光灯之下。

从整个自动驾驶行业来看，Google、苹果、微软等科技公司都在建立自己的汽车生态体系，不过智能汽车对于他们来说都不是核心业务，更为重要的是，他们并没有真正进入汽车供应链体系。与之相反，英伟达的 Drive PX 系列自动驾驶解决方案，已经进入了汽车的上游供应链中，并创造了利润，这也意味着英伟达将在汽车芯片市场与英特尔、高通、恩智浦、瑞萨电子等做 CPU 的公司正面碰撞，自动驾驶的风口让英伟达在汽车市场从“边缘人”变成了挑战者。

随着特斯拉 Model S 等备受瞩目的车型更加智能化与多媒体化，英伟达有了弯道超车的机会，并有望在汽车产业的上游供应链占据更有优势的地位。最新款的 Tegra 系列处理器功耗只有 10 瓦，几乎与同等级的 FPGA 产品功耗持平甚至更低，这对于车载移动芯片来说是巨大的优势。

但同样的，单移动处理器的架构和极低的功耗必然无法支撑起超大规模的运算，目前英伟达计算平台的功能定位仅聚焦于高速公路上的自动巡航，而 CPU 的应用可以拓展至车机娱乐信息系统层面。未来自动驾驶的发展方向必然是整车的控制中心，从目前英伟达基于 Tesla 架构的主流芯片来看，低功耗、极速运算与逻辑控制是可以同时实现的，英伟达公司在自动驾驶领域的优势非常明显。

8.Nvidia 的产业优势：完善的生态系统

与其它芯片公司相比，带有 CUDA 的重点软件生态系统是英伟达占领人工智能市场的关键促成因素。从 2006 年开始，英伟达发布了一个名叫 CUDA 的编程工具包，该工具包让开发者可以轻松编程屏幕上的每一个像素。在 CUDA 发布之前，给 GPU 编程对程序员来说是一件极其痛苦的事，因为这涉及到编写大量低层面的机器码以实现渲染每一个不同像素的目标，而这样的微型计算操作通常有上万个。CUDA 在经过了英伟达的多年开发之后，成功将 Java 或 C++这样的高级语言开放给了 GPU 编程，从而让 GPU 编程变得更加轻松简单，研究者也可以更快更便宜地开发他们的深度学习模型。

四、未来市场：半定制芯片 FPGA

技术世界正在迈向一个全新的轨道，我们对于人工智能的想象已经不再局限于图片识别与声音处理，机器，将在更多领域完成新的探索。不同领域对计算的需求是差异的，这就要求深度学习的训练愈发专业化与区别化。芯片的发展趋势必将是在每一个细分领域都可以更加符合我们的专业需求，但是考虑到硬件产品一旦成型便不可再更改这个特点，我们不禁开始想，是不是可以生产一种芯片，让它硬件可编程。

也就是说，这一刻我们需要一个更适合图像处理的硬件系统，下一刻我们需要一个更适合科学计算的硬件系统，但是我们又不希望焊两块板子，我们希望一块板子便可以实现针对每一个应用领域的不同需求。这块板子便是半定制芯片 FPGA，便是未来人工智能硬件市场的发展方向。

1.FPGA 是什么？

场效可编程逻辑闸阵列 FPGA 运用硬件语言描述电路，根据所需要的逻辑功能对电路进行快速烧录。一个出厂后的成品 FPGA 的逻辑块和连接可以按照设计者的需要而改变，这就好像一个电路试验板被放在了一个芯片里，所以 FPGA 可以完成所需要的逻辑功能。

FPGA 和 GPU 内都有大量的计算单元，因此它们的计算能力都很强。在进行神经网络运算的时候，两者的速度会比 CPU 快很多。但是 GPU 由于架构固定，硬件原生支持的指令也就固定了，而 FPGA 则是可编程的。其可编程性是关键，因为它让软件与终端应用公司能够提供与其竞争对手不同的解决方案，并且能够灵活地针对自己所用的算法修改电路。

2.FPGA 和 GPU 的性能差异

同样是擅长并行计算的 FPGA 和 GPU，谁能够占领人工智能的高地，并不在于谁的应用更广泛，而是取决于谁的性能更好。**在服务器端，有三个指标可供对比：峰值性能、平均性能与功耗能效比。**当然，这三个指标是相互影响的，不过还是可以分开说。

峰值性能：GPU 远远高于 FPGA

GPU 上面成千上万个核心同时跑在 GHz 的频率上是非常壮观的，最新的 GPU 峰值性能甚至可以达到 10TFlops 以上。GPU 的架构经过仔细设计，在电路实现上是基于标准单元库而在关键路径上可以用手工定制电路 ,甚至在必要的情形下可以让半导体 fab 依据设计需求微调工艺制程，因此可以让许多 core 同时跑在非常高的频率上。

相对而言，FPGA 首先设计资源受到很大的限制，例如 GPU 如果想多加几个核心只要增加芯片面积就行，但 FPGA 一旦型号选定了逻辑资源上限就确定了。而且，FPGA 里面的逻辑单元是基于 SRAM 查找表，其性能会比 GPU 里面的标准逻辑单元差很多。最后，FPGA 的布线资源也受限制 ,因为有些线必须要绕很远 ,不像 GPU 这样走 ASIC flow 可以随意布线，这也会限制性能。

平均性能：GPU 逊于 FPGA

FPGA 可以根据特定的应用去编程硬件 ,例如如果应用里面的加法运算非常多就可以把大量的逻辑资源去实现加法器，而 GPU 一旦设计完就不能改动了，所以不能根据应用去调整硬件资源。

目前机器学习大多使用 SIMD 架构，即只需一条指令可以平行处理大量数据，因此用 GPU 很适合。但是有些应用是 MISD，即单一数据需要用许多条指令平行处理，这种情况下用 FPGA 做一个 MISD 的架构就会比 GPU 有优势。

所以 ,对于平均性能 ,看的就是 FPGA 加速器架构上的优势是否能弥补运行速度上的劣势。如果 FPGA 上的架构优化可以带来相比 GPU 架构两到三个数量级的优势 ,那么 FPGA 在平均性能上会好于 GPU。

功耗能效比：

功耗方面，虽然 GPU 的功耗远大于 FPGA 的功耗，但是如果要比较功耗应该比较在执行效率相同时需要的功耗。如果 FPGA 的架构优化能做到很好以致于一块 FPGA 的平均性能能够接近一块 GPU，那么 FPGA 方案的总功耗远小于 GPU，散热问题可以大大减轻。反之，如果需要二十块FPGA才能实现一块GPU的平均性能 ,那么FPGA在功耗方面并没有优势。

能效比的比较也是类似，能效指的是完成程序执行消耗的能量，而能量消耗等于功耗乘以程序执行的时间。虽然 GPU 的功耗远大于 FPGA 的功耗，但是如果 FPGA 执行相同程序需要的时间比 GPU 长几十倍，那 FPGA 在能效比上就没有优势了；反之如果 FPGA 上实现的硬件架构优化得很适合特定的机器学习应用，执行算法所需的时间仅仅是 GPU 的几倍或甚至于接近 GPU，那么 FPGA 的能效比就会比 GPU 强。

3.FPGA 市场前景

随着科技的进展，制造业走向更高度的自动化与智能化，对工业控制技术等领域不断产生新的需求，在未来的工业制造领域，FPGA 将有更大的发展空间。目前来看，有两个领域的应用前景十分巨大：

工业互联网领域

作为未来制造业发展的方向，工业大数据、云计算平台、MES 系统等都是支持工业智能化的重要平台，它们需要完成大数据量的复杂处理，FPGA 在其中可以发挥重要作用。

工业机器人设备领域

在多轴向运作的精密控制、实时同步的连接以及设备多功能整合等方面，兼具弹性和整合性的 FPGA，更能展现设计优势。如汽车 ADAS 需要对实时高清图像进行及时的分析识别与处理；在人工智能方面，深度学习神经网络也需要进行大量并行运算。

4.FPGA 现有市场

FPGA 市场前景诱人，但是门槛之高在芯片行业里无出其右。全球有 60 多家公司先后斥资数十亿美元，前赴后继地尝试登顶 FPGA 高地，其中不乏英特尔、IBM、德州仪器、摩托罗拉、飞利浦、东芝、三星这样的行业巨鳄，但是最终登顶成功的只有位于美国硅谷的两家公司：Xilinx 与 Altera。这两家公司共占有近 90%的市场份额，专利达到 6000 余项之多，如此之多的技术专利构成的技术壁垒当然高不可攀。

2015 年 6 月，英特尔用史无前例的 167 亿美元巨款收购了 Altera，当时业内对于英特尔此举的解读主要集中在服务器市场、物联网市场的布局上，英特尔自己对收购的解释也没有明确提到机器学习。但现在看来，或许这笔收购在人工智能领域同样具有相当大的潜力。

5.FPGA 行业的开拓者

英特尔能不能通过 FPGA 切入 AI 硬件市场？要讲清楚这个问题，我们必须要把视角从人工智能身上拉远，看看英特尔的整体战略布局。最近几年，英特尔的核心盈利业务 CPU 同时遭到了三个因素的狙击：PC 市场增长放缓、进军移动市场的尝试失败以及摩尔定律逐渐逼近极限。单纯的卖 CPU 固然也能赚到钱，但只有研发更高端的芯片，形成自己领导者的形象，才能赚更多的钱，支撑公司的发展。

上述三个因素的同时出现，已经让英特尔发现，如果自己仍然只是安心的守着自己的 CPU 业务，很快就会面临巨大的危机，事实上在过去的一年里，利润下降、裁员的新闻也一直围绕在英特尔的身边，挥之不去。

因而英特尔十分渴望不要错过下一个深度学习的潮流，不过它缺乏自己最先进的人工智能研究，所以在过去的两年中疯狂地收购。2015 年，英特尔用史无前例的 167 亿美元拍下了 FPGA 制造商 Altera，2016 年又相继兼并了人工智能芯片初创公司 Nervana 与 Movidius。目前的英特尔正在试图将他们整合在一起。

6.Intel 的产品布局

英特尔斥巨资收购 Altera 不是来为 FPGA 技术发展做贡献的，相反，它要让 FPGA 技术为英特尔的发展做贡献。表现在技术路线图上，那就是从现在分立的 CPU 芯片+分立的 FPGA 加速芯片，过渡到同一封装内的 CPU 晶片+FPGA 晶片，到最终的集成 CPU+FPGA 芯片。预计这几种产品形式将会长期共存，因为分立器件虽然性能稍差，但灵活性更高。

如果简单的将英特尔对于人工智能的产品布局，可以分以下几层：

- Xeon Phi+ Nervana：用于云端最顶层的高性能计算。

- Xeon+FPGA：用于云端中间层/前端设备的低功耗性能计算。

英特尔下一代的 FPGA 和 SoC FPGA 将支持 Intel 架构集成，大致如下：代号为 Harrisville 的产品采用 Intel 22nm 工艺技术，用于工业 IoT、汽车和小区射频等领域；代号为 Falcon Messa 的中端产品采用 Intel 10nm 工艺技术，用于 4G/5G 无线通信、UHD/8K 广播视频、工业 IoT 和汽车等领域；代号为 Falcon Mesa 的高端产品采用 Intel 10nm 工艺技术，用于云和加速、太比特系统和高速信号处理等领域。

- Core（GT）：用于消费级前端设备的性能计算、图形加速。

- Euclid：提供给开发者/创客的开发板，集成 Atom 低功耗处理器、RealSense 摄像头模块、接口，可用做无人机、小型机器人的核心开发部件。

- Curie：提供给开发者/创客的模块，其内置 Quark SE 系统芯片、蓝牙低功耗无线电、以及加速计、陀螺仪等传感器，可用做低功耗可穿戴设备的核心部件。

从产品线来看，包含了 CPU 与 FPGA 的异构计算处理器将是 Intel 盈利的重点。预计到 2020 年 Intel 将有 1/3 的云数据中心节点采用 FPGA 技术，CPU+FPGA 拥有更高的单位功耗性

能、更低时延和更快加速性能，在大数据和云计算领域有望冲击 CPU+GPU 的主导地位，而 Intel 的至强处理器 Xeon +FPGA 也将在 2017 年下半年量产。

7.Intel 的痛点：生态不完善

FPGA 对 GPU 的潜力在于其计算速度与 GPU 不相上下，却在成本和功耗上对 GPU 有着显著优势。当然，劣势也有，但是 FPGA 的潜力是非常明显的。作为一个想要推向市场的商品来说，FPGA 最需要克服，也是最容易克服的问题是普及程度。

大部分 PC 都配有或高端或低端的独立 GPU，对于个人进行的中小规模神经网络开发和训练来说，其实它们的性能已经基本足够。而 FPGA 却不是在电脑里能找得到的东西，而多见于各种冰箱、电视等电器设备及实验室中，因此想要搞到一块能用来开发深度学习的 FPGA 其实还挺麻烦的。不仅如此，**FPGA 的不普及还体现在以下三个方面：**

OpenCL 编程平台应用不广泛

即使 GPU 有着种种不足，它也不是能够轻易被取代的。从深度学习应用的开发工具角度，具备 CUDA 支持的 GPU 为用户学习 Caffe、Theano 等研究工具提供了很好的入门平台。自 2006 年推出 CUDA 以来，已有超过 5 亿的笔记本电脑、工作站、计算集群和超级计算机安装了支持 CUDA 的 GPU。

如果 FPGA 想要攻占深度学习的市场，那么产业链下游的编程平台必不可少。目前较为流行的异构硬件编程的替代性工具是 OpenCL。不同于 CUDA 单一供应商的做法，OpenCL 对开发者开源、免费，这是一大重要竞争力。但目前来看，其获得的支持相较 CUDA 还略逊一筹。

实现硬件编程困难

除了软件编程的不普及之外，吸引偏好上层编程语言的研究人员和应用科学家来开发 FPGA 尤为艰难。虽然能流利使用一种软件语言常常意味着可以轻松地学习另一种软件语言，但对于硬件语言翻译技能来说却非如此。针对 FPGA 最常用的语言是 Verilog 和 VHDL，两者均为硬件描述语言（HDL）。这些语言和传统的软件语言之间的主要区别是，HDL 只是单纯描述硬件，而例如 C 语言等软件语言则描述顺序指令，并无需了解硬件层面的执行细节。

有效地描述硬件需要对数字化设计和电路的专业知识，尽管一些下层的实现决定可以留给自动合成工具去实现，但往往无法达到高效的设计。因此，研究人员和应用科学家倾向于选择软件设计，因其已经非常成熟，拥有大量抽象和便利的分类来提高程序员的效率。

部署环节需要定制复杂套件

FPGA 需要有一个完善的复杂生态系统才能保证其使用，不只体现在软件与硬件编程平台上，更体现在部署环节中。FPGA 在安装过程中需要针对不同的 IP 核定制一系列复杂的工具套

件，相比之下，GPU 通过 PCI-e 接口可以直接部署在服务器中，方便而快速。因此，嵌入式 FPGA 概念虽好，想要发展起来仍将面临十分严峻的挑战。

8.Intel 的优势

目前在深度学习市场 FPGA 尚未成气候，谷歌这样的超级大厂又喜欢自己研发专用芯片，因此可以说对于深度学习芯片来说，个人开发者及中小型企业内还有相当大的市场。这个市场目前几乎只有英伟达一家独大，英特尔想要强势进入未必没有机会。**而相比于英伟达来说，英特尔有两个明显的优势：**

更熟悉 CPU

尽管目前的人工智能市场几乎只有英伟达一家独大，但英伟达的芯片也不是能够自己完成深度学习训练的。或者说，英伟达的 GPU 芯片还不足以取代那些英特尔的 CPU，大多数环境下它们暂时只能加速这些处理器。所以，GPGPU 暂时只是概念上的，GPU 还不足以在大多数复杂运算环境下代替 CPU，而随着人工智能技术的进步，对硬件的逻辑运算能力只会更高不会降低，所以搭载强大 CPU 核心的多核异构处理器才是更长期的发展方向。而论对 CPU 的熟悉，没有一家芯片厂商能过胜过英特尔，英特尔是最有可能让搭载了 FPGA 与 CPU 的异构处理器真正实现多核心相辅相成的芯片公司。

曾涉足云计算

算法的训练应该是贯穿整个应用过程的，这样可以随时为消费者提供最好体验的服务。但是如果要将所有算法都集中于本地训练，不仅会面临计算瓶颈的问题，也容易面临从单个用户处收集到的数据量太少的尴尬。我们暂时不考虑很久以后可能出现的基于小样本的无监督学习的 AI，毕竟那其实已经跟人差不多了，在目前 AI 的发展状况下，将所有数据集中于云端进行计算显然是更理性且有效的做法。这就对通信提出了极高的要求，而英特尔恰巧在这个领域有着相当多的积累。虽然英特尔的通信部门连年亏损，但在现在的形势下，它却意外地有了新的价值与潜力。