浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD

专题

行业公司研究—半导体行业—

证券研究报告

半导体行业

报告日期：2016 年 5 月 27 日

风起于青萍之末——GPU,FPGA，ASIC

——人工智能芯片行业深度报告

：杨云　　执业证书编号：S0860510120006
：021-80108643
：chenjunjie@stocke.com.cn

报告导读

从 3 月份智能机器人 AlphaGo 战胜 李世石，到近期谷歌的最新用于人工智能深度学习的芯片 TPU 曝光，一个千亿级的市场应用逐渐从水底浮向了水面。我们将深度剖析，在人工智能领域，有可能爆发的芯片——GPU、FPGA、ASIC 及相关的市场和公司。

投资要点

❑ **人工智能——风起于青萍之末**

人工智能市场将保持高速增长，根据艾瑞咨询的数据，2020 年全球人工智能市场规模约 1190 亿人民币。而未来 10 年，人工智能将会是一个 2000 亿美元的市场。空间非常巨大。其中在硬件市场方面，将会有 30%的市场份额。对人工智能的实现来说，算法是核心，计算、数据是基础。人工智能之所以在当下能得到长足进展，高性能的计算能力（CPU，GPU，FPGA，ASIC）的突破至关重要。

❑ **GPU——厚积薄发正当时**

我们认为人工智能时代的 GPU 已经不再是传统意义上的图形处理器，而更多的应该赋予专用处理器的头衔，具备强大的并行计算能力。因为 GPU 的特点特别适合于大规模并行运算，GPU 在“深度学习”领域发挥着巨大的作用，因为 GPU 可以平行处理大量琐碎信息。深度学习所依赖的是神经系统网络——与人类大脑神经高度相似的网络——而这种网络出现的目的，就是要在高速的状态下分析海量的数据。

❑ **FPGA——“万能芯片”在人工智能时代复苏**

FPGA 之所以能有潜力成为人工智能深度学习方面的计算工具，主要原因就在于其本身特性：可编程专用性，高性能，低功耗。比 CPU 和 GPU，FPGA 凭借比特级细粒度定制的结构、流水线并行计算的能力和高效的能耗，在深度学习应用中展现出独特的优势，在大规模服务器部署或资源受限的嵌入式应用方面有巨大潜力。此外，FPGA 架构灵活，使得研究者能够在诸如 GPU 的固定架构之外进行模型优化探究。Intel 收购 Altera,目的也是看中 FPGA 在深度学习体想出的性能优势。

❑ **ASIC——后起之秀，不可估量**

ASIC 将性能和功耗完美结合。ASIC 著名应用之一：比特币挖矿。比特币挖矿和人工智能深度学习有类似之处，都是依赖于底层的芯片进行大规模的并行计算。而 ASIC 在比特币挖矿领域，展现出了得天独厚的优势。从“比特币挖矿机 ASIC 发展”推导“ASIC 在人工智能领域大有可为”。而谷歌最新推出的人工智能专用芯片其实也是一款 ASIC。

行业评级

半导体	看好

相关报告

1《一周观点：人工智能硬件（首篇）：从 GPU 到 TPU》2016.05.23

2《一周观点：寻找半导体行业的投资机会》2016.05.16

3《一周观点：从 SOX 和 B/B 值看半导体行业投资》2016.05.08

4《一周观点：未来可能颠覆 CPU 的几类芯片》2016.05.02

5《一周观点：聊聊毫米波雷达芯片》2016.04.24

报告撰写人：杨云
数据支持人：陈俊杰

正文目录

图表目录

1. 人工智能——风起于青萍之末

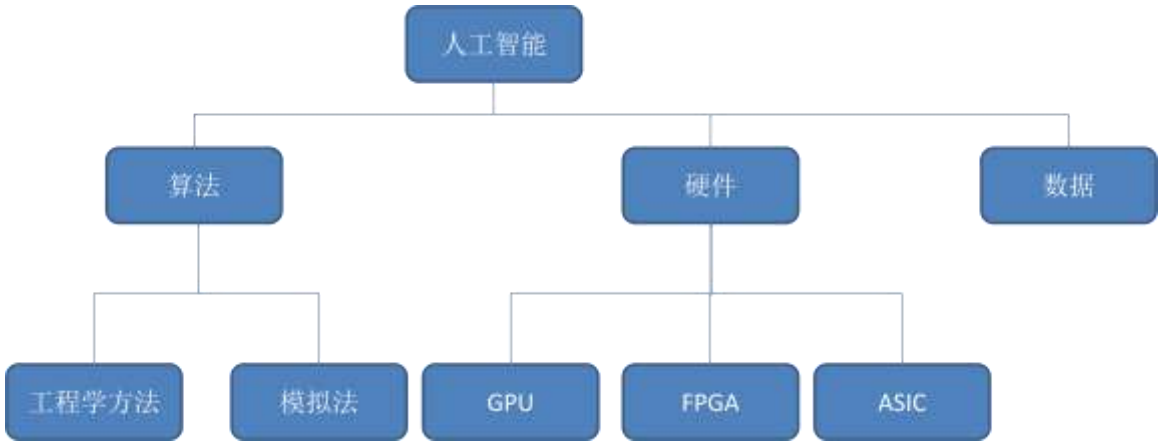
从 3 月份智能机器人 AlphaGo 战胜 李世石，到近期谷歌的最新用于人工智能深度学习的芯片 TPU 曝光，一个千亿级的市场应用逐渐从水底浮向了水面。我们将深度剖析，在人工智能领域，有可能爆发的芯片——GPU、FPGA、ASIC 及相关的市场和公司。

1.1. 人工智能——下一个千亿级市场

人工智能会成为未来的趋势吗？答案是会。人工智能，简单地说，就是用机器去实现目前必须借助人类智慧才能实现的任务。**人工智能包括三个要素：算法，计算和数据**。

对人工智能的实现来说，算法是核心，计算、数据是基础。在算法上来说，主要分为工程学法和模拟法。工程学方法是采用传统的编程技术，利用大量数据处理经验改进提升算法性能；模拟法则是模仿人类或其他生物所用的方法或者技能，提升算法性能，例如遗传算法和神经网络。而在计算能力来说，目前主要是使用 GPU 并行计算神经网络，同时，FPGA 和 ASIC 也将是未来异军突起的力量。

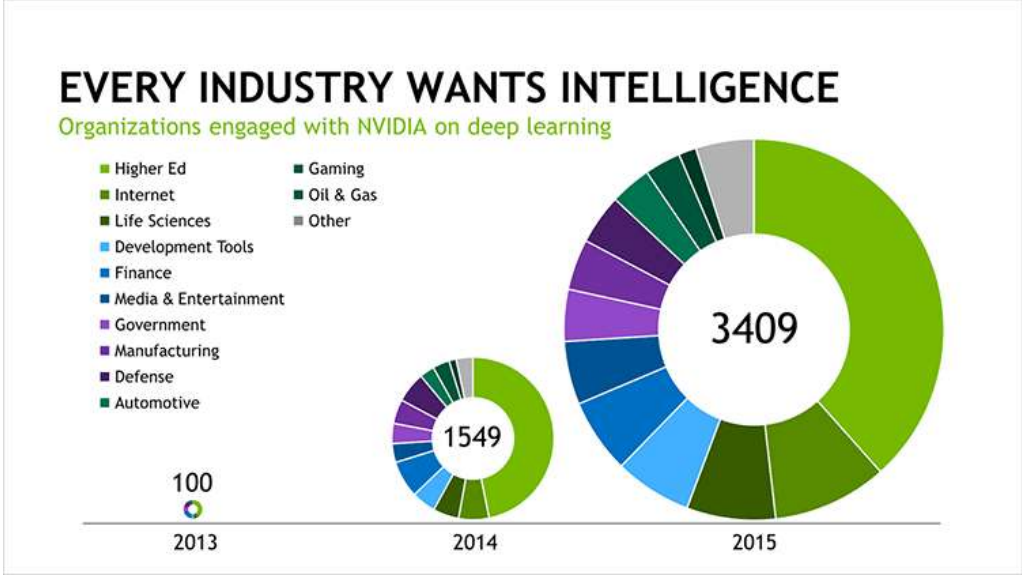
图 1：人工智能实现三要素



资料来源：浙商证券研究所

随着百度，Google，Facebook ，Microsoft 等企业开始切入人工智能，人工智能可应用的领域非常广泛。2013 年 100 多家组织开始研发深度学习与人工智能，到 2015 年，短短 2 年间，研发机构已经迅速激增到 3409 家。可以看到，未来人工智能的应用将呈几何级数的倍增。应用领域包括互联网，金融，娱乐，政府机关，制造业，汽车，游戏等。从产业结构来讲，人工智能生态分为基础、技术、应用三层。应用层包括人工智能+各行业（领域），技术层包括算法、模型及应用开发，基础层包括数据资源和计算能力。

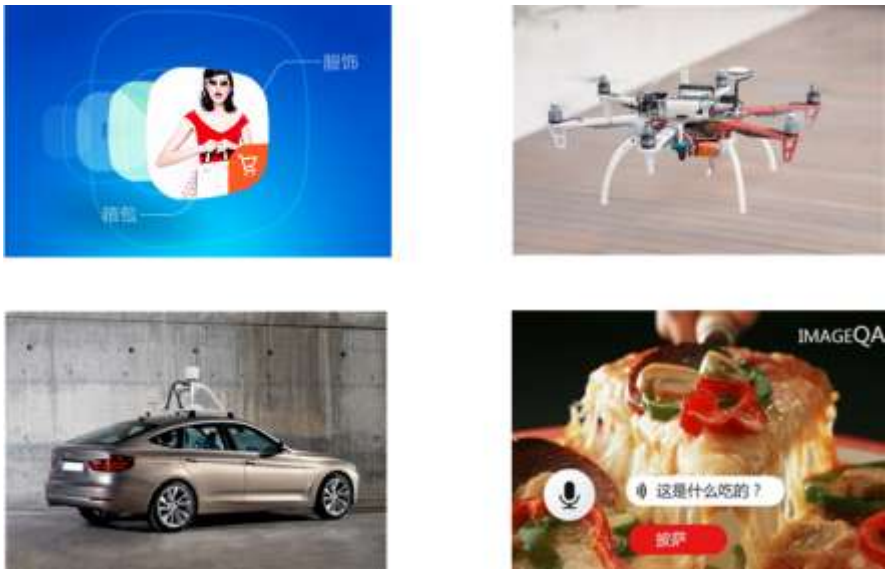
图 2：人工智能应用广泛



资料来源：NVIDIA 官网，浙商证券研究所

人工智能将在很多领域得到广泛的应用。目前重点部署的应用有：语音识别，人脸识别，无人机，机器人，无人驾驶等。

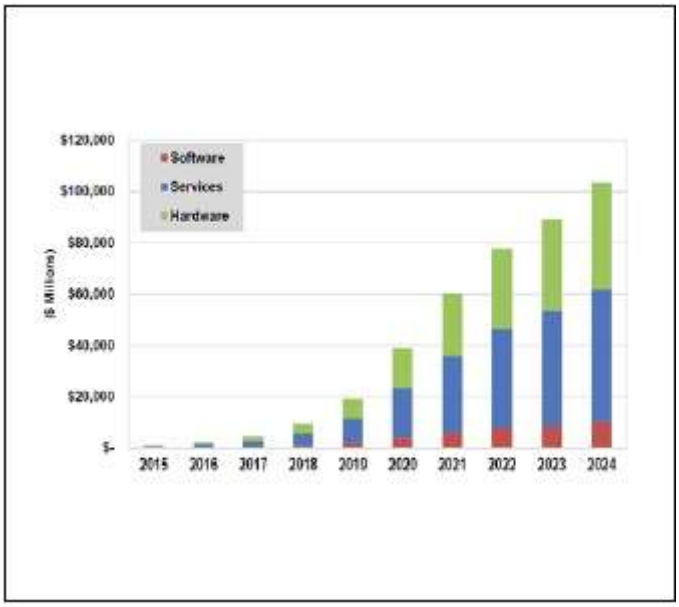
图 3：人工智能应用领域



资料来源：浙商证券研究所

人工智能市场将保持高速增长，根据艾瑞咨询的数据，2020 年全球人工智能市场规模约 1190 亿人民币。而未来 10 年，人工智能将会是一个 2000 亿美元的市场。空间非常巨大。其中在硬件市场方面，将会有 30%的市场份额。

图 4：人工智能市场规模



资料来源：浙商证券研究所

1.2. 深度学习

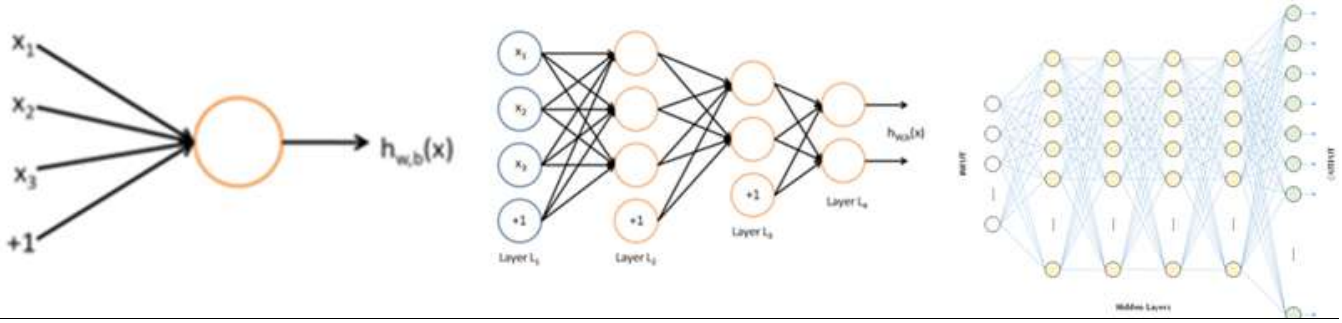
人工智能的核心是算法，深度学习是目前最主流的人工智能算法。深度学习在 1958 年就被提出，但直到最近，才真正火起来，主要原因在于：数据量的激增和计算机能力/成本。

深度学习是机器学习领域中对模式（声音、图像等等）进行建模的一种方法，它也是一种基于统计的概率模型。在对各种模式进行建模之后，便可以对各种模式进行识别了，例如待建模的模式是声音的话，那么这种识别便可以理解为语音识别。而类比来理解，如果说将机器学习算法类比为排序算法，那么深度学习算法便是众多排序算法当中的一种，这种算法在某些应用场景中，会具有一定的优势。

深度学习的学名又叫深层神经网络（Deep Neural Networks ），是从很久以前的人工神经网络（Artificial Neural Networks）模型发展而来。这种模型一般采用计算机科学中的图模型来直观的表达，而深度学习的“深度”便指的是图模型的层数以及每一层的节点数量，相对于之前的神经网络而言，有了很大程度的提升。

从单一的神经元,再到简单的神经网络,到一个用于语音识别的深层神经网络。层次间的复杂度呈几何倍数的递增。

图 5：单一神经元 VS 简单神经网络 VS 深层神经网络

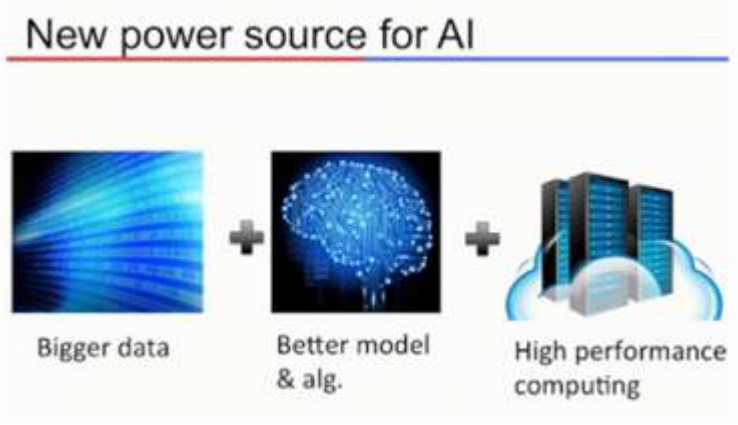


资料来源：浙商证券研究所

以图像识别为例，图像的原始输入是像素，相邻像素组成线条，多个线条组成纹理，进一步形成图案，图案构成了物体的局部，直至整个物体的样子。不难发现，可以找到原始输入和浅层特征之间的联系，再通过中层特征，一步一步获得和高层特征的联系。想要从原始输入直接跨越到高层特征，无疑是困难的。而整个识别过程，所需要的数据量和运算量是十分巨大的。2012 年，由人工智能和机器学习顶级学者 Andrew Ng 和分布式系统顶级专家 Jeff Dean，用包含 16000 个 CPU 核的并行计算平台训练超过 10 亿个神经元的深度神经网络，在语音识别和图像识别等领域取得

了突破性的进展。该系统通过分析 YouTube 上选取的视频，采用无监督的方式训练深度神经网络，可将图像自动聚类。在系统中输入“cat”后，结果在没有外界干涉的条件下，识别出了猫脸。可以看到，深度学习之所以能够在今天得到重要的突破，**原因在于：1 海量的数据训练 2 高性能的计算能力（CPU，GPU，FPGA，ASIC）**。两者缺一不可。

图 6：人工智能新驱动



资料来源：浙商证券研究所

1.3. 算力

衡量芯片计算性能的重要指标称为算力。通常而言，将每秒所执行的浮点运算次数（亦称每秒峰值速度）作为指标来衡量算力，简称为 FLOPS。现有的主流芯片运算能力达到了 TFLOPS 级别。一个 TFLOPS（teraFLOPS）等於每秒万亿（=10^12）次的浮点运算。

增加深度学习算力需要多个维度的齐头并进的提升：1 系统并行程度 2 时钟的速度 3 内存的大小（包括 register,cache,memory）；4 内存带宽（memory bandwidth）5 计算芯片同 CPU 之间的带宽 6 还有各种微妙的硬件里的算法改进。

我们这篇报告将主要关注人工智能的芯片领域，着重讨论 GPU,FPGA,ASIC 等几种类型的芯片在人工智能领域的应用和未来的发展。

2. GPU——厚积薄发正当时

2.1. GPU 简介

GPU，又称显示核心、视觉处理器、显示芯片，是一种专门在个人电脑、工作站、游戏机和一些移动设备（如平板电脑、智能手机等）上图像运算工作的微处理器，与 CPU 类似，只不过 GPU 是专为执行复杂的数学和几何计算而设计的，这些计算是图形渲染所必需的。随着人工智能的发展，如今的 GPU 已经不再局限于 3D 图形处理了，GPU 通用计算技术发展已经引起业界不少的关注，事实也证明在浮点运算、并行计算等部分计算方面，GPU 可以提供数十倍乃至于上百倍于 CPU 的性能。

GPU 的特点是有大量的核（多达几千个核）和大量的高速内存，最初被设计用于游戏，计算机图像处理等。GPU 主要擅长做类似图像处理的并行计算，所谓的“粗粒度并行（coarse-grain parallelism）”。这个对于图像处理很适用，因为像素与像素之间相对独立，GPU 提供大量的核，可以同时对很多像素进行并行处理。但这并不能带来延迟的提升（而仅仅是处理吞吐量的提升）。比如，当一个消息到达时，虽然 GPU 有很多的核，但只能有其中一个核被用来处理当前这个消息，而且 GPU 核通常被设计为支持与图像处理相关的运算，不如 CPU 通用。GPU 主要适用于在数据层呈现很高的并行特性（data-parallelism）的应用，比如 GPU 比较适合用于类似蒙特卡罗模拟这样的并行运算。

图 7：GPU VS CPU

资料来源：浙商证券研究所

CPU 和 GPU 本身架构方式和运算目的不同导致了 CPU 和 GPU 之间的不同，主要不同点列举如下。

表 1：CPU VS GPU

	CPU	GPU
架构区别	70%晶体管用来构建 Cache 还有一部分控制单元，负责逻辑算数的部分并不多	整个就是一个庞大的计算阵列（包括 alu 和 shader 填充
	非常依赖 Cache	不依赖 Cache
	逻辑核心复杂	逻辑核心简单
计算目的	适合串行	适合大规模并行
	运算复杂度高	运算复杂度低

资料来源：浙商证券研究所

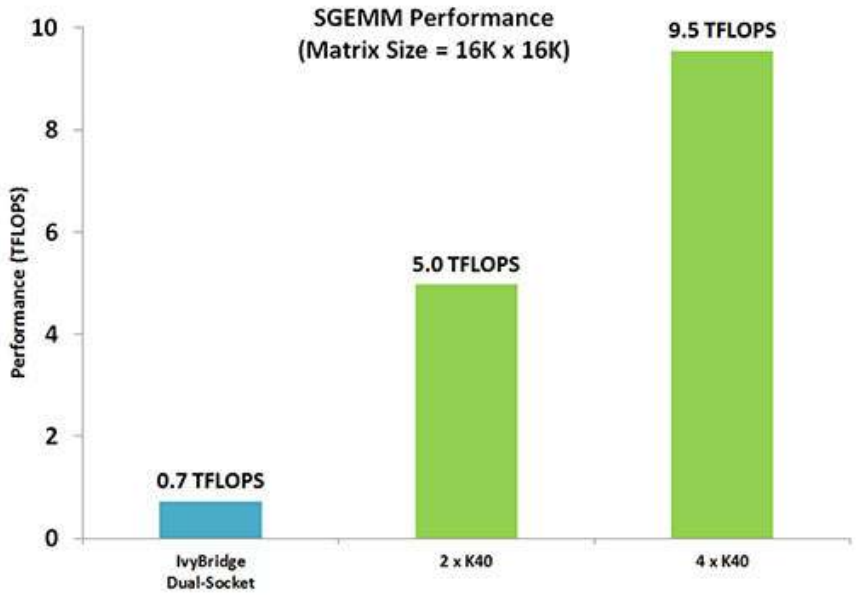
正是因为 GPU 的特点特别适合于大规模并行运算，GPU 在“深度学习”领域发挥着巨大的作用，因为 GPU 可以平行处理大量琐碎信息。深度学习所依赖的是神经系统网络——与人类大脑神经高度相似的网络——而这种网络出现的目的，就是要在高速的状态下分析海量的数据。例如，如果你想要教会这种网络如何识别出猫的模样，你就要给它提供无数多的猫的图片。而这种工作，正是 GPU 芯片所擅长的事情。而且相比于 CPU，GPU 的另一大优势，就是它对能源的需求远远低于 CPU。GPU 擅长的是海量数据的快速处理。

工业与学术界的数据科学家已将 GPU 用于机器学习以便在各种应用上实现开创性的改进，这些应用包括图像分类、视频分析、语音识别以及自然语言处理等等。 尤其是深度学习，人们在这一领域中一直进行大力投资和研究。深度学习是利用复杂的多级「深度」神经网络来打造一些系统，这些系统能够从海量的未标记训练数据中进行特征检测。

虽然机器学习已经有数十年的历史，但是两个较为新近的趋势促进了机器学习的广泛应用：**海量训练数据的出现以及 GPU 计算所提供的强大而高效的并行计算。** 人们利用 GPU 来训练这些深度神经网络，所使用的训练集大得多，所耗费的时间大幅缩短，占用的数据中心基础设施也少得多。 GPU 还被用于运行这些机器学习训练模型，以便在云端进行分类和预测，从而在耗费功率更低、占用基础设施更少的情况下能够支持远比从前更大的数据量和吞吐量。

将 GPU 加速器用于机器学习的早期用户包括诸多规模的网络和社交媒体公司，另外还有数据科学和机器学习领域中一流的研究机构。 与单纯使用 CPU 的做法相比，GPU 具有数以千计的计算核心、可实现 10-100 倍应用吞吐量，因此 GPU 已经成为数据科学家处理大数据的处理器。

图 8：GPU 性能展示



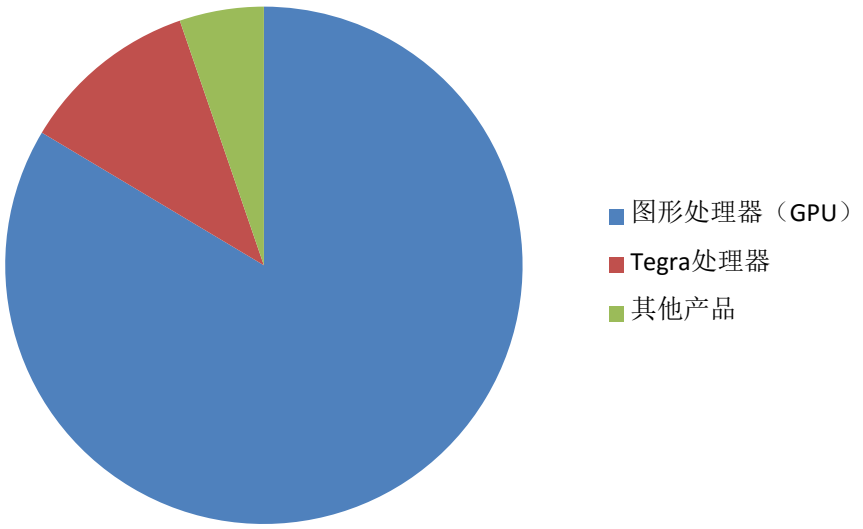
资料来源：浙商证券研究所

综上而言，我们认为人工智能时代的 GPU 已经不再是传统意义上的图形处理器，而更多的应该赋予专用处理器的头衔，具备强大的并行计算能力。

2.2. 王者归来的 NVIDIA

NVIDIA 是一家以设计 GPU 芯片为主业的半导体公司。其主要产品包括游戏显卡 GeForce GPU，工作站 Quadro，可用于深度学习计算的 Tesla GPU，为移动以及汽车处理设计 Tegra GPU。NVIDIA 的产品在应用领域来划分，主要包括图形处理器（GPU），Tegra 处理器（用于车载），以及其他。各块业务所占比重如图所示。

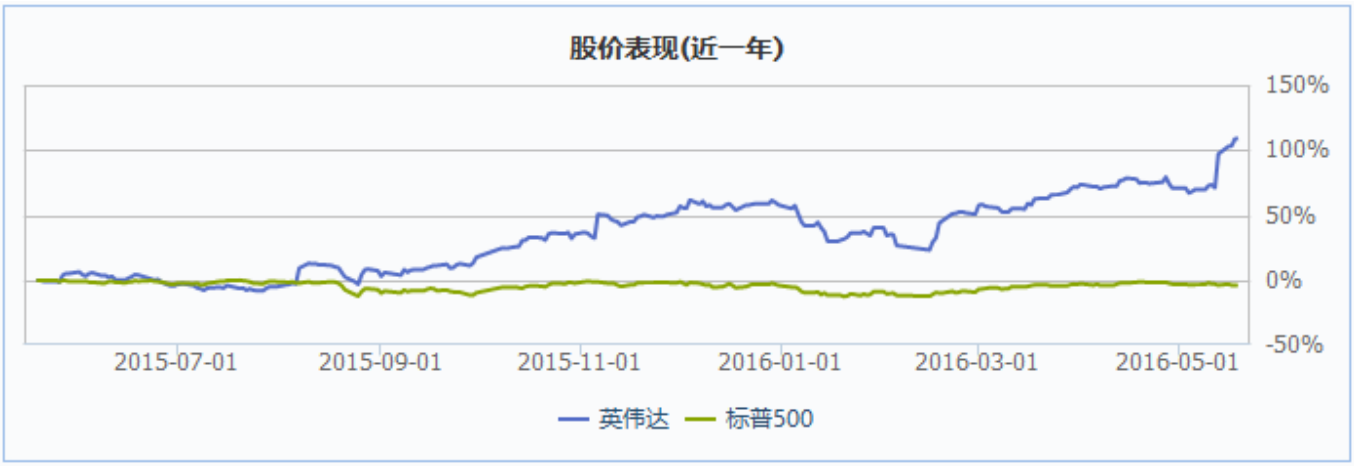
图 9：NVIDIA2015 主营构成



资料来源：NVIDIA 官网，浙商证券研究所

NVIDIA 在最近 12 日发布的财报显示，2016 年第一财报季度内，公司整体利润激增 46%，至 1.96 亿美元，营收同比增长 13%至 13.05 亿美元。财报公布后，NVIDIA 股价一度大涨 7.7%，盘中最高触及 38.81 美元。

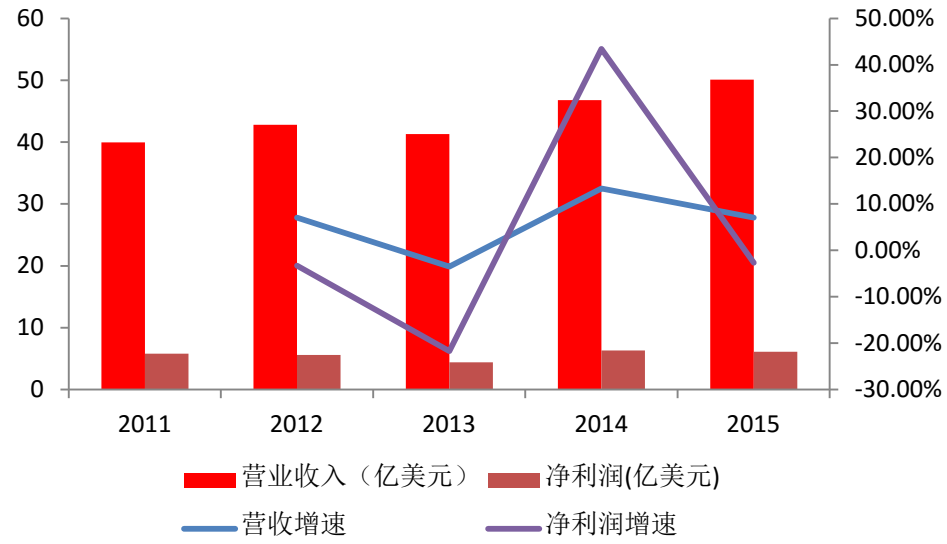
图 10：NVIDIA 股价表现强势



资料来源：wind,浙商证券研究所

NVIDIA2011-2015 年的营收和净利润如图所示。从 2013 年起，业绩明显呈现出上升的势头。这个时间点其实也契合人工智能在深度学习领域开始使用 GPU 来进行大规模并行计算。2016 年 NVIDIA 的一季报更是呈现出爆发的迹象，整体利润激增 46%，**我们认为财报的数据是最有力的证明，GPU 正在受益于人工智能深度学习的需求，而广泛地得到应用。**

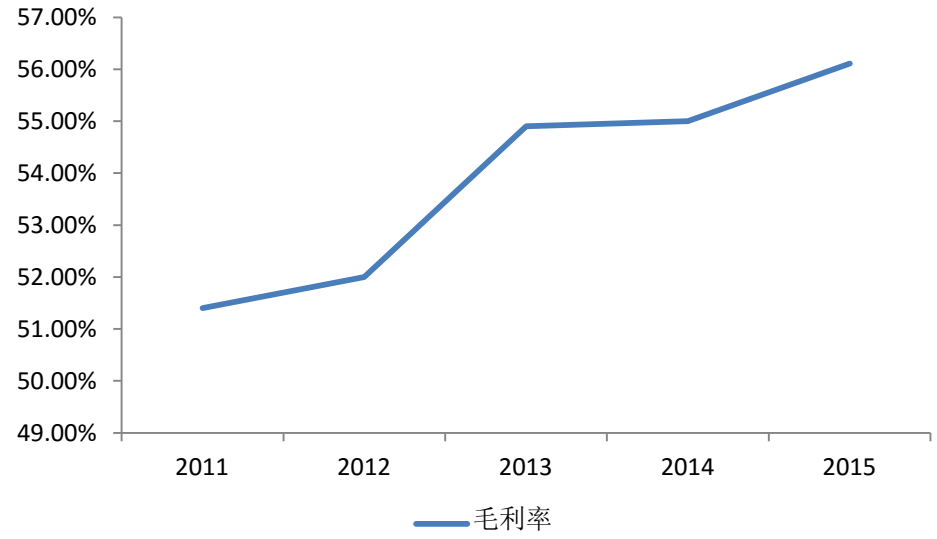
图 21：NVIDIA2011-2015 年营收 VS 净利润



资料来源：浙商证券研究所

NVIDIA 的产品毛利率从 2011 年开始，保持了连续上升的势头。NVIDIA 的核心产品是 GPU,从毛利率水平不断提升来看，NVIDIA 的 GPU 产品始终保持了核心竞争力和更新换代的能力。产品的结构也在不断优化，从独立专显到服务器再到大规模并行计算，随着应用的不断升级，产品的结构也越来越优化。从毛利率水平看，NVIDIA 的产品保持了不断更新和竞争力。随着并行计算在深度学习中的广泛应用，NVIDIA 的产品毛利率还将进一步提升。

图 32：NVIDIA2011-2015 毛利率



资料来源：浙商证券研究所

在高性能计算机、深度学习、人工智能等领域，NVIDIA 的 Tesla 芯片有十分关键的作用。NVIDIA 的 CUBA 技术，大幅度提高了纯 CPU 构成的超级计算机的性能。人工智能和深度学习需要大量的浮点计算，在高性能计算领域，GPU 需求在不断增强。目前 NVIDIA 的高性能显卡已经占有 84%的市场份额。亚马逊的 AWS，Facebook，Google 等世界一级数据中心都需要用 NVIDIA 的 Tesla 芯片，随着云计算和人工智能的不断发展，我们认为 NVIDIA 的高性能 GPU 也能在未来 5 年保持 20%以上的增长速度。

2.3. GPU 国内行业现状及公司

国内在 GPU 芯片设计方面，还处于起步阶段，与国际主流产品尚有一定的差距。不过星星之火，可以燎原。有一些企业，逐渐开始拥有自主研发的能力，比如国内企业景嘉微。景嘉微拥有国内首款自主研发的 GPU 芯片 JM5400，专用于公司的图形显控领域。JM5400 为代表的图形芯片打破外国芯片在我国军用 GPU 领域的垄断，率先实现军用 GPU 国产化。

公司的 GPU JM5400 主要替代 AMD 的 GPU M9,两者在性能上的比较如下。相比而言，公司的 JM5400 具有功耗低，性能优的优势。

表 2：M9 VS JM5400

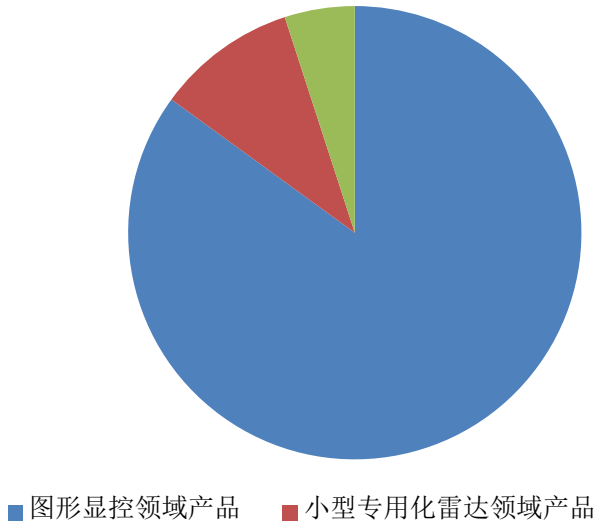
产品	工艺	外存类型	外存位宽（bit）	外存容量（MB）	外存带宽（GB/S）
M9（AMD）	150（nm）	DDR	128	64	6.4
JM5400（景嘉微）	65(nm)	DDR3	128	1024	9.6

资料来源：公司招股说明书，浙商证券研究所

虽然景嘉微的 GPU 芯片主要用于军用显示，尚无法达到人工智能深度学习的算力要求，但随着研发和支持的投入，参照 NVIDIA 当年的发展历史，景嘉微也会有潜力成长起来。

分析景嘉微的主营构成，主要分为图形显控领域产品、小型专用化雷达领域产品，以及其他。其中图形显控领域产品占公司收入的 85%多。公司的主要下游客户是军用飞机，目前我国大多数军用飞机都使用公司的图形显空产品

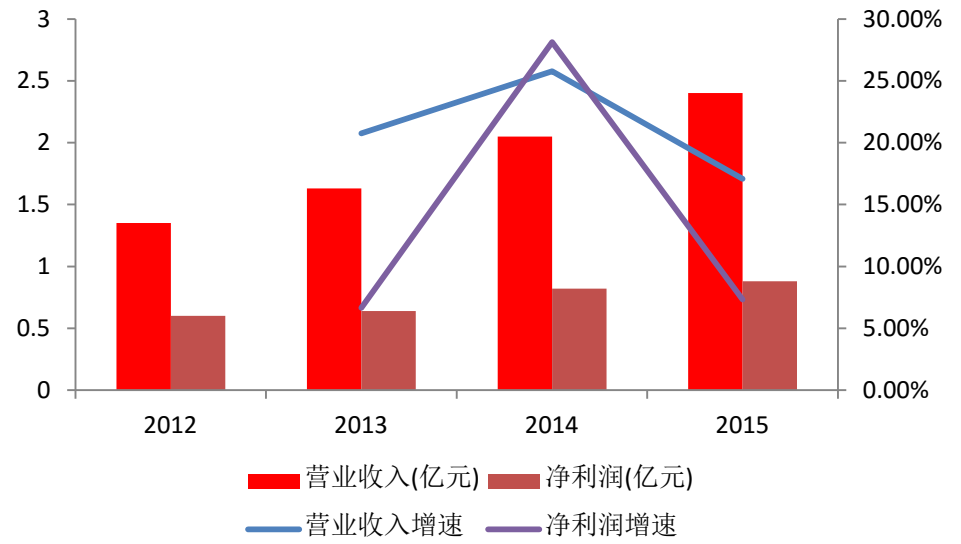
图 43：景嘉微 2015 年主营构成



资料来源：公司招股说明书，浙商证券研究所

公司从 2012-2015 年的营收和净利润都保持稳定的成长，呈现出向上的趋势。随着公司国产 GPU 的量产和替代，我们预计后续产品的营收和净利润将会得到进一步的提升。

图 54：景嘉微 2012-2015 年营收 VS 净利润



资料来源：Wind，浙商证券研究所

我们认为，国内的 GPU 发展尚处于起步阶段，目前的产品还是主要用于 GPU 原先的图形显控领域，虽然还不能跟现在人工智能深度学习所需要的 GPU 所媲美，但走在正确方向的道路上，未来也有可能得到突破。

3. FPGA——“万能芯片”在人工智能时代复苏

FPGA（Field－Programmable Gate Array），即现场可编程门阵列，它是在 PAL、GAL、CPLD 等可编程器件的基础上进一步发展的产物。FPGA 芯片主要由 6 部分完成，分别为：可编程输入输出单元、基本可编程逻辑单元、完整的时钟管理、嵌入块式 RAM、丰富的布线资源、内嵌的底层功能单元和内嵌专用硬件模块。

FPGA 还具有静态可重复编程和动态在系统重构的特性，使得硬件的功能可以像软件一样通过编程来修改。FPGA 能完成任何数字器件的功能，甚至是高性能 CPU 都可以用 FPGA 来实现。

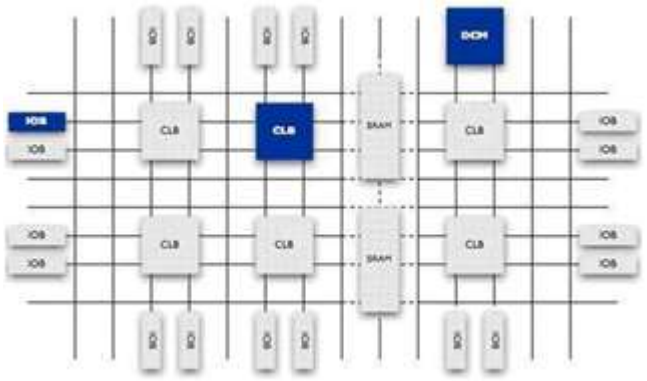
Intel 在 2015 年以 161 亿美元收购了 FPGA 龙头 Altera，其目的之一也是看中 FPGA 的专用计算能力在未来人工智能领域的发展。

3.1. FPGA——高性能、低功耗的可编程芯片

FPGA 之所以能有潜力成为人工智能深度学习方面的计算工具，主要原因就在于其本身特性：**可编程专用性，高性能，低功耗。**

先来看一下 FPGA 的内部架构。FPGA 拥有大量的可编程逻辑单元，可以根据客户定制来做针对性的算法设计。除此以外，在处理海量数据的时候，FPGA 相比于 CPU 和 GPU，独到的优势在于：**FPGA 更接近 IO**。换句话说，FPGA 是硬件底层的架构。比如，数据采用 GPU 计算，它先要进入内存，并在 CPU 指令下拷入 GPU 内存，在那边执行结束后再拷到内存被 CPU 继续处理，这过程并没有时间优势；而使用 FPGA 的话，数据 I/O 接口进入 FPGA，在里面解帧后进行数据处理或预处理，然后通过 PCIE 接口送入内存让 CPU 处理，一些很底层的工作已经被 FPGA 处理完毕了（FPGA 扮演协处理器的角色），且积累到一定数量后以 DMA 形式传输到内存，以中断通知 CPU 来处理，这样效率就高得多。

图 65：FPGA 内部架构



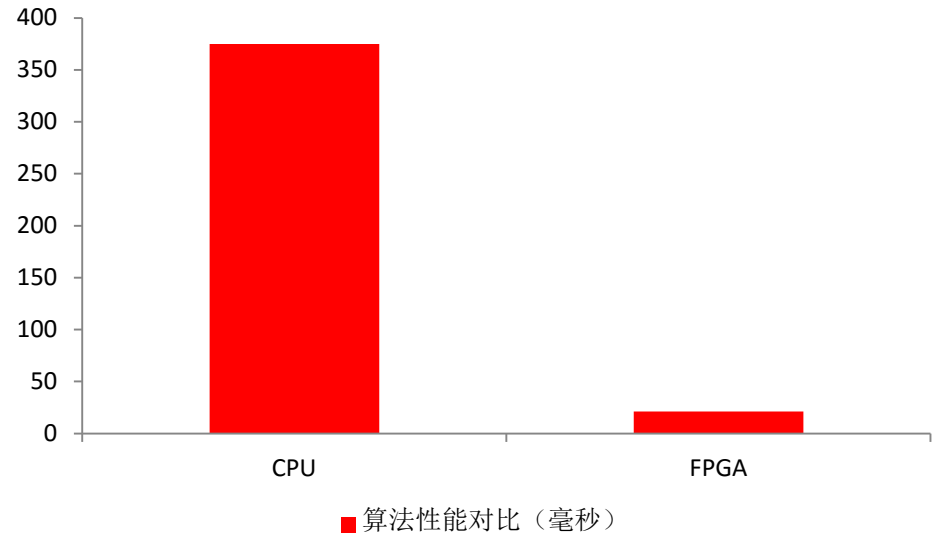
资料来源：互联网，浙商证券研究所

专用计算领域强过 CPU

虽然 FPGA 的频率一般比 CPU 低，但 CPU 是通用处理器，做某个特定运算(如信号处理，图像处理)可能需要很多个时钟周期，而 FPGA 可以通过编程重组电路，直接生成专用电路，加上电路并行性，可能做这个特定运算只需要一个时钟周期。比如一般 CPU 每次只能处理 4 到 8 个指令，在 FPGA 上使用数据并行的方法可以每次处理 256 个或者更多的指令，让 FPGA 可以处理比 CPU 多很多的数据量。举个例子，CPU 主频 3GHz，FPGA 主频 200MHz，若做某个特定运算 CPU 需要 30 个时钟周期，FPGA 只需一个，则耗时情况：CPU：30/3GHz =10ns；FPGA：1/200MHz = 5ns。可以看到，FPGA 做这个特定运算速度比 CPU 块，能帮助加速。

北京大学与加州大学的一个关于 FPGA 加速深度学习算法的合作研究。展示了 FPGA 与 CPU 在执行深度学习算法时的耗时对比。在运行一次迭代时，使用 CPU 耗时 375 毫秒，而使用 FPGA 只耗时 21 毫秒，取得了 18 倍左右的加速比

图 76：CPU,FPGA 算法性能对比



资料来源：浙商证券研究所

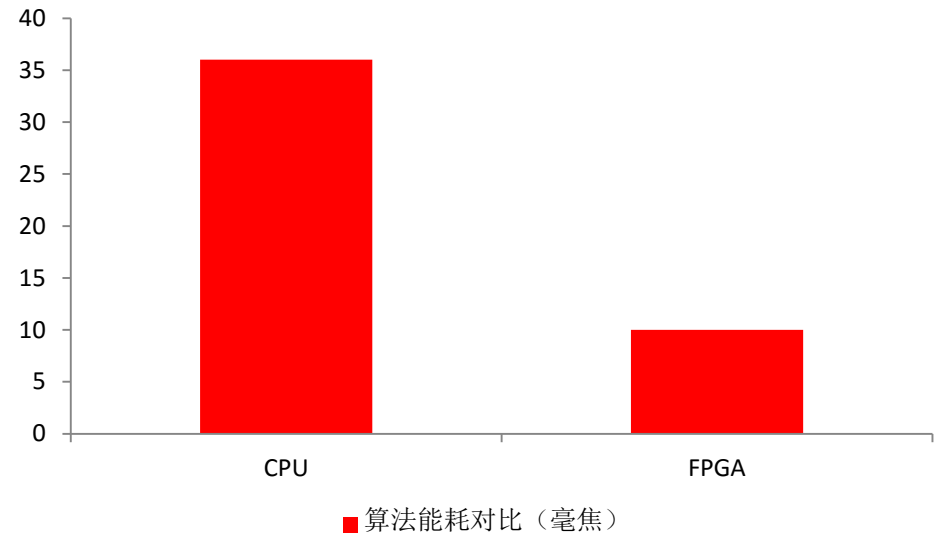
能耗显著降低

FPGA 相对于 CPU 与 GPU 有明显的能耗优势，主要有两个原因。首先，在 FPGA 中没有取指令与指令译码操作，在 Intel 的 CPU 里面，由于使用的是 CISC 架构，仅仅译码就占整个芯片能耗的 50%；在 GPU 里面，取指令与译码也

消耗了 10%~20%的能耗。其次，FPGA 的主频比 CPU 与 GPU 低很多，通常 CPU 与 GPU 都在 1GHz 到 3GHz 之间，而 FPGA 的主频一般在 500MHz 以下。如此大的频率差使得 FPGA 消耗的能耗远低于 CPU 与 GPU。

FPGA 与 CPU 在执行深度学习算法时的耗能对比。在执行一次深度学习运算，使用 CPU 耗能 36 焦，而使用 FPGA 只耗能 10 焦，取得了 3.5 倍左右的节能比。通过用 FPGA 加速与节能，让深度学习实时计算更容易在移动端运行。

图 87：CPU，FPGA 算法能耗对比



资料来源：浙商证券研究所

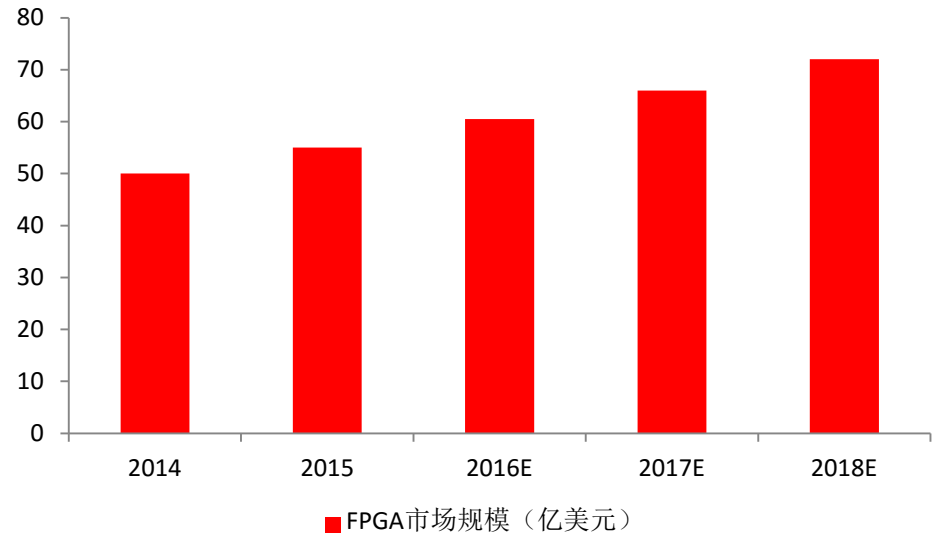
相比 CPU 和 GPU，FPGA 凭借比特级细粒度定制的结构、流水线并行计算的能力和高效的能耗，在深度学习应用中展现出独特的优势，在大规模服务器部署或资源受限的嵌入式应用方面有巨大潜力。此外，FPGA 架构灵活，使得研究者能够在诸如 GPU 的固定架构之外进行模型优化探究。

3.2. Intel 收购 Altera 分析

众所周知，在深度神经网络计算中运用 CPU、GPU 已不是什么新鲜事。虽然 Xilinx 公司早在 1985 年就推出了第一款 FPGA 产品 XC2064，但该技术真正应用于深度神经网络还是近几年的事。英特尔 167 亿美元收购 Altera，IBM 与 Xilinx 的合作，都昭示着 FPGA 领域的变革，未来也将很快看到 FPGA 与个人应用和数据中心应用的整合。

目前而言，FPGA 的应用领域以当前的通信、图像处理、IC 原型验证、汽车电子、工业等为主。整个 FPGA 市场由 Xilinx 和 Altera 主导，两者共同占有 85％的市场份额。FPGA 市场规模预计在 2016 年将达到 60 亿美元，并保持年复合增速 9%。

图 98：全球 FPGA 市场规模



资料来源：浙商证券研究所

根据 Altera 内部文件显示，Altera 很早就在研发使用 FPGA 针对深度学习算法的应用，并在 2015 年 Intel 的论坛上展示了产品的性能。结论是在功耗和性能上相对同等级的 CPU,有较大的优势。

图 109：Altera FPGA VS CPU

Projected AlexNet Performance and Power

CNN Classification Platform	Power (W)	Performance (Image/s)	Efficiency (Images/Sec/W)
E52699 Dual Xeon CPU (18 core per Xeon)	321	1320	4.11
PCIe w/Dual Arria 10 1150	130*	1200	9.27

资料来源：浙商证券研究所

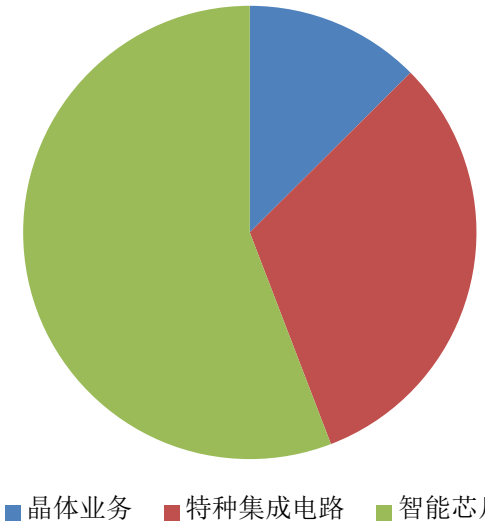
我们认为，Intel 之所以收购 Altera，主要原因就在于看中人工智能的发展，但 CPU 在计算能力上的先天不足，让其需要寻找一个合作伙伴。Altera 的 FPGA 正好弥补了 CPU 在这方面的缺陷，我们认为，CPU+FPGA 在人工智能深度学习领域，将会是未来的一个重要发展方向。

3.3. FPGA 国内行业与公司

FPGA 整个市场被国外的两大巨头所寡占，Xilinx 和 Altera 占了 85%的份额。国内目前也有一些公司在 FPGA 领域有所建树，其中比较优秀的有同方国芯。

同方国芯的主营业务包括晶体业务，特种集成电路，智能芯片这几块。各业务所占比重如下

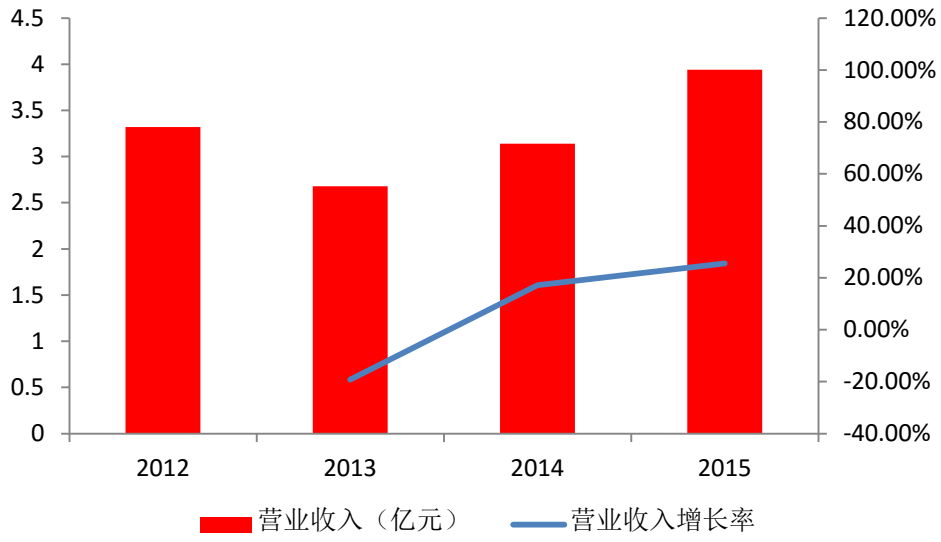
图 20：同方国芯 2015 年业务占比



资料来源：浙商证券研究所

同方国芯的 FPGA 归属于特种集成电路业务，我们统计上市以来这块业务的营收状况，保持着非常迅速的增长，年复合增速在 20%左右。

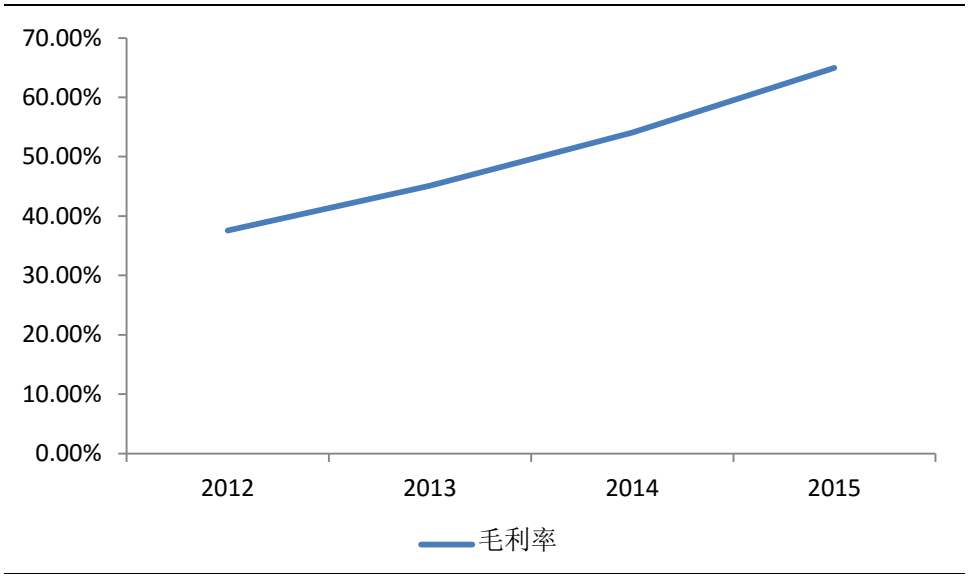
图 21：同方国芯特种集成电路（FPGA 等）业务营收



资料来源：浙商证券研究所

同时，特种集成电路的毛利率也一直维持在很高的水准，并持续往上升。我们预计随着 FPGA 在人工智能深度学习领域的应用增长，还会给公司带来持续性的利好增长。

图 22：同方国芯特种集成电路（FPGA 等）毛利率



资料来源：浙商证券研究所

4. ASIC——后起之秀，不可估量

4.1. 性能与功耗完美结合的 ASIC

ASIC(Application Specific Integrated Circuits，专用集成电路)，是指应特定用户要求或特定电子系统的需要而设计、制造的集成电路。 严格意义上来讲，ASIC 是一种专用芯片，与传统的通用芯片有一定的差异。是为了某种特定的需求而专门定制的芯片。

ASIC 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，具有以下几个方面的优越性：体积更小、功耗更低、可靠性提高、性能提高、保密性增强、成本降低。

回到深度学习最重要的指标：算力和功耗。我们对比 NVIDIA 的 GK210 和某 ASIC 芯片规划的指标，如下所示

表 3：GK210 指标 VS ASIC 指标

	GK210 指标	某 ASIC 芯片 2017 年规划指标
计算能力（TFLOPS）	4	10
内部存储器带宽（TB/S）	NA	3
内部存储器大小（MB）	10	256
外部 DDR 带宽（GB/S）	240	120
功耗（W)	150	10
成本（美金）	50	5

资料来源：浙商证券研究所

从算力上来说，ASIC 产品的计算能力是 GK210 的 2.5 倍。第二个指标是功耗，功耗做到了 GK210 的 1/15。第三个指标是内部存储容量的大小及带宽。这个内部 MEMORY 相当于 CPU 上的 CACHE。深度雪地的模型比较大，通常能够到几百 MB 到 1GB 左右，会被频繁的读出来，如果模型放在片外的 DDR 里边，对 DDR 造成的带宽压力通常会到 TB/S 级别。

因为全定制芯片 ASIC 综合考虑了工艺和性能方面的权衡，随着工艺的进步，性能和价格的进展如下：

图 24：工艺 VS 性能 VS 功耗

process vs speed/cost*speed/power						
item	65nm	40nm	28nm	16nm	10nm	7nm
speed	1.00	1.25	1.56	1.88	2.16	2.48
power	1.00	0.60	0.36	0.29	0.24	0.21
wafer(US$)	3500.00	4000.00	5200.00	10000.00	15000.00	24000.00
area	1.00	0.50	0.25	0.14	0.08	0.05
cost	0.050	0.028	0.018	0.019	0.018	0.017
full mask($)	0.6	1.0	1.8	5.0	?	?
speed/cost*speed/power	20.19	91.99	368.56	627.22	1084.31	1757.35
ratio		4.56	4.01	1.70	1.73	1.62

资料来源：浙商证券研究所

全定制设计的 ASIC，因为其自身的特性，相较于非定制芯片，拥有以下几个优势：

- 同样工艺，同样功能，第一次采用全定制设计性能提高 7.6 倍
- 普通设计，全定制和非全定制的差别可能有 1～2 个数量级的差异
- 采用全定制方法可以超越非全定制 4 个工艺节点（采用 28nm 做的全定制设计，可能比 5nm 做的非全定制设计还要好）

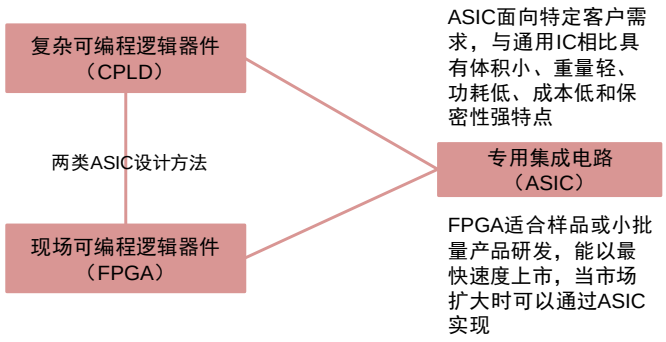
我们认为，ASIC 的优势，在人工智能深度学习领域，具有很大的潜力。

4.2. 从“比特币挖矿机 ASIC 发展”推导“ASIC 在人工智能领域大有可为”

ASIC 在人工智能深度学习方面的应用还不多，但是我们可以拿比特币矿机芯片的发展做类似的推理。**比特币挖矿和人工智能深度学习有类似之处，都是依赖于底层的芯片进行大规模的并行计算**。而 ASIC 在比特币挖矿领域，展现出了得天独厚的优势。

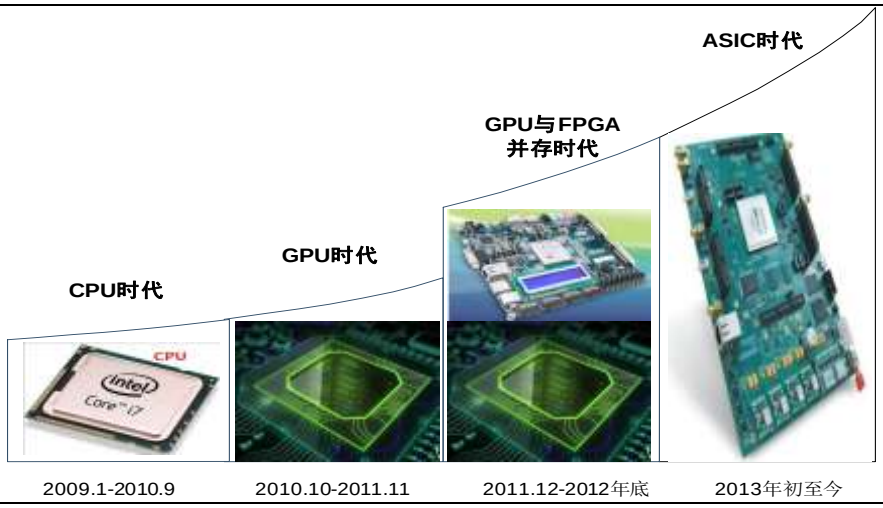
比特币矿机的芯片经历了四个阶段：CPU、GPU、FPGA 和 ASIC。其中 2009 年 1 月比特币创始人中本聪利用电脑 CPU 挖出了第一个创世块，其后大约一年时间 BTC 网络主要依靠 CPU 挖矿，CPU 设计中需要大量的逻辑判断和很强的通用性来处理不同类型的数据，而 GPU 处理简单的 SHA-256 算法速度更具优势； GPU 由于采用了大量并行处理的核心架构，对于简单的 SHA256 算法处理速度较快，2010 年 9 月挖矿进入了 GPU 时代，但是 GPU 也存在功耗高、搭建部署困难的缺陷，不适合大规模部署；2011 年 12 月出现了基于 FPGA 芯片的挖矿设备，其功耗为同类型的 GPU 的 1/40，但是 FPGA 芯片价格昂贵、部署也很复杂，主要被少数具备专业背景的矿工所使用，这个阶段 FPGA 和 GPU 成为挖矿的主力军； 2013 年首台基于 ASIC 芯片的 Avalon 矿机面世，挖矿进入了 ASIC 时代。ASIC 芯片是专为挖矿量身定制的芯片，它将 FPGA 芯片中在挖矿时不会使用的功能去掉，与同等工艺的 FPGA 芯片相比执行速度块，大规模生产后的成本也要低于 FPGA 芯片。

图 25：ASIC 芯片专为矿机量身定做，执行速度快于 FPGA



资料来源：浙商证券研究所

图 26：比特币矿机芯片经历了从 CPU、GPU、FPGA 和 ASIC 四个阶段



资料来源：互联网公开资料、浙商证券研究所

在 CPU、GPU 时代，挖矿门槛较低，家庭的普通台式机或者带有独立显卡的笔记本都可以用来挖矿，2012 年以前挖矿还是大众可以参与的相对公平对等阶段；随着 FPGA、ASIC 芯片的出现，挖矿逐渐开始向一些专业人士聚集。ASIC 芯片是为挖矿量身定制的，与同等工艺的 FPGA 芯片相比 ASIC 芯片的执行速度更快，大规模生产后成本也会比 FPGA 芯片低。目前 ASIC 芯片已成为主流的矿机芯片，挖矿速度基本都达到了 GH/S 的级别，比如 BITMAIN 的第四代芯片 BM1385，单颗芯片算力可达 32.5GH/S，在 0.66V 的核心电压下功耗仅为 0.216W/GH/S。ASIC 芯片随着硅片加工精度的提升，其性能更好，功耗更低。目前硅片加工精度已经 130nm 提升至 14nm，基本接近现有半导体技术的极限。

表 4：各种挖矿芯片的性能比较

比较项目	电脑 CPU	独立 GPU	FPGA	早期 ASIC
挖矿速度（MH/S）	20-40	300-400	200	289
矿机功耗（W）	100	130	10	6.6
价格（元/块）	1600	2000-3000	500 左右	60 左右
挖矿门槛	低	低	高	高
主要生产商	Intel、AMD	AMD、Nvidia	Altera、Xilinx、Actel、Lattice、Atmel	Alchip、KnCMiner、Avalon、BITMAIN、ASICMiner、BitFury

资料来源：中关村在线、互联网公开资料、浙商证券研究所

以上，从 ASIC 在比特币挖矿机时代的发展历史，可以看出 ASIC 在专用并行计算领域所具有的得天独厚的优势：**算力高，功耗低，价格低，专用性强。谷歌最近曝光的专用于人工智能深度学习计算的 TPU,其实也是一款 ASIC。**

4.3. ASIC 国内行业与公司

我们认为，国内的比特币芯片生产厂商，都有可能在人工智能时代华丽转身，成为拥抱深度学习的定制芯片供应商。**在这块领域有所深耕建树的公司有，国内的深圳烤猫、迦南耘智、比特大陆和龙矿科技**。拥有自产芯片的矿机生产商的盈利能力强，普遍的毛利率达到 50%以上。

5. 总结

综上，我们的观点：**人工智能时代逐步临近，GPU,FPGA,ASIC 这几块传统领域的芯片，将在人工智能时代迎来新的爆发。风起于青萍之末，一起关注人工智能时代芯片的大机会！**

股票投资评级说明

以报告日后的 6 个月内，证券相对于沪深 300 指数的涨跌幅为标准，定义如下：
1、买入 ：相对于沪深 300 指数表现 +20％以上；
2、增持 ：相对于沪深 300 指数表现 +10％～+20％；
3、中性 ：相对于沪深 300 指数表现－10％～+10％之间波动；
4、减持 ：相对于沪深 300 指数表现－10％以下。

行业的投资评级：

以报告日后的 6 个月内，行业指数相对于沪深 300 指数的涨跌幅为标准，定义如下：
1、看好 ：行业指数相对于沪深 300 指数表现 +10%以上；
2、中性 ：行业指数相对于沪深 300 指数表现－10%～+10%以上；
3、看淡 ：行业指数相对于沪深 300 指数表现－10%以下。
我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重。
建议：投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者不应仅仅依靠投资评级来推断结论

法律声明及风险提示

浙商证券研究所

上海市浦东南路 1111 号新世纪办公中心 16 层
邮政编码：200120
电话：(8621)80108518
传真：(8621)80106010
浙商证券研究所：http://research.stocke.com.cn