人工智能芯片行业

证券研究报告
2017 年 11 月 29 日

人工智能立夏已至，AI 芯片迎接蓝海；首推：英伟达 GPU 王者风范，Google TPU 破局科技

AI 芯片迎接蓝海，GPU 引领主流，ASIC 割据一地，看好未来各领风骚

在人工智能立夏将至的大趋势下，芯片市场蛋糕越做越大，足以让拥有不同功能和定位的芯片和平共存，百花齐放。后摩尔定律时代，我们强调 AI 芯片市场不是零和博弈。我们认为在 3-5 年内深度学习对 GPU 的需求是当仁不让的市场主流。行业由上至下传导形成明显的价值扩张，英伟达和 AMD 最为受益。

在深度学习上游训练端(主要用在云计算数据中心里)，GPU 是当仁不让的第一选择，但以 ASIC 为底芯片的包括谷歌的 TPU、寒武纪的 MLU 等，也如雨后春笋。而下游推理端更接近终端应用，需求更加细分，我们认为除了 GPU 为主流芯片之外，包括 CPU/FPGA/ASIC 等也会在这个领域发挥各自的优势特点。

深度学习下游推理端，FPGA 依靠电路级别的通用性，加上可编程性，适用于开发周期较短的 IoT 产品、传感器数据预处理工作以及小型开发试错升级迭代阶段等。以 TPU 为代表的 ASIC 定制化芯片，针对特定算法深度优化和加速，将在确定性执行模型(deterministic execution model)的应用需求中发挥作用。我们认为深度学习 ASIC 芯片，包括英特尔的 Nervana Engine、Wave Computing 的数据流处理单元、以及英伟达的 DLA 等逐步面市，将依靠特定优化和效能优势，未来在细分市场领域发挥所长。

行业首推：英伟达 GPU 王者风范，Google TPU 破局科技

1、英伟达：1）AI 起锚，数据中心业务 2020 年前有望翻 4 倍：我们一直强调数据中心升级加速过程中的巨大空间。深度学习上游训练端由 GPU 主导并基本为英伟达所垄断。我们预测今年数据中心收入将达 18.4 亿美元，同比涨超 120%。2）自动驾驶广泛布局，3-5 年期长期驱动：自动驾驶领域正在向“车企+供应商+芯片巨头+打车软件+物流公司”的组合格局发展，英伟达与英特尔-Mobileye 联盟形成两大上游竞争者。3）以游戏业务为现金马，三驾马车齐发力。我们认为英伟达将持续巩固 GPU 市场龙头地位，保持现有业务充沛活力的同时自上而下推动 AI 浪潮，1300 亿美元市值还仅是 AI 立夏开端。公司 27 日收盘价 214.14 美元，我们给予公司 2018/19 年 EPS 分别为 5.42、6.61 美元，对应 2018/19 年 PE 52/42x，目标价从 250 上调至 280 美元，重申“买入”评级。

2、AMD 在 GPU 和 CPU 市场，都屈居行业老二，人工智能芯片布局上也慢英伟达一步。但作为唯一拥有 GPU 和 x86 硅芯片技术的公司，我们认为随着公司产品线上移，重回高端市场，并且从零到一破局数据中心市场，利用 GPU+CPU 异构计算技术储备的协同效应，进一步修复利润率，在公司 CEO Lisa Su 的带领下，打开与英伟达、英特尔正面竞争之外的市场，200 亿美元市值亦可期。公司 2018 年 PS 为 2.18x，对比英伟达 10x 仍被低估，公司 27 日收盘价 11.55 美元，我们预测公司 2018 年营收/EPS 分别为 56.85 亿美元/0.35 美元，我们认为 2.65x PS 和 45x PE 较合理，重申“买入”评级，目标价维持 16 美元。

3、Google 依靠 TPU 自下而上冲入云端，针对 TensorFlow 深度优化，提高浮点运算精度，将 TPU 部署在云计算中以云服务形式进行销售共享，进一步激活中小企业的云计算需求市场，另辟 AWS、Azure 之外蹊径。我们看好谷歌的新征途——云+YouTube+硬件持续助推转型，长期看 AI 和 Other Bets 创新业务厚积薄发：谷歌是人工智能的龙头标的，我们长期看好语音识别和无人驾驶的发力。公司 27 日收盘价 1072.01 美元，根据彭博一致预期 2018 年 EPS 41.46 美元，给予 31x PE，目标价从 1200 上调至 1300 美元，重申“买入”评级。

风险提示：芯片开发周期过长，市场需求不达预期等。

作者

何翩翩 分析师
SAC 执业证书编号：S1110516080002
hepianpian@tfzq.com

雷俊成 联系人
leijuncheng@tfzq.com

马赫 联系人
mahe@tfzq.com

董可心 联系人
dongkexin@tfzq.com

相关报告

1 《谷歌 TPU 及强化学习：谷歌 TPU 以时间换吞吐量，加速云端 AI 帝国；AlphaGo 从 Lee 到零，探索强化学习新起点》2017-11-22

2 《GPU 行业点评：GPU 需求和虚拟货币的关系：“微小但不是零的”，Jensen 强调 5 次也不为过》2017-11-20

3 《英伟达（NVDA.US）深度：人工智能立夏已至，AI 芯片迎接蓝海；英伟达 AI 时代最强音，重申买入，TP 上调至 250 美元》2017-11-12

4 《AMD（AMD.US）深度：扭亏为盈，Q2 超预期，CPU+GPU 双剑合璧的唯一，重申“买入”，目标价维持 16 美元》2017-08-17

5 《人工智能芯片行业点评：英伟达 GPU 王者风范，Google TPU 破局科技；人工智能冲入云霄，看好 GPU、ASIC 各领风骚》2017-05-31

6 《Mobileye（MBLY.US）深度报告：ADAS 龙头获 Intel 收购，打造无人驾驶的 Android，驰骋于强化机器学习的征途上》2017-03-14

7 《2017 MIT 人工智能 5 大趋势预测：寒梅傲香春寒料峭，人工智能立夏将至》2017-01-25

8 《谷歌人工智能深度解剖：从 HAL 的太空漫游到 AlphaGo，AI 的春天来了》2017-01-05

内容目录

图表目录

1. 人工智能“脑力”基础：AI 芯片繁荣共生，创造非零和博弈的一片蓝海

正如 20 年前多媒体应用及 3D 游戏蓬勃发展倒逼显卡硬件升级一样，互联网大数据的兴起对超算芯片提出了新的需求。人工智能的“脑力”核心在于算法和芯片，目前深度学习/强化学习算法已在 AlphaGo Master 和 Zero 身上淋漓彰显。深度学习算法意味着通过构建一种深层非线性网络结构，来实现复杂函数逼近及自动特征提取，具有强大的从少数样本集中挖掘数据统计规律的能力。

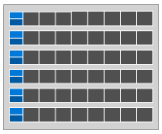
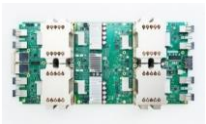
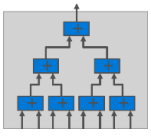
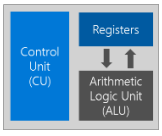
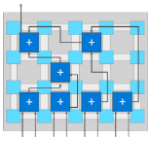
但是，人工智能、深度学习从来不是新的概念。深度学习为何突然重新回到大众视野？而且随着模型的逐渐复杂化，浮点运算的数量也呈指数级增长至 ExaFLOPS。2015 年微软 ResNet 含有 6000 万个参数，运算量为 7 ExaFLOPS（百亿亿次浮点运算）。2016 年百度语音识别系统 Deep Speech 2 的参数量上升到 3 亿个，运算量提升至 20 ExaFLOPS。而今年 Google 的 NMT 神经网络机器翻译系统，参数量达 87 亿个，需要 105 ExaFLOPS 的运算量。

本质上，摩尔定律的突破和并行计算以及云计算的发展，让人工智能开始得以普及。或者一言以蔽之，就是 GPU 的加入。没有 GPU，人们就无法快速的处理海量数据，而数据训练的匮乏，会让深度学习的效率还不如人类工程算法(human engineering algorithm)。所以很长的一段时间，人们认为深度学习很有趣，但效率不高，直到 GPU 和 CUDA 的部署加入，深度学习/神经网络才获得惊人发展。这其实是个“chicken and egg”的问题，大数据、深度学习、GPU 的一同出现造就了当前的人工智能繁景，这也是为什么人们把英伟达比作深度学习的三大建立者之一。

1.1. AI 芯片繁荣共生，GPU 引领主流，ASIC 割据一地，看好未来各领风骚

AI 芯片市场蛋糕越做越大，足以让拥有不同功能和定位的芯片和平共存，百家争鸣非零和博弈。“通用性和功耗的平衡”——在深度学习上游训练端（主要用在云计算数据中心里），GPU 是当仁不让的第一选择，ASIC 包括谷歌 TPU、寒武纪 NPU 也如雨后春笋。而下游推理端更接近终端应用，需求更加细分，GPU 主流芯片之外，包括 CPU/FPGA/ASIC 也会在这个领域发挥各自的优势特点。

图 1：目前深度学习领域常用的四大芯片类型，“通用性和功耗的平衡”

	训练端	推理端	
	**GPU**：以英伟达为主，AMD为辅标榜通用性，多维计算及大规模并行计算架构契合深度学习的需要。在深度学习上游训练端（主要用在云计算数据中心里），GPU是当仁不让的第一选择。	**GPU**：英伟达Volta GPU也开始布局推理端。深度学习下游推理端虽可容纳CPU/FPGA/ASIC等芯片，但竞争态势中英伟达依然占主导。	
	**ASIC**：以谷歌的TPU、英特尔的Nervana Engine为代表，针对特定框架进行深度优化定制。但开发周期较长，通用性较低。比特币挖矿目前使用ASIC专门定制化矿机。	**ASIC**：下游推理端更接近终端应用，需求也更加细分，英伟达的DLA，寒武纪的NPU等逐步面市，将依靠特定优化和效能优势，未来在深度学习领域分一杯羹。	
	**CPU**：通用性强，但难以适应于人工智能时代大数据并行计算工作。	**FPGA**：依靠可编程性及电路级别的通用性，适用于开发周期较短的IoT产品、传感器数据预处理工作以及小型开发试错升级迭代阶段等。但较成熟的量产设备多采用ASIC。	

资料来源：微软 Build，谷歌官网，天风证券研究所整理

1、 **GPU**：以英伟达为主，AMD 为辅。依靠通用及灵活的强大并行运算能力，广泛契合当前人工智能监督深度学习以及生成式对抗网络(GAN)/强化学习所需要的密集数据和多维并算处理需求，在 3-5 年内 GPU 仍然是深度学习市场的第一选择。

AI 将至的大趋势下，数据中心市场空间巨大。深度学习上游训练端由 GPU 主导并基本为英伟达所垄断。下游推理端虽可容纳 CPU/FPGA/ASIC 等芯片，但竞争态势中英伟达依然占大头。英伟达依靠 Volta 构架升级以及广泛成熟的开发生态环境，自上而下的

对训练、推理兼顾，扩张版图。以 2016 年为例，全年服务器市场出货量约在 1110 万台，在只有 7%用于人工智能 workload，其中约 3.4%配置 GPU，总量仅 2.6 万台。所以全球新增服务器中 GPU 的渗透率仅为 0.24%，我们预计在 2020 年前全球服务器 GPU 渗透率将达 4 倍以上增长。

2、**ASIC（Application Specific Integrated Circuit，专用集成电路）：**细分市场需求确定后，以 TPU 为代表的 ASIC 定制化芯片（或者说针对特定算法深度优化和加速的 DSA, Domain-Specific-Architecture），将在确定性执行模型(deterministic execution model)的应用需求中发挥作用。例如比特币早年间的挖矿热潮就从 GPU 通用算力堆积逐步转向了 ASIC 专用矿机。我们认为深度学习 ASIC 包括英特尔的 Nervana Engine，Wave Computing 的数据流处理单元，英伟达的 DLA，寒武纪的 NPU 等逐步面市，将依靠特定优化和效能优势，未来在深度学习领域分一杯羹。

3、**FPGA（Field Programmable Gate Array，现场可编程门阵列）：**依靠电路级别的通用性，加上可编程性，适用于开发周期较短的 IoT 产品、传感器数据预处理工作以及小型开发试错升级迭代阶段等。但一般较成熟的量产设备大多采用 ASIC。FPGA 厂商包括 Xilinx、Altera（英特尔）、Lattice 及 Microsemi。

图 2：目前深度学习领域常用的四大芯片类型及主要芯片商

类别	GPU	ASIC	ASIC：TPU	CPU	FPGA
特点	1.可多达上千个简单核心，上千个并行硬件线程 2.并行运算能力、浮点运算能力强大 3.最大化浮点运算数据吞吐量	1. 需求确定后可进行专门优化设计 2. 优秀的功耗控制 3. 性能稳定、可靠性高	1.与TensorFlow深度结合，更接近DSA (Domain-Specific-Architecture) 2.已能同时用于高性能计算和浮点计算 3.结合谷歌云提供云计算服务	1.通用性强 2.核心复杂程度高 3.串行运算能力强，单线程性能优化 4.晶体管空间用于复杂并行性指令(Complex ILP)	1.电路级别的通用性 2.可编程性 3.适用于开发周期较短的IoT产品、传感器数据预处理工作以及小型开发试错升级迭代阶段
主要厂商	英伟达、AMD、Imagination等	英特尔、德州仪器、三星、高通等	谷歌	英特尔、AMD、高通等	Xilinx、Altera（已被Intel收购）、Lattice、Microsemi

资料来源：微软 Build，谷歌官网，天风证券研究所整理

图 3：四大芯片的“通用性和功耗的平衡”

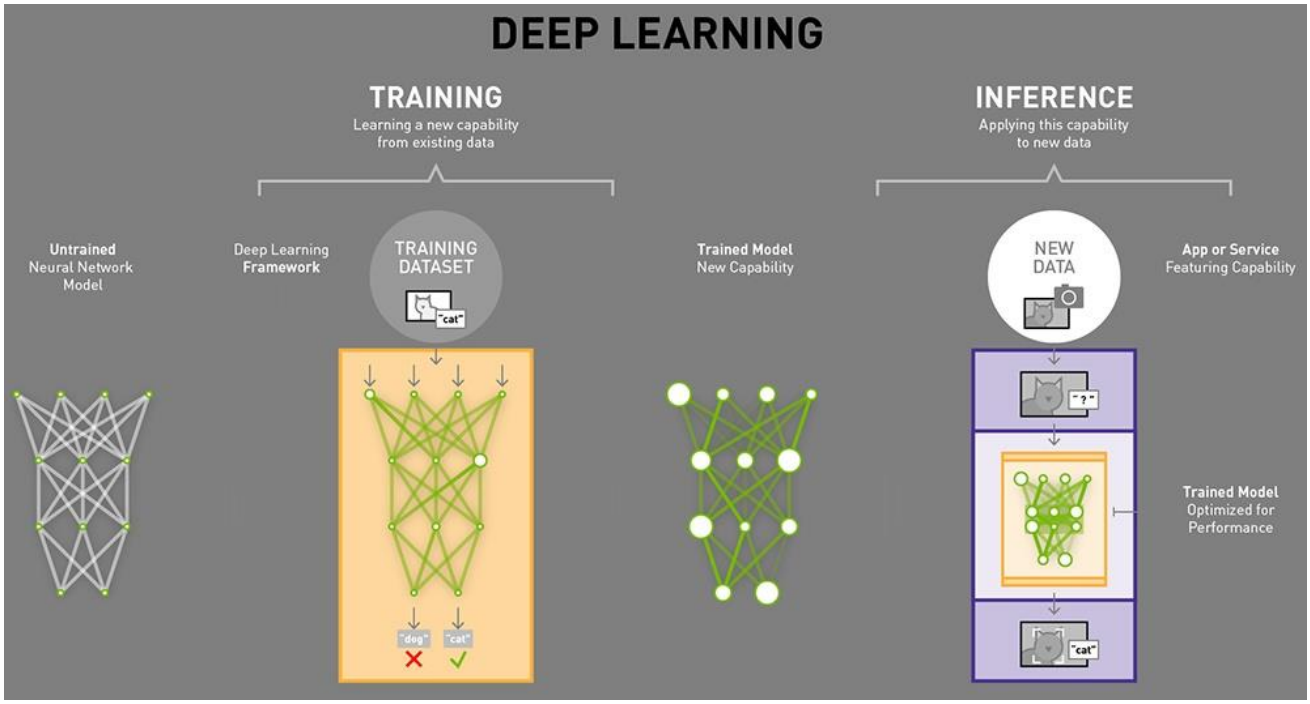
资料来源：微软 Build，天风证券研究所

我们认为，AI 芯片市场的花团锦簇，百家争鸣非零和博弈。“通用性和功耗的平衡”——在深度学习上游训练端，GPU 是当仁不让的第一选择。而下游推理端更接近终端应用，需求也更加细分，包括 CPU/FPGA/ASIC 和 GPU 都会在这个领域发挥各自的优势特点。

英伟达拥有目前最为成熟的开发生态环境（CUDA 因统一而完整的开发套件，丰富的库以及对英伟达 GPU 的原生支持而成为开发主流，目前已开发至第 9 代，开发者人数超过 51 万）。Google 的 TPU 也结合 TensorFlow 开源开发环境。而 AMD 通过 CPU+GPU+ROCm 的开源生态打造 GPU 计算最通用开源平台。开源时代生态为天，硬件厂商以开源之态，本质上是抢夺业界事实标准的控制权，但随之而来的也是整个芯片行业设计门槛和研发成本的不断降低。

在未来，随着 AI 算法的推进、芯片材料的多样化、芯片供电方法、能耗的节约和工艺水平等都逐渐将芯片的设计门槛“平民化”。量子计算机和人脑模拟芯片等新型硬件也将改变市场格局。AI 芯片正进入下一片蓝海，风物长宜放眼量，我们认为 AI 芯片会走出“CPU+FPGA+GPU—特制芯片—量子芯片”的征途。

图 4：深度学习在神经网络模型的应用中主要分为上游训练端和下游推理端



资料来源：英伟达官网，天风证券研究所

1.2. 行业首推：英伟达 GPU 王者风范，Google TPU 破局科技

1.2.1. 英伟达 GPU 王者风范，TP 上调至 280 美元

我们重申英伟达三大投资亮点：1）AI 起锚，数据中心业务 2020 年前有望翻 4 倍；2）无人驾驶业务有望跟 Mobileye 平分秋色；3）以游戏业务为现金马，三驾马车齐发力。

我们认为英伟达持续巩固 GPU 市场龙头地位，保持现有业务充沛活力的同时自上而下推动 AI 浪潮，1300 亿市值仅是 AI 立夏开端。

1）未来 2-3 年数据中心业务仍是公司的爆发增长点，我们看好人工智能立夏将至，数据中心增长空间巨大，深度学习上游训练端由 GPU 主导并基本为英伟达所垄断，下游推理端虽可容纳 CPU/FPGA/ASIC 等芯片，但竞争态势中英伟达依然占主导。我们预计未来三年增速 CAGR 可达 70%，在 2020 年前将实现 40 亿美元收入，对比 2016 年 8.3 亿美元。

2）汽车业务将在未来 3-5 年内，随着无人驾驶的普及接棒数据中心成为新的爆发点。自动驾驶领域正在向「车企+供应商+芯片巨头+打车软件+物流公司」的组合格局发展，形成英伟达与英特尔-Mobileye 联盟的两大竞争者。英伟达目前仍在布局阶段，重视车载本地超级电脑的研发，未来发展有待驾驶决策软件算法落地，以及 Drive PX 平台成本降低。

3）游戏业务是公司业务的“现金马”，稳固增长依托高端 PC 游戏、VR 以及电子竞技热情以及用户基数升级周期。

我们认为英伟达将持续巩固 GPU 市场龙头地位，保持现有业务充沛活力的同时自上而下推动 AI 浪潮，1300 亿美元市值还仅是 AI 立夏开端。**估值方面：我们给予公司 2018/19 年 EPS 分别为 5.42、6.61 美元，对应 2018/19 年 PE 52/42x，目标价从 250 上调至 280 美元，重申“买入”评级。**

1.2.2. AMD 行业老二不遑多让，TP 16 美元

AMD 在 GPU 和 CPU 市场，都屈居行业老二的位置，当下在人工智能芯片布局上也慢英伟达一步。但是作为唯一拥有 GPU 和 x86 硅芯片技术的公司，我们认为随着公司产品线上移，重回高端市场，并且从零到一破局数据中心市场，利用 GPU+CPU 异构计算技术储备的协同效应，进一步修复利润率，在公司 CEO Lisa Su 的带领下，打开与英伟达、英特尔正面竞争之外的市场，200 亿美元市值亦可期。

公司 2018 年 PS 为 2.3x，对比英伟达 11.7x，我们预计公司今明年毛利率将进入 34-36%区间，并实现 Non-GAAP EPS 转正，重新走上盈利正轨。利用 GPU+CPU 异构计算技术储备的协同效应，自上而下的产品线羽翼渐丰，云计算和人工智能布局为 AMD 带来更高估值弹性。并依托中科院/中科曙光等国内芯片最高生产力和资源抢占国内数据中心处理器先机，加速切入国内 AI 发展快车道。**我们预测公司 2018 年营收/EPS 分别为 56.85 亿美元/0.35 美元，我们认为 2.65x PS 和 45x PE 较合理，重申“买入”评级，目标价维持 16 美元。**

1.2.3. Google 软硬兼施，打造 AI 帝国，TP 上调至 1300 美元

Google 依靠 TPU 自下而上冲入云端，针对 TensorFlow 深度优化，提高浮点运算精度，将 TPU 部署在云计算中以云服务形式进行销售共享，进一步激活中小企业的云计算需求市场，另辟 AWS、Azure 之外蹊径。我们早在年初已经开始强调，人工智能巨头新征途——云+YouTube+硬件。YouTube & 云计算的巨大增长动力将是谷歌持续转型的助推器，长期看好 AI 和 Other Bets 创新业务厚积薄发。**根据彭博一致预期 2018 年 EPS 41.46 美元，给予 31x PE，目标价从 1200 上调至 1300 美元，重申“买入”评级。**

另外，传统巨头英特尔依靠大举收购，全面布局 AI 底层生态，以及 Xilinx、Broadcom 等芯片厂在市场扩张形势下的长期增长前景也不容忽视。

2. 英伟达和 AMD：GPU 巨头的引领 AI 盛夏

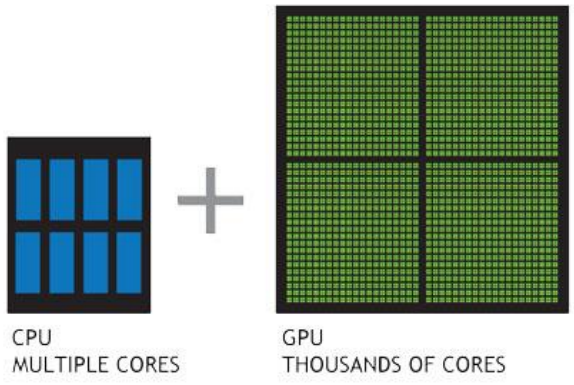
机器学习得以普及起来，其中一个重要原因是计算能力的提升和 GPU 的出现。图形处理器(Graphics Processing Unit, GPU)是主要用来处理图形数据的芯片处理单元。GPU 在执行复杂的多维计算和几何计算中十分有效，因此被广泛地运用在图像和图形处理中。现在 GPU 的运用不单在 3D 图形处理，由于 GPU 调用多个核心，让 GPU 可以同时对多像素进行处理。这个特点正广泛地运用在有并行特性(Data-Parallelism)的应用中，比如说大数据的多任务处理。

在 1999 年，英伟达带来了被认为是世界上第一款消费者级别的 3D 图形 GPU，GeForce 256，GPU 首次被当作计算机中的一个独立处理芯片。进入 2000 年后，GPU 在 3D 图形的渲染计算能力上得到巨大发展，可编程着色技术(programmable shading)和浮点运算能力(floating point abilities)是实时处理 3D 图形加速技术中最大的飞跃。

随着英伟达 GeForce 8 系列的推出，GPU 成为了一种用途更为广泛的计算设备。现今，并行 GPU 已经开始在计算方面与传统的 CPU 竞争。其中随着研究的细分，GPU 计算以及通用型 GPU（简称 GPGPU）的应用领域已经十分多样化，其中包括深度学习、石油开发、科学图片处理、线性计算、统计、3D 构建甚至股票期权定价策略等。

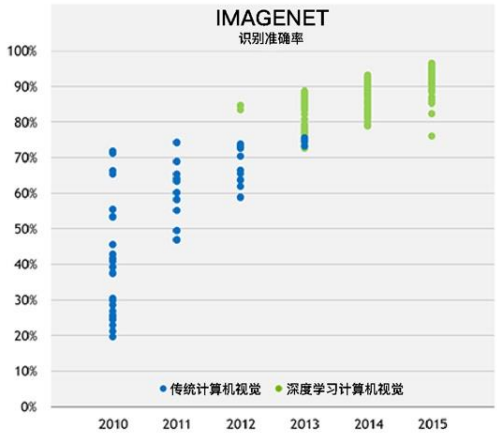
简单理解 CPU 与 GPU 之间的区别的话，那就是对比二者的任务处理能力。CPU 有着较少量的核来进行最大化的连续串行处理，而 GPU 则有着极大的并行结构，这种结构包含了数以千计的、高效计算能力的核，因此在同时处理多个问题中显得更加得心应手。

图 5：GPU 和 CPU 结构上的区别



资料来源：英伟达官网，天风证券研究所

图 6：深度学习在计算机视觉领域的优越表现



资料来源：英伟达官网，天风证券研究所

GPU 比 CPU 拥有更多的运算器(ALU, Arithmetic Logical Unit)，因此在处理庞大的数据中，GPU 可以做得更高效。一个 CPU 核可以同时执行 4 项 32 位指令（用 CPU 中的 128 位 SSE 指令集）或者通过 256 位高级矢量扩展指令集(AVX)执行 8 个指令集。但 GPU 如 AMD 的 Radeon HD 5970，则可以执行 3200 个 32 位的指令（通过其 3200 个运算器）。二者之间的运算效率的差距达到 800 倍（如果使用 AVX 则是 400 倍）之多。GPU 的高运算性能让它能够应用在科学计算、密码破解、数值分析、海量数据处理等方面。

GPU 大规模并行计算(parallel computing)的能力得到充分利用，被运用在当前最前沿的人工智能神经网络算法中。2007 年开始，为游戏中 3D 实时处理而设计的显卡为 GPU 的每秒浮点运算次数(FLOPS)带来了突破性的进展，计算速度的突飞猛进让科学家大量的将 GPU 运用到人工智能当中。在 2012 年，谷歌的人工智能团队进行人工神经网络的研究中，首次将人工智能用于分析 YouTube 的视频内容。通过由英伟达 GPU 组成的神经网络去识别 YouTube 中有猫的视频，并成功地做出了数以万计的识别。《自然》杂志表示，随着 GPU 的出现，研究人员在对深度学习神经网络进行训练的速度得到了 10-20 倍的提升。英伟达也表示，他们在对 GPU 设计、系统架构、编译器、算法等方面进行改进后，在短短三年时间内，将深度神经网络训练的速度提高了 50 倍。

2012 年，多伦多大学 Alex Krizhevsky 团队获得了当年计算机图像识别大赛的冠军。他们利用英伟达的 GPU 构建神经网络，由于其大模块数据处理的特性，该团队的 Alexnet 神经网络在没有构建别的程序的情况下得以利用深度学习来识别手画。

到了 2015 年，人工智能获得里程碑式的发展。通过深度学习，谷歌和微软分别在 ImageNet 比赛中战胜人类选手，获得最高分；微软与中科大合作建立的深度学习神经网络在 IQ 测试中达到了大学毕业生的标准；百度宣布 Deep Speech 2 可以同时对英语和普通话进行深度学习。英伟达 GPU 的并行多任务处理能力让深度学习神经网络增速，为研究人员节约数周的迭代训练。另外，英伟达的 CUDA 平台的可编程性与丰富性帮助研究者在深度学习神经网络(DNN)、递归神经网络(RNN)以及 LSTM 网络中进行创新。

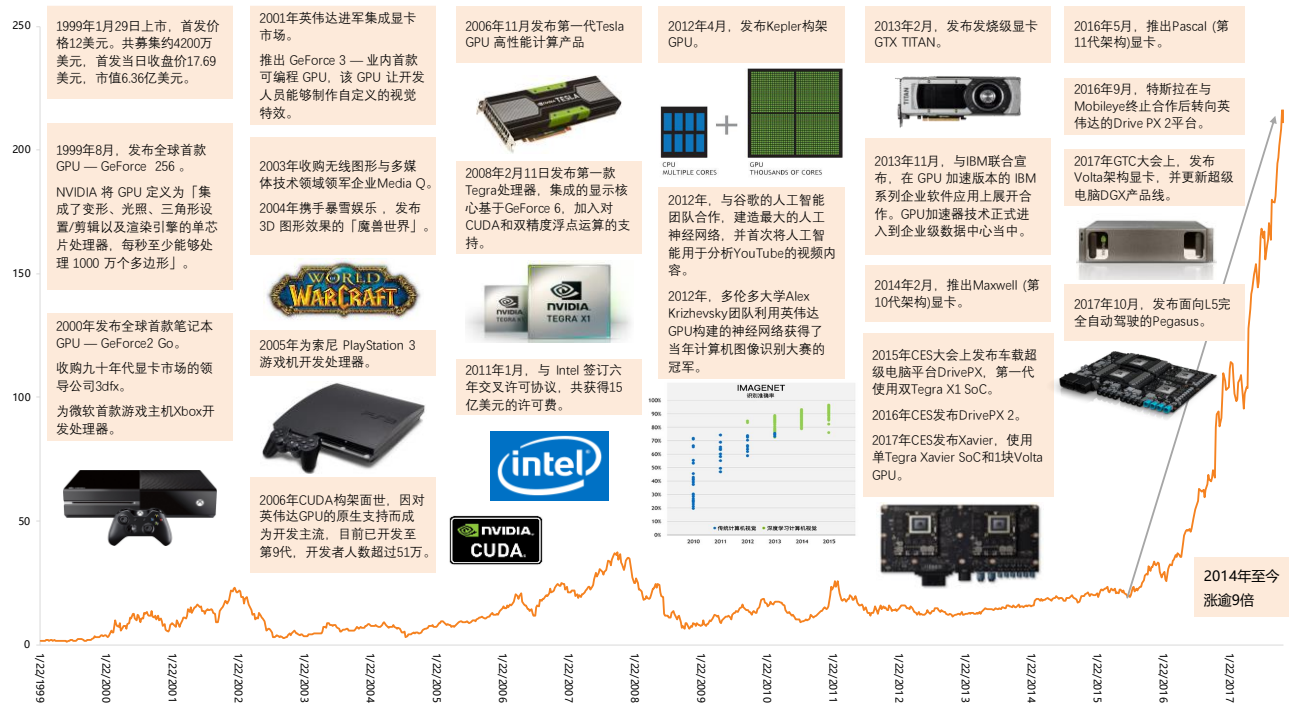
2.1. 英伟达：GPU 巨头的 AI 时代最强音

1999 年，英伟达带来了被认为是世界上第一款消费者级别的 3D 图形 GPU，GeForce 256，GPU 首次被当作计算机中的一个独立处理芯片。21 世纪的前十年，英伟达和 ATI（之后的 AMD）难解难分，曾在主板芯片市场于 AMD 合作共击英特尔，也有 ATI-AMD 合并后割据 GPU 市场。2004 年到 2007 年，英伟达首席科学家 David Kirk 一直在思考让只做 3D 渲染的 GPU 技术通用化，于是包括 CUDA 构架，大规模并行运算芯片 Tesla 系列应运而生。

2012 年，英伟达与谷歌的人工智能团队合作，建造最大的人工神经网络，并首次将人工智能用于分析 YouTube 的视频内容。深度神经网络技术在通用 GPU 的支持下实现重大突破，英伟达为包括计算机视觉、语音识别、自然语言处理在的人工智能里程碑式的发展提供了长足的计算动力。

2016 年 4 月，英伟达发布 Pascal 架构 GPU Tesla P100。英伟达 CEO 黄仁勋表示，该款 GPU 较之以往产品可以让深度学习神经网络的处理速度提高 12 倍。P100 的研发耗资 20 亿美元。以往的 Tesla GPU 产品，例如 Quadro Plex 2200 S4，最大只有 4 核 GT200GL 芯片，而 DGX-1 则有 8 块 P100 芯片，配置成本更达到了 12.9 万美元。Tesla P100 被首先配送到 MIT、斯坦福大学和加州大学伯克利分校的人工智能研究部门。

图 7：英伟达历史大事件



资料来源：公司官网，Bloomberg，天风证券研究所整理，数据截止至 2017 年 11 月 27 日，采用调整后收盘价

2.1.1. 数据中心抢滩战，英伟达先拔头筹

我们一直强调数据中心升级加速过程中的巨大空间。深度学习上游训练端由 GPU 主导并基本为英伟达所垄断。下游推理端虽可容纳 CPU/FPGA/ASIC 等芯片，但竞争态势中英伟达依然占主导。以 2016 年为例，全年服务器市场出货量约在 1110 万台，只有 7%用于人工智能 workload，其中约 3.4%配置 GPU，总量仅 2.6 万台。所以全球服务器中 GPU 的渗透率仅有 0.24%。我们预计英伟达数据中心业务在 2020 年前将达 40 亿美元，对应全球服务器 GPU 渗透率也将达 4 倍以上增长。

英伟达数据中心业务去年 Q2 实现同比 110%增长，下半年开始成为公司第二大业务，Q3、Q4 均实现约 200%的同比增长，耀眼增速的背后是去年 4 月发布的 Pascal 架构 Tesla P100 GPU 在 6 月正式出货，提供了前任 Kepler 架构 K40、K80 超过 50 倍的性能提升，为深度学习神经网络的训练端带来了真正的强大算力。P100 的惊艳出世收获的不只是用户的大面积升级换新，也为英伟达 2016 年数据中心业务带来全年 8.3 亿美元收入，同比增长 145%。

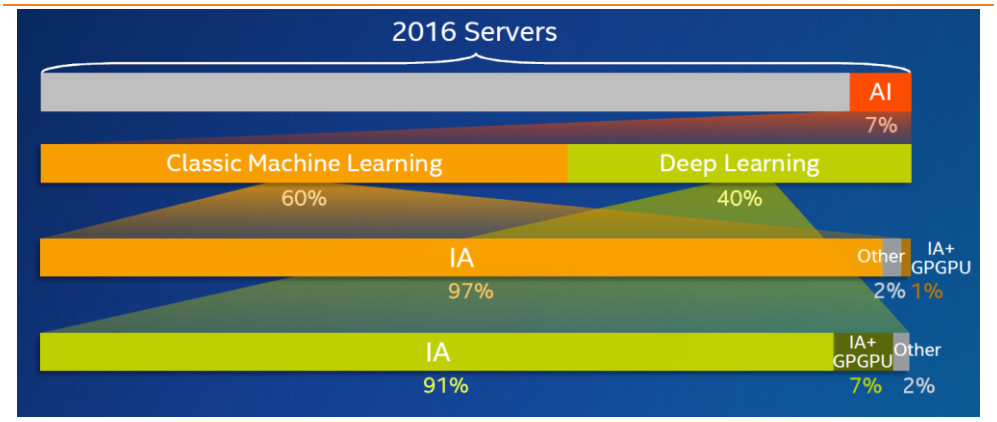
今年数据中心业务增长的动力则落在了 Volta 架构 V100 的身上，随着亚马逊宣布部署英伟达 V100 GPU 实例，包括微软、谷歌、Oracle、国内 BAT，以及所有服务器 OEM 厂商的合作推进都在稳步进行。今年前 9 个月英伟达数据中心收入已达 13.26 亿美元，同比增长 148%。我们认为 Volta 的放量将会进一步对 Q4 和明年的数据中心业务增长产生贡献。

目前全球云计算巨头基本使用英伟达 GPU 进行深度学习与算法加速。英伟达发布全新 Volta 构架 Tesla V100 GPU，配合张量处理指令 Tensor Core，将训练吞吐量提高至上代 Pascal 的 12 倍。针对推理端推出的 TensorRT 推理引擎，自上而下的对训练、推理兼顾，扩张版图。我们认为，英伟达将依靠 Volta 构架升级以及广泛成熟的开发生态环境，在数据中心加速市场中抢滩训练端，与 AMD 的竞争中稳固先发优势，并向推理端加速布局。

2.1.2. 数据中心 GPU 空间测算：蓝海正当时

根据英特尔给出的当前数据中心市场的人工智能渗透率，2016 年全球服务器约有 7%用于人工智能相关工作负载。根据 Gartner 统计，2016 年全年服务器市场出货量约在 1110 万台，代表约 78 万台用于人工智能。这其中约 3.4%配置有 GPU，即 2.6 万台。考虑到英伟达在训练端基本处于垄断地位，这 2.6 万台服务器应该基本使用英伟达 GPU。另外我们也可以看到，全球服务器中 AI workload 对 GPU 的使用率仅为 0.24%。如今 AI workload 渗透率尚低，我们认为 GPU 拥有更为广阔的渗透空间。

图 8：英特尔估算的 AI 相关工作及 GPU 在服务器中的使用率



资料来源：英特尔官网，天风证券研究所

图 9：当前 GPU 在数据中心使用情况

	2016
英伟达2016年数据中心收入（百万美元）	830
深度学习占比%	59%
英伟达深度学习收入（百万美元）	**492**
2016年全球服务器出货量（千台）	11,104
服务器AI使用率%	7%
服务器AI使用量（千台）	777
AI服务器中GPU使用率%	3.4%
AI服务器中GPU使用量（千台）	26
GPU占全球服务器使用率%	0.24%
每台GPU服务器ASP（千美元）	**$19**
英伟达单片GPU价格（假设每台服务器搭载4片GPU）（千美元）	$4.7
英伟达单片GPU价格（假设每台服务器搭载8片GPU）（千美元）	$2.3

资料来源：Gartner，天风证券研究所预测

对数据中心 GPU 市场进行测算，英伟达表示目前数据中心内所有 GPU 使用中，约 60%用于深度学习/AI 相关工作，所以 2016 年全球数据中心新增 GPU 使用量约为 4.5 万台。假设英伟达占据 98%的市场份额，单台服务器 ASP 约为 1.9 万美元/台；AMD 基本上 2016Q4 才进入云计算市场，且单台服务器 ASP 也要低于英伟达（约 1.9 万美元/台），以每台 6 块 GPU，GPU 单价 2000 美元/片计，假设 AMD 的 ASP 为 1.2 万美元。

图 10：全球服务器 GPU 市场估计

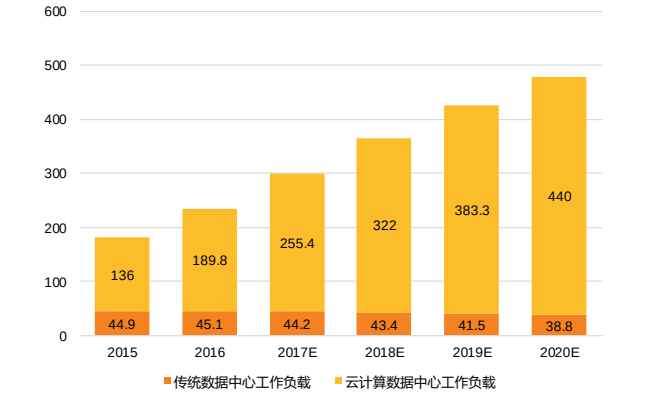
	2013	2014	2015	2016	2017E	2018E	2019E	2020E
全球服务器出货量（千）	9,887	10,091	11,091	11,104	11,881	12,594	13,123	13,648
增长率%		*2%*	*10%*	*0%*	*7%*	*6%*	*4%*	*4%*
服务器AI使用率%				7.0%	10.0%	12.5%	14.5%	16.0%
服务器AI使用量（千台）				777	1,188	1,574	1,903	2,184
AI服务器中GPU使用率%				3.4%	5.1%	6.3%	8.0%	9.3%
AI服务器中GPU使用量（千台）				26	60	99	152	203
GPU占全球服务器使用率%				**0.24%**	**0.51%**	**0.85%**	**1.16%**	**1.49%**
英伟达数据中心深度学习使用率%				59%	64%	68%	70%	72%
数据中心GPU使用量（千台）				44.6	93.9	145.8	217.5	282.0
英伟达市场份额%				98%	98%	96%	94%	92%
数据中心英伟达GPU使用量（千台）				43.7	92.1	140.0	204.4	259.5
每台英伟达GPU服务器ASP（千美元）				19	20	20	20	20
英伟达数据中心收入（百万美元）				**830**	**1,843**	**2,800**	**4,088**	**5,190**
增长率%					*122%*	*52%*	*46%*	*27%*
AMD市场份额%				2%	2%	4%	6%	8%
数据中心AMD GPU使用量（千台）				.9	1.8	5.8	13.0	22.6
每台AMD GPU服务器ASP（千美元）				12	12	12	12	12
AMD数据中心GPU收入（百万美元）				**11**	**21**	**70**	**157**	**271**
增长率%					*100%*	*227%*	*124%*	*73%*
GPU数据中心市场空间（百万美元）				**840**	**1,864**	**2,870**	**4,245**	**5,460**
增长率%					*122%*	*54%*	*48%*	*29%*

资料来源：公司财报，Gartner，英特尔，天风证券研究所预测

我们再从全球服务器工作负载维度来考量英伟达 GPU 的市场空间。我们把加速计算的范畴广义概括为 AI、深度学习和推理等计算元素，应用层面包括搜索、社交网络、流媒体视频 ERP、数据库大数据分析、IoT 和企业协作等商业应用，我们也可以看到整个数据中心市场正向着计算量需求扩张的方向前进，并为英伟达带来巨大的市场空间。

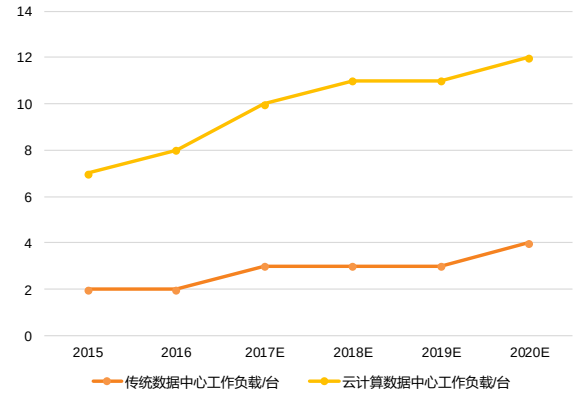
首先，根据南加大学者 Martin Hilbert 的研究表明，从上世纪 80 年代开始，全球服务器计算能力的增长保持着 CAGR 约 50%的速度。但是数据中心整体开支并没有表现明显扩张，我们看到这是摩尔定律的不断进步，数据中心逐渐向标准的 x86 构架服务器转变的过程中（根据 IDC 估计，x86 服务器使用率已经达到了 82%），以及硬件资源虚拟化带来的隔离性、可扩展性、安全性、资源可充分利用性，我们将看到云计算数据中心成为不可逆转的趋势。传统数据中心的工作负载呈平缓的形态（Cisco 估计 CAGR 为-3%），云计算数据中心的工作负载将占据绝大部分（Cisco 估计 CAGR 为+26%）。同时云计算服务中心每台服务器的工作负载效率也明显高于传统数据中心（Cisco 估计为 3x）。

图 11：全球数据中心工作负载变化



资料来源：Cisco，天风证券研究所

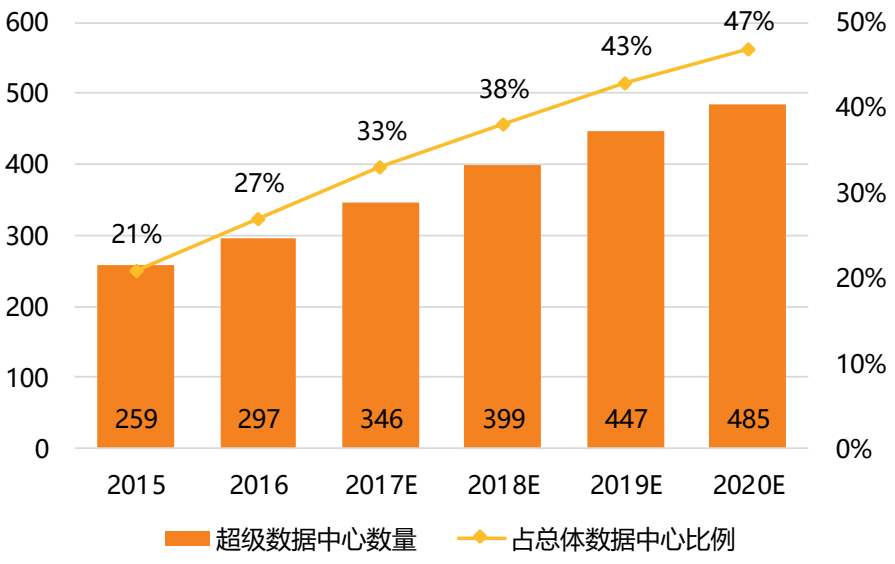
图 12：云计算数据中心的工作负载效率明显高于传统数据中心



资料来源：Cisco，天风证券研究所

此外，我们认为超级数据中心也越来越依赖 GPU 来更快地处理高要求的工作负载，成为 GPU 的需求主力。根据 Cisco 预测，超级数据中心在全球数据中心的比例将从 2015 年的 21%翻倍到 2020 年 47%，CAGR 为 13%。（Cisco 定义超级数据中心：PaaS 年收入超过 10 亿美元，或 SaaS 年收入超过 20 亿美元，或应用于搜索、社交网络的年收入超过 40 亿美元，或应用于电商、电子支付的年收入超过 80 亿美元）

图 13：全球超级数据中心数量



资料来源：Cisco，天风证券研究所

2.1.2.1. Volta 更新迸发澎湃动力

英伟达在今年 5 月 15 日举行的 GTC 开发者大会上，正式发布 Volta 构架 GPU。作为 Pascal 构架之后的全新构架，今后 2-3 年的 GPU 产品线，全都要由 Volta 供血。本次发布的 Volta 构架 GPU Telsa V100，采用台积电 12 纳米工艺，在 815 平方毫米面积的硅片上集成了 210 亿个晶体管，5120 个 CUDA 核心，其单精度浮点运算性能达到 15 TFLOP/s，双精度浮点运算性能达到 7.5 TFLOP/s。

图 14：英伟达人工智能布局平台

资料来源：英伟达官网，天风证券研究所

作为一款面向深度学习专门设计的 GPU，Tesla V100 加入了全新的张量运算指令 Tensor Core，这是一个针对深度学习张量运算专门优化的 4*4 矩阵处理阵列，将 Tesla V100 的张量运算能力提升到 120 TFLOPS。

图 15：英伟达发布 Volta 构架 GPU Tesla V100

资料来源：英伟达官网，天风证券研究所

同时 V100 加入了 TensorRT 神经网络推理编译器，更好地面向深度学习下游推理阶段的计算需求，更快速地优化、验证、部署训练好的神经网络。也是从此次 V100 开始，英伟达将注意力扩展到深度学习推理端的需求上，对比前代 Pascal 构架，V100 在 HPC 的常规浮点运算速度提高 1.5 倍，深度学习训练速度提高了 12 倍，推理速度提高 6 倍。

图 16：V100 对比前代 P100 的性能提升明显

资料来源：英伟达官网，天风证券研究所

此外英伟达更新了基于 Tesla V100 的高性能计算机 DGX 产品线。升级后的 DGX-1V 搭载 8 块 Tesla V100，运算能力达到 960 Tensor TFLOPS，黄仁勋表示相当于把 400 个服务器装到一个盒子里。英伟达还发布了个人版 DGX——DGX Station，内置四块 Tesla V100，水冷降温，运算能力 480 Tensor TFLOPS。另外英伟达与微软合作开发了公有云服务器 HGX-1 超大规模 GPU 加速器，将更好地适应于基于云端的人工智能计算需求。

图 17：英伟达 DGX 产品线更新：DGX-1V、DGX Station 以及公有云服务器 HGX-1

资料来源：英伟达官网，天风证券研究所

图 18：英伟达历代 Tesla GPU 性能对比

Tesla Product	Tesla K40	Tesla M40	Tesla P100	Tesla V100
GPU	GK180 (Kepler)	GM200 (Maxwell)	GP100 (Pascal)	GV100 (Volta)
SMs	15	24	56	80
TPCs	15	24	28	40
FP32 Cores / SM	192	128	64	64
FP32 Cores / GPU	2880	3072	3584	5120
FP64 Cores / SM	64	4	32	32
FP64 Cores / GPU	960	96	1792	2560
Tensor Cores / SM	NA	NA	NA	8
Tensor Cores / GPU	NA	NA	NA	640
GPU Boost Clock	810/875 MHz	1114 MHz	1480 MHz	1455 MHz
Peak FP32 TFLOP/s*	5.04	6.8	10.6	15
Peak FP64 TFLOP/s*	1.68	0.21	5.3	7.5
Peak Tensor Core TFLOP/s*	NA	NA	NA	120
Texture Units	240	192	224	320
Memory Interface	384-bit GDDR5	384-bit GDDR5	4096-bit HBM2	4096-bit HBM2
Memory Size	Up to 12 GB	Up to 24 GB	16 GB	16 GB
L2 Cache Size	1536 KB	3072 KB	4096 KB	6144 KB
Shared Memory Size / SM	16 KB/32 KB/48 KB	96 KB	64 KB	Configurable up to 96 KB
Register File Size / SM	256 KB	256 KB	256 KB	256KB
Register File Size / GPU	3840 KB	6144 KB	14336 KB	20480 KB
TDP	235 Watts	250 Watts	300 Watts	300 Watts
Transistors	7.1 billion	8 billion	15.3 billion	21.1 billion
GPU Die Size	551 mm²	601 mm²	610 mm²	815 mm²
Manufacturing Process	28 nm	28 nm	16 nm FinFET+	12 nm FFN

资料来源：英伟达官网，天风证券研究所

2.1.3. 自动驾驶开启「黄金十年」

我们认为，以 2020 年为界，全球将开启无人驾驶「黄金十年」。

L3 半自动驾驶水平以上的行业发展，需要整个汽车行业供应商关系的重组和整合。包括：

1、形成“车企+供应商+芯片巨头+打车软件+物流公司”的格局；

2、共享经济下的租车、打车以及商业货运物流领域会最快落地得到应用；

3、L4 相对比 L1、L2，单车系统零部件支出会增长 470%，从 545 美元升至 3100 美元/车。

我们认为全球 L1-L5 智能驾驶市场的渗透率会在接下来 5 年内依靠 ADAS 市场的高速发展而处于高速渗透期，然后伴随半无人驾驶的普及进入稳速增长期，在来到 2025 年无人驾驶放量阶段后，依赖全产业链的配合而进入市场成熟期。我们预测到 2030 年，全球 L4/5 级别的自动驾驶车辆渗透率将达到 15%，单车应用成本的显著提升之外，从 L1-L4 级别的智能驾驶功能全面渗透为汽车产业带来全面的市场机会。英伟达指引到 2025 年自动驾驶市场空间可达 80 亿美元，包括 500 万辆 L4/5 自动驾驶汽车的 50 亿美元空间，1500 万辆 L2/3 等级 20 亿美元空间，以及 500 万辆 AI Co-Pilot 的 10 亿美元。

1、2020 年开启无人驾驶「黄金十年」

我们预测到 2030 年，全球 L4/5 级别的自动驾驶车辆渗透率将达到 15%，单车应用成本的显著提升之外，从 L1-L4 级别的智能驾驶功能全面渗透为汽车产业带来全面的市场机会。

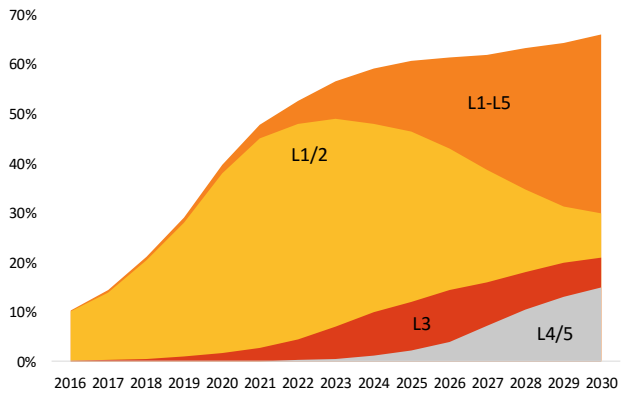
针对 L1/2 的 ADAS 功能，从 2017 年 15%左右的渗透率开始，到 2020 年看到加速向上的渗透率提升，并在 2025 年达到 50%左右的顶峰。

L3 的半自动驾驶级别，今年已经出现产品落地。我们认为渗透率将从现在开始显著提高，并保持稳定上升的态势。

L4/5 的全自动驾驶级别，实质性的 1%市场渗透率需要到 2024 年左右才能看到，配套政策法规的完善会给完全自动驾驶带来有力促进。不过完全无人驾驶还是会先从利基市场开始获取消费者，但是会拥有较半无人驾驶更为有力的渗透率提升速度。

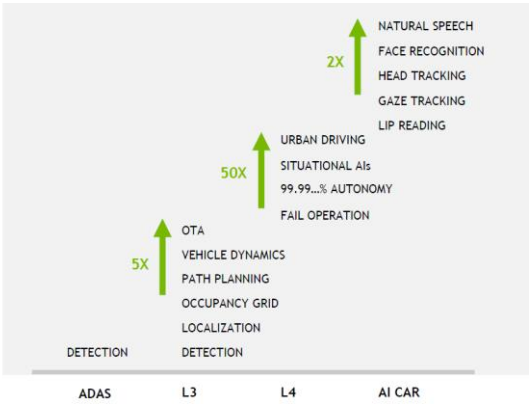
综上，我们看到全球 L1-L5 智能驾驶市场的渗透率会在接下来 5 年内依靠 ADAS 市场的高速发展而处于高速渗透期，然后伴随半无人驾驶的普及进入稳速增长期，在来到 2025 年无人驾驶放量阶段后，依赖全产业链的配合而进入市场成熟期。

图 19：全球自动驾驶 L1-L5 渗透率预测



资料来源：IHS，天风证券研究所预测

图 20：英伟达公布的自动驾驶算力提升路径



资料来源：公司官网，天风证券研究所

英伟达认为从 ADAS 提升到 L3 半自动驾驶所需的计算难度会提升 5 倍，而关键的 L3 向 L4 提升需要 50 倍，从 L4 提升到 L5 则需要 2 倍。因此，汽车电子化和智能化的方向将持续提高科技类公司在汽车产业链内的重要程度，我们看到三星收购哈曼，高通收购 NXP，到现在英特尔收购 Mobileye，我们最为看好的产业方向就是掌握关键技术和客户资源的技术

公司，在上游汽车电子方向上发力，然后是下游驾驶服务软件层面布局，包括地图数据、用户出行数据等等，另外就是车联网对于基础架构建设的需求。

图 21：全球自动驾驶渗透率预测

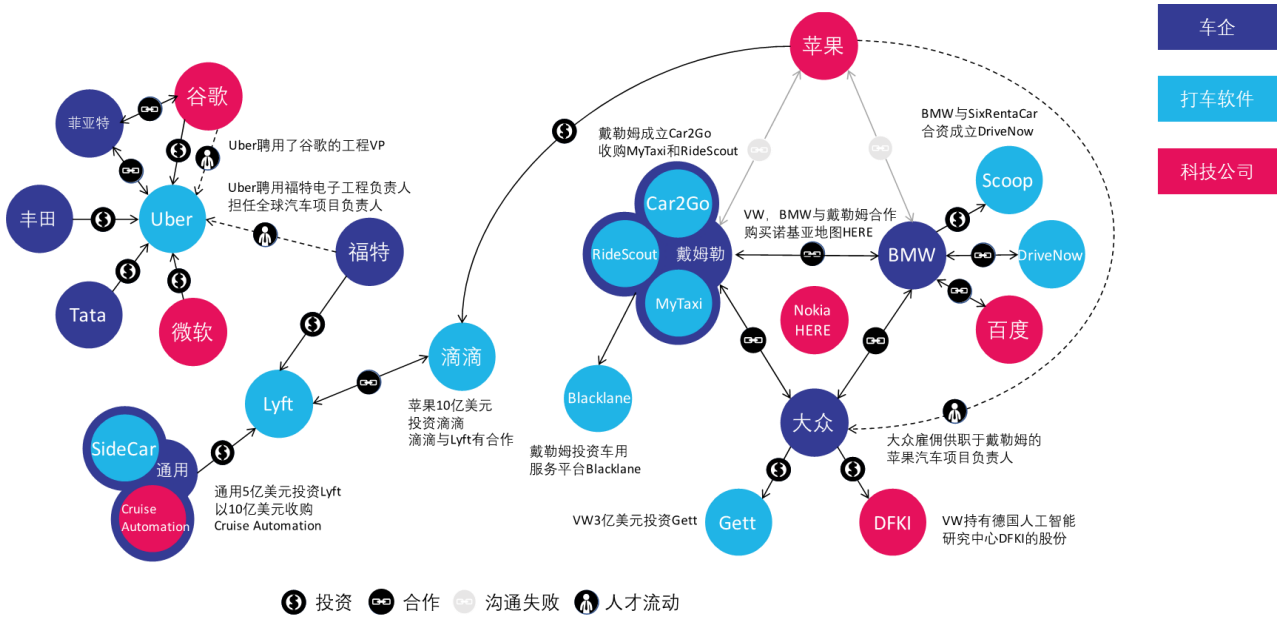
	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030
ADAS功能 (L1/2)															
全球轻型车产量（千辆）	92125	93512	94835	97480	99294	101089	103093	105165	107423	109572	111763	113998	115708	117444	119206
渗透率	**10%**	**14%**	**21%**	**30%**	**40%**	**45%**	**48%**	**49%**	**48%**	**47%**	**43%**	**39%**	**35%**	**31%**	**30%**
半自动驾驶 (L3)															
全球轻型车产量（千辆）	92125	93512	94835	97480	99294	101089	103093	105165	107423	109572	111763	113998	115708	117444	119206
渗透率	**0.2%**	**0.4%**	**0.6%**	**1.0%**	**1.8%**	**2.8%**	**4.4%**	**7.0%**	**10.0%**	**12.0%**	**14.5%**	**16.0%**	**18.0%**	**20.0%**	**21.0%**
自动驾驶 (L4/5)															
全球轻型车产量（千辆）	92125	93512	94835	97480	99294	101089	103093	105165	107423	109572	111763	113998	115708	117444	119206
渗透率						**0.0%**	**0.3%**	**0.6%**	**1.2%**	**2.3%**	**4.0%**	**7.2%**	**10.5%**	**13.0%**	**15.0%**
L1-L5总渗透率	**10%**	**14%**	**21%**	**31%**	**41%**	**48%**	**53%**	**57%**	**59%**	**61%**	**62%**	**62%**	**63%**	**64%**	**66%**
L1-L5总车辆数	9581	13466	20010	30219	41008	48320	54330	59523	63595	66620	68734	70565	73278	75572	78676

资料来源：IHS，天风证券研究所预测

2、「车企+供应商+芯片巨头+打车软件+物流公司」新格局

我们认为，虽然学术研发水平逐步进入 L4 高度自动驾驶，但在行业落地到 L3 等级以上，还需要整个汽车行业供应商关系的资源重组和整合。自动驾驶领域正在向“车企+供应商+芯片巨头+打车软件+物流公司”的组合格局发展。

图 22：「车企+供应商+芯片巨头+打车软件」新格局



资料来源：《中国人工智能产业发展报告》，天风证券研究所整理

3、共享经济下的租车、打车以及商业货运物流领域会最快落地得到应用

共享经济我们常说的自动驾驶出租车(self-driving cabs)的共享经济概念。Uber 一直以来的打算就是未来的叫车服务将不再需要人类驾驶员，Uber 完全控制驾驶系统以及自有车辆，这样可以极大地节约人力成本与提高车辆的使用率，从而减低路面的拥堵。

通用与 Lyft 以及 Cruise Automation 的合作就被看作是通用希望未来人们使用 Lyft 打车时搭乘无人交通工具。通用表示非常看好未来自动驾驶出租车市场的巨大发展空间，他们也认为无人驾驶技术最先应用的地方就是拼车及租车服务。

Elon Musk 的大师计划第二部中也看到了共享经济的概念，由于当前大部分私家车每天的使用率仅为 5-10%，Musk 希望自动驾驶的特斯拉能加入特斯拉网络(Tesla Network)中。这样的共享经济将大部分的价值归还到车主手中，降低特斯拉的使用成本。

谷歌最近也有一个“无人车接送旅客决定接送地点”的专利被披露出来(专利号 US 20160370194)，乘客提供接送或目的地，车辆操纵自身到该位置，整个驾驶过程处于完全自动的驾驶模式。

图 23：打车软件紧抓数据，把握共享经济

资料来源：MarketWatch，公司财报，天风证券研究所

商业货运领域：MIT Tech Review 预测 2017 年全球十大新兴科技趋势之一，就是在高速公路上长途行驶的自动驾驶货车很有可能会成为最快落地的无人驾驶项目，并对百万计的货车司机职位产生冲击。

去年 10 月，Uber 旗下无人驾驶货车公司 Otto 完成首次商用无人卡车货运项目，高速公路上货车司机全程没有碰过方向盘。Otto 的硬件设备包括 3 个激光雷达，1 个普通雷达以及 1 个高精度摄像头，可适用于任何具备自动换挡功能的卡车。无人卡车的需求来自现实的司机缺口，美国卡车货运量占到了美国本土货运量的 70%，而到 2024 年司机缺口会提升到 175000 人。包括 Uber、Volvo、Daimler 和 Peterbilt 都在研发相应技术。

图 24：商业用车的无人驾驶场景最快落地

Tesla于11月揭晓电动半挂卡车，我们认为会进一步推动汽车业S.E.A.变革发展，并为商用车无人驾驶带来新风

资料来源：Otto、公司官网，天风证券研究所

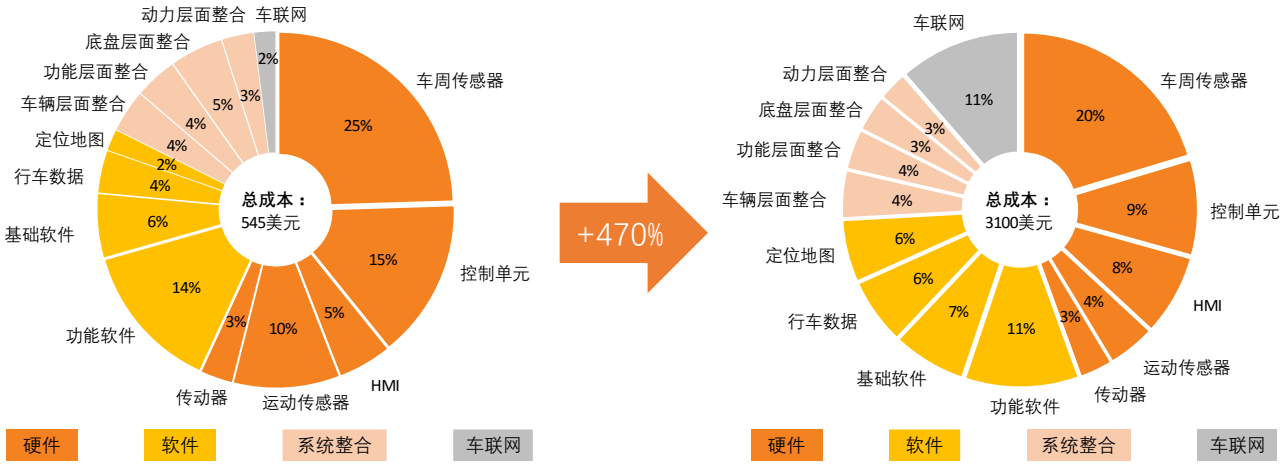
4、L4 相对比 L1、L2，单车系统零部件支出会增长 470%，从 545 美元升至 3100 美元/车

我们认为从 L1/2 升级到 L4，单车系统零部件支出会增长 470%，从 545 美元升至 3100 美元/车。在车辆的系统零部件上呈两大趋势：

1、车载摄像头的数量明显提升：从 L1/2 的每车 2 个提升到 L3/4 的每车 8-12 个；

2、激光雷达的成本瓶颈得以解决：激光雷达技术的发展将打破自动驾驶在 L3/4 级别的成本瓶颈。目前谷歌 Waymo 通过自行打造全套传感器设备，将激光雷达成本从 8 万美元下降到约 7500 美元。我们认为实现可商业量产级别的激光雷达成本应在 1000 美元以下。

图 25：L1 到 L4 的单车零部件成本变化



资料来源：Strategy Engineers，天风证券研究所

2.1.4. 英伟达成自动驾驶先行军，接棒数据中心

目前，无人驾驶上游系统解决方案逐渐形成英伟达与英特尔-Mobileye 联盟两大竞争者。英伟达方面此前汽车业务主要集中在汽车显示屏和影音系统，我们一直强调自动驾驶业务当前还处在合作布局阶段。而随着以奥迪 A8 为首的高端车型在明年开始逐步配置 L3 基本半自动驾驶，市场放量会对汽车业务带来明显营收贡献。

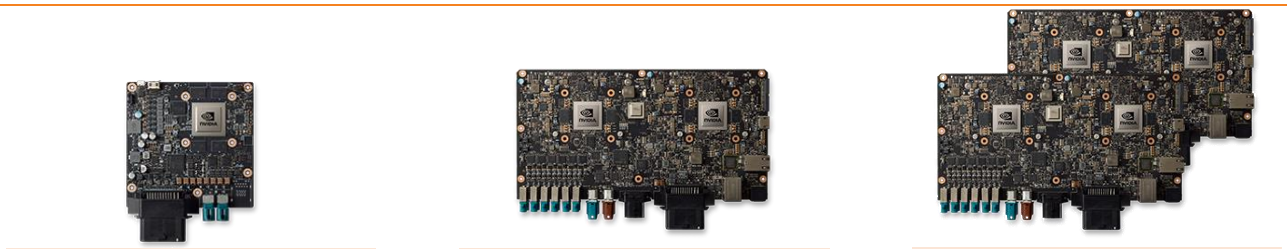
我们认为英伟达通过 L2-L5 统一的 Drive PX 底层平台以及开放的上层传感器布局和自定义模块为车企和 Tier 1 留出充足的可定制以及溢价空间。英伟达强调车载电脑的本地计算可靠性，在网络带宽不足或信号不好的情况下，本地硬件提供足够的算力、安全冗余和低延迟，同时保障行车数据的本地储存的隐私性。不过目前英伟达在软件层面产品有限，未来发展有待驾驶决策软件算法落地，以及 Drive PX 平台成本的降低。

在今年 1 月的 CES 大会上，英伟达发布无人驾驶的整体布局——从车载超级电脑平台以及人工智能驾驶系统发力，展示了包括 Xavier 下一代车载超级电脑，使用 DRIVE PX 2 车载电脑平台的 BB8 无人驾驶汽车，包含四大感知功能的人工智能协同驾驶系统 AI Co-Pilot，同时还有基于 Drive PX 2 和 Tesla GPU 云数据的端到端高清制图产品。

10 月英伟达在德国慕尼黑的 GTC Europe 大会上，发布了面向完全自动驾驶 L5 级别的新一代 Drive PX 人工智能车载计算平台 Pegasus。Pegasus 平台基于两块 Xavier 片上系统以及两块下一代 GPU，每秒操作超过 320 万亿次，达到 Drive PX 2 的 10 倍以上。

我们认为，目前英伟达自动驾驶计算平台已经拥有行业最强大的计算性能，对比竞争对手 Mobileye 的 EyeQ 芯片，预计 2020 年推出算力为 15 万亿次的 EyeQ5。英伟达在硬件层面算力和研发节奏上成为当仁不让的先行军。

图 26：Drive PX 2 平台的三款芯片



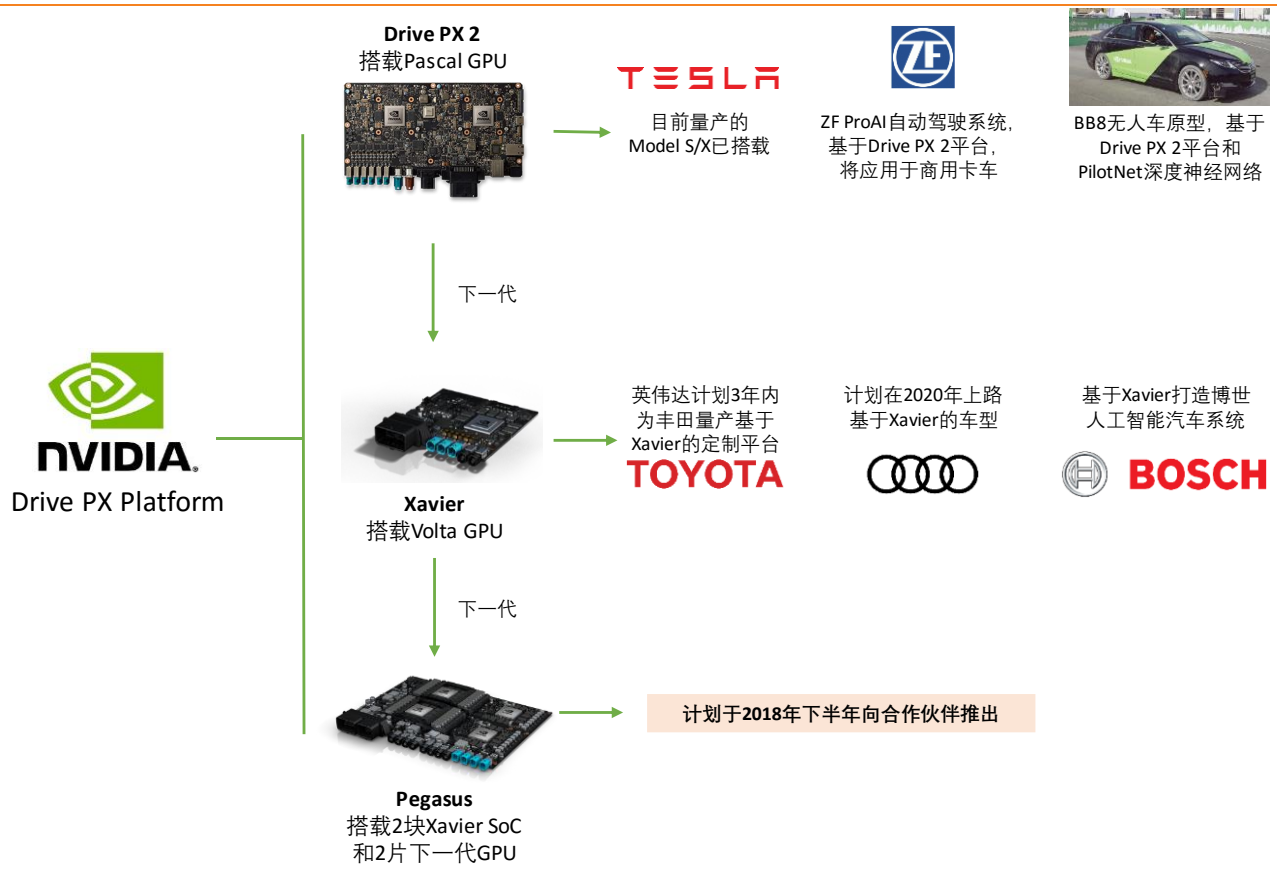
AUTOCRUISE 自动巡航级别：高速公路自动巡航、高精度地图绘制，2016Q4推出

AUTOCHAUFFEUR 自动私人司机级别：双独立Pascal独显、两个SoCs实现点对点的自动驾驶。

FULLY AUTONOMOUS 完全自动驾驶级别：多块DRIVE PX 2主板整合达到完全自动驾驶算力

资料来源：公司官网，天风证券研究所

图 27：英伟达 Drive PX 车载计算平台情况



资料来源：Geekcar，天风证券研究所

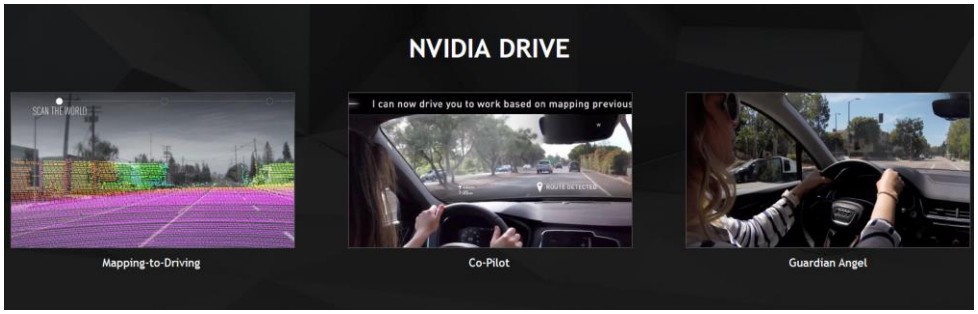
图 28：英伟达三代自动驾驶平台性能比较

NVIDIA Drive PX Specification Comparison						
	Drive CX	Drive PX	Drive PX 2 (AutoCruise)	Drive PX 2 (AutoChauffeur)	Xavier AI Car Supercomputer	Drive PX Pegasus
Generation	First		Second		Third	
Introduced	January 2015		September 2016	January 2016	January 2017	October 2017
Computing	1x Tegra X1	2x Tegra X1	1x Tegra X2 (Parker)	2x Tegra X2 (Parker)+ 2x Pascal GPU	1x Tegra Xavier	2x Xavier
CPU Cores	4x Cortex A574x Cortex A53	8x Cortex A578x Cortex A53	2x Denver4x Cortex A57	4x Denver8x Cortex A57	8x Custom ARM64	16x NVIDIA Custom ARM
GPU	2 SMM Maxwell256 CUDA cores	4 SMM Maxwell512 CUDA cores	1x Parker GPGPU(1x 2 SM Pascal, 256 CUDA cores)	2x Parker GPGPU (2x 2 SM Pascal, 512 CUDA cores) + 2x dedicated MXM modules	1x Volta GPGPU512 CUDA Cores	2x Xavier Volta iGPU 2x post-Volta dGPUs
TDP			(Parker SoC only: 10 W)	250W	30W	500W

资料来源：英伟达官网，wccftech，天风证券研究所

英伟达在自动驾驶汽车产业链内的定位是提供一个自动驾驶基础运算平台——也就是一台车载超级电脑。无论是与车企直接合作还是 Tier 1 合作，都会通过统一的底层构架以及开放的上层传感器布局和自定义模块的做法为 OEM 和 Tier 1 留出充足的可选择以及溢价空间。Xavier 运算平台上，各个厂商可以加上自己的算法，并通过 DriveWorks SDK 来进一步开发不同的功能。在服务层面，英伟达为合作伙伴提供了 3 方面的应用：1、从感知到制图到驾驶策略的完整解决方案；2、包含四大感知功能的人工智能协同驾驶系统 AI Co-Pilot；3、对驾驶环境进行感知，辅助司机安全驾驶的 Guardian Angel。

图 29：英伟达自动驾驶服务

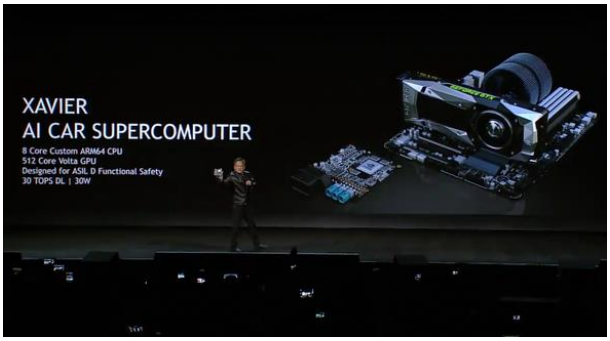


资料来源：英伟达官网，天风证券研究所

2.1.4.1. Xavier：“装进手提箱”的车载超级电脑

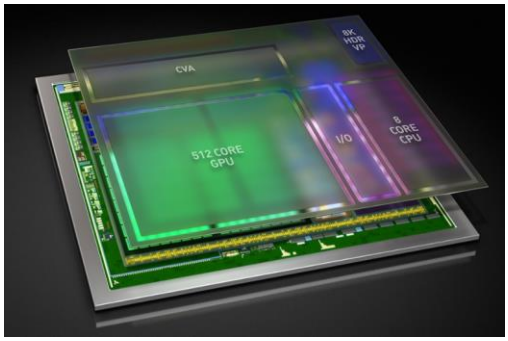
今年 1 月推出的 Xavier 片上系统由 8 核 ARM 64 位 CPU 和 512 核的 Volta 架构 GPU 构成，整合了开源计算机视觉加速器。Xavier 采用 16nm FinFET 制程，单 AI 处理器将取代 Drive PX 2 里的双移动 SoCs 和双独显的构架，保证了 1 TOPS/瓦的功耗比。Xavier 芯片本身为 ASIL C，但其模块可通过设计实现 ASIL D 安全功能。

图 30：英伟达 Xavier 下一代车载超级电脑



资料来源：公司官网，天风证券研究所

图 31：英伟达 Xavier 包含 8 核 CPU 和 512 核 GPU



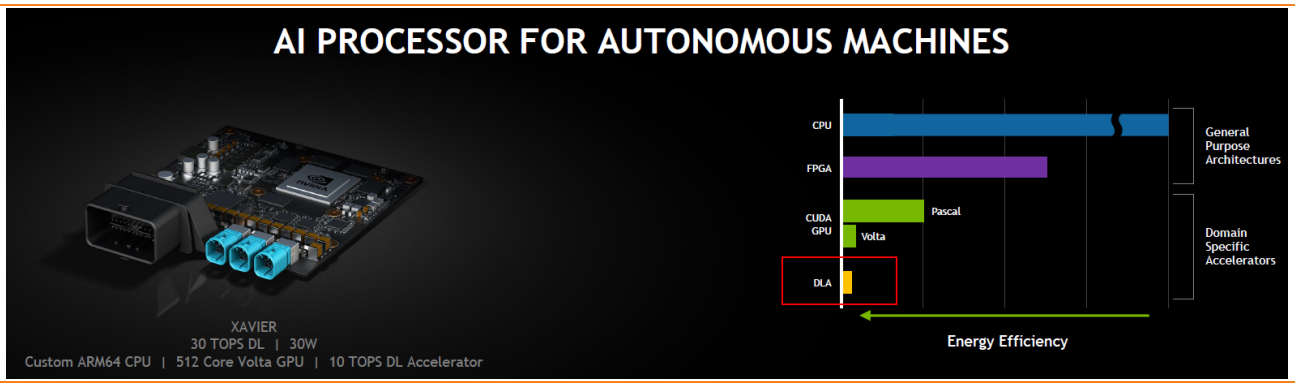
资料来源：公司官网，天风证券研究所

相较于去年 9 月首次披露时，Xavier 的算力从 20 TOPS 提高到 30 TOPS（万亿次运算/秒）性能且能耗低至 30 瓦。英伟达希望 Xavier 成为未来自动驾驶的主力军，特斯拉 Autopilot 2.0 硬件升级后，搭载的是英伟达 DRIVE PX 2 车载电脑平台，我们认为在往 Level 4/5 无人车阶段升级的过程中，特斯拉应该会进一步升级至 Xavier 平台。Xavier 同时具备 CPU 的单线程性能，CUDA 的并行加速能力，以及 DLA 的计算机视觉特殊功能，目前已经小批量试产，在今年开始 4Q 给包括车企、一级供应商、初创公司以及研究室进行无人驾驶研发。

2.1.4.2. 开源 DLA，加速自动驾驶研发生态

英伟达在 5 月的 GTC 大会上与丰田宣布合作，打造端到端的全栈开源深度学习平台，同时宣布将 Xavier DLA（深度学习加速器）开源，并在 9 月份正式发布，进一步吸引开发者进入开发生态圈。我们在前文提到，ASIC 仍然面临通用性弱、开发成本高企等局限，短期对 GPU 需求的影响十分有限。英伟达同时也在针对性提升 ASIC 领域的竞争力，包括开源 Xavier 自动驾驶系统中英伟达自己的推理端 TPU——硬件加速模块 DLA（深度学习加速器）。我们认为，Xavier 的发布将很好地契和自动驾驶参与者的研发节奏，公司汽车业务应该从明年开始看到显著的收入贡献。

图 32：英伟达 Xavier 下一代车载超级电脑，硬件加速模块 DLA 拥有最好的能效比

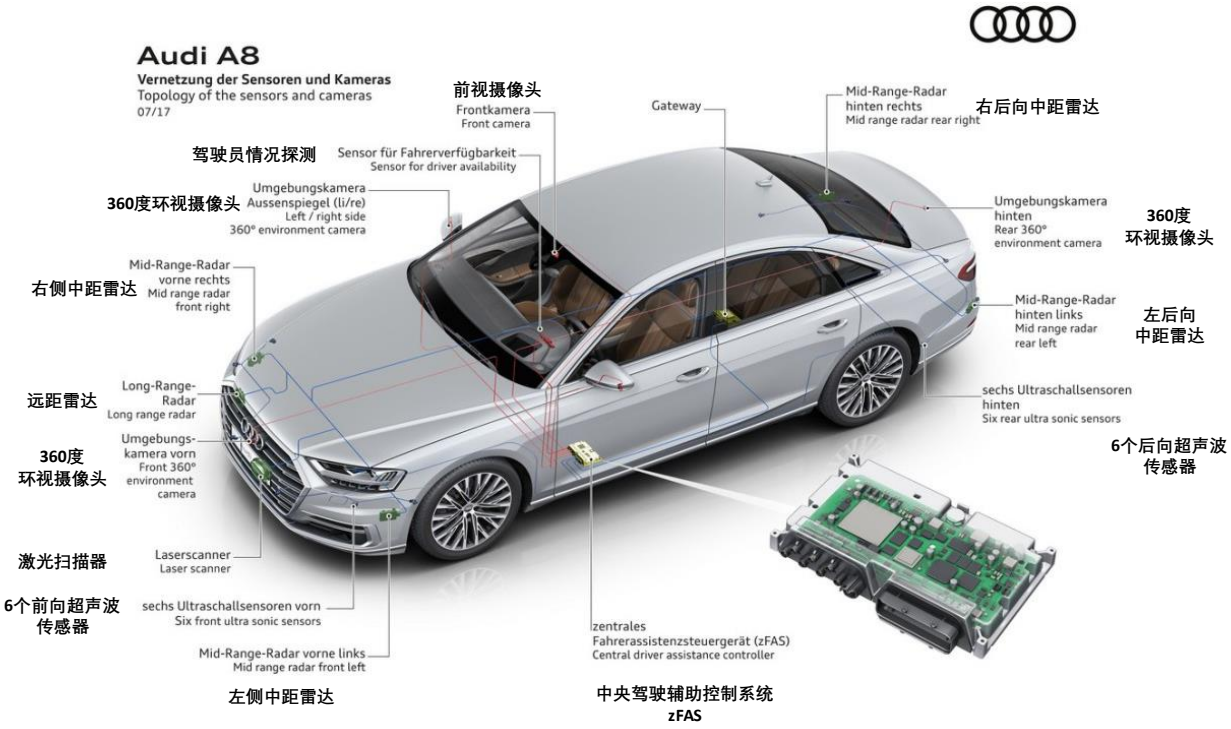


资料来源：公司官网，天风证券研究所

最近英伟达还宣布与大众在整个人工智能深度学习领域正式扩大合作伙伴关系。大众集团的人工智能野心不只在自动驾驶上，还囊括了在集团业务流程与移动出行服务领域的探索。大众与英伟达的合作重点会从优化城市交通入手，在人工智能的帮助下实现的人机协同智能工作。

英伟达长期合作伙伴奥迪也在日前推出了与英伟达、Mobileye、Delphi 等公司合作设计的全球首款搭载 L3 级自动驾驶的量产车——新一代 A8。新 A8 一共搭载了 4 个中距雷达，12 个超声波传感器，5 个车载摄像头，1 个激光扫描器，所有传感器数据都会传输至车辆核心的 zFAS 中央驾驶辅助控制系统。

图 33：新一代奥迪 A8 搭载了超过 22 个传感器设备



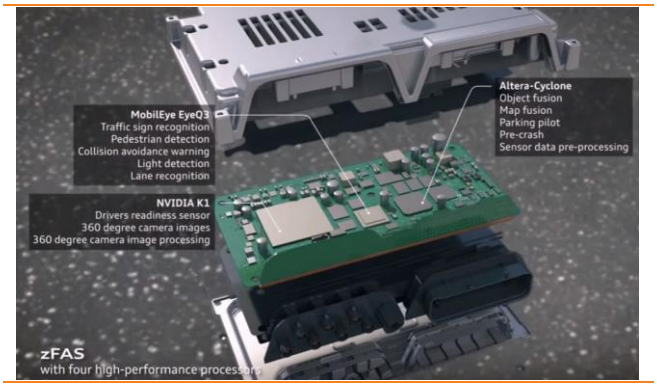
资料来源：奥迪发布会，天风证券研究所整理

zFAS 中央驾驶辅助控制系统使用了包括英伟达 Tegra K1 处理器，Mobileye 的 EyeQ3 芯片，Altera 的 Cyclone V 芯片，Infineon 的 Aurix 以及 TTTech 的通信模块等。核心系统 Traffic Jam Pilot 是目前市面上首款达到三级自动驾驶水平的拥堵导航功能，能够在 60 公里/小时时速下实现无人驾驶功能。

图 34：奥迪新 A8 的无人驾驶按钮

资料来源：公司发布会，天风证券研究所

图 35：奥迪的 zFAS 中央驾驶辅助控制系统



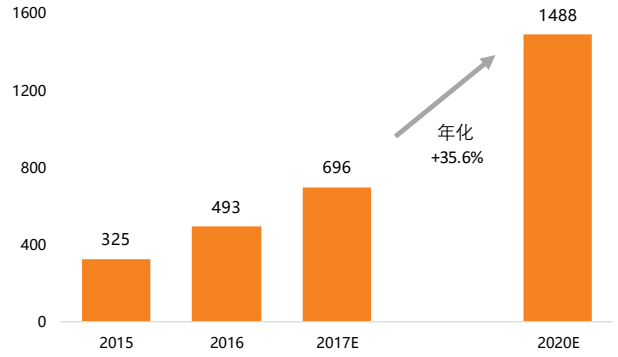
资料来源：auto connected car news，天风证券研究所

2.1.5. 游戏业务“现金马”扬鞭奔腾

英伟达游戏业务过去 5 年年化增速 25%，包含 ASP 年化增速 11%和销量年化增速 12%。依托高端 PC 游戏（3A 游戏）、VR 以及电子竞技热情以及用户基数升级周期，我们看好英伟达游戏业务的稳固增长，公司“现金马”扬鞭奔腾。

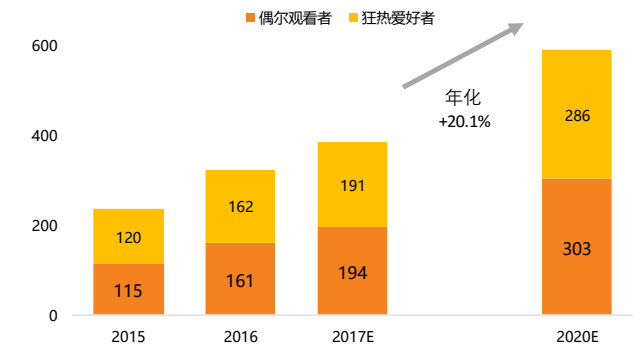
电竞游戏市场正进入活力迸发的新纪元，根据 NewZoo 的统计，2017 年全球电子竞技市场规模将达到 6.96 亿美元，并在 2020 成长值 15 亿美元，年化增速超过 35%。观众人数也在进一步增长，NewZoo 预计 2017 年电竞狂热爱好观众的人数约在 1.9 亿人，并在 2020 年增加至 2.86 亿人，这其中亚太地区观众占据 51%。我们认为中国电竞市场伴随赛事推动、直播兴起和移动电竞的流行，已经进入成熟发展阶段，在向泛娱乐化方向发展的过程中不断吸引游戏爱好者加入，刺激游戏电脑的购买和升级。

图 36：全球电子竞技市场收入（百万美元）



资料来源：NewZoo，天风证券研究所

图 37：全球电竞游戏观众人数（百万人）



资料来源：NewZoo，天风证券研究所

另外高端 PC 游戏（3A 游戏）以及 VR 游戏的纷至沓来，今年《绝地求生》打造的“大逃杀”模式成为继 MOBA 之后又一现象级的游戏模式，创造了 Steam 在线人数记录，最多 290 万玩家同时在线，其中近一半玩家来自中国。该游戏推荐显卡配置就要求英伟达 GTX 960/1060 或以上。另外今年下半年的 3A 游戏大作包括《最终幻想世界》、《使命召唤 14》、《命运 2》、《NBA 2K18》、《FIFA18》等。另外 VR 游戏尚处在用户习惯培养阶段，目前已经看到 VR 电竞赛事的尝试，同时传统电竞游戏通过 VR 直播搭上了 VR 概念，我们认为大量射击类、动作类 VR 游戏会凭借游戏性+社交性获得口碑效应。英伟达表示 81%的视频游戏用户对 VR 游戏都表达了兴趣，同时 VR 游戏对显卡的高性能需求会进一步拉动 ASP 的提升。

图 38：3A 游戏大作及现象级游戏频出刺激玩家升级电脑配置

资料来源：各公司官网，天风证券研究所

游戏的精细程度升级以及高清画质的追求，成为游戏玩家不断更新游戏显卡的动力，运行 1080P 游戏的显卡均价从 2014 年的 140 美元增长至 2017 年的 180 美元，接下来再叠加 VR 游戏对高性能显卡的巨大需求，我们认为英伟达游戏业务的稳固增长，作为公司“现金马”扬鞭奔腾。英伟达最近发布新的 Pascal 架构高端显卡 1070 Ti，借今年最后两个月的销售旺季进一步推动游戏业务在 Q4 的增长。同时预计明年上半年才会发布 Volta 架构游戏显卡。公司 CEO Jensen 就表示，希望视频游戏会成为全球最大的单一娱乐市场，同时 VR 虚拟现实能实现各种场景，游戏的选择和可能性被极大的拓展，因此视频游戏在未来数十年中都会有极大的增长动力。

图 39：英伟达新发游戏显卡 1070 Ti 性能比较

	GTX 1080	GTX 1070 Ti	GTX 1070
CUDA Cores	2560	2432	1920
Texture Units	160	152	120
ROPs	64	64	64
Core Clock	1607MHz	1607MHz	1506MHz
Boost Clock	1733MHz	1683MHz	1683MHz
Memory Clock	10Gbps GDDR5X	8Gbps GDDR5	8Gbps GDDR5
Memory Bus Width	256-bit	256-bit	256-bit
VRAM	8GB	8GB	8GB
FP64	1/32	1/32	1/32
TDP	180W	180W	150W
GPU	GP104	GP104	GP104
Transistor Count	7.2B	7.2B	7.2B
Manufacturing Process	TSMC 16nm	TSMC 16nm	TSMC 16nm
Launch Date	05/27/2016	11/02/2017	06/10/2016
Launch Price	MSRP: $599 Founders: $699	MSRP: $449 Founders: $449	MSRP: $379 Founders: $449

资料来源：英伟达官网，天风证券研究所

2.1.6. 估值：重申“买入”，TP 上调至 280 美元

英伟达 FY18Q3 营收同比增长 32%至 26.4 亿美元，Non-GAAP EPS 同比涨 41%至 1.33 美元，高于预期的 23.6 亿美元和 0.94 美元。游戏业务增速稳定，收入同比增长 25%达 15.6 亿美元，高于预期的 13.1 亿美元，营收占比提升至 59%。数据中心业务随着 Volta GPU 的正式放量实现营收 5.01 亿美元，同比涨 101%，环比涨 20%，高于预期的 4.74 亿美元；汽车业务营收 1.44 亿美元，同比涨 13%，略低于预期的 1.49 亿美元。Q2 整体毛利率 59.7%，高于指引的 58.8%。FY18Q4 指引营收中位数 26.5 亿美元，高于预期的 24.4 亿美元。对应全年营收指引将达 94.5 亿美元，实现 37%增长。

数据中心业务 2020 年前有望翻 4 倍：我们一直强调数据中心市场巨大空间。深度学习上游训练端由 GPU 主导并基本为英伟达所垄断。英伟达数据中心业务去年收入 8.3 亿美元，

我们预测今年收入将达 18.4 亿美元，全年实现超 120%增长。随着亚马逊宣布部署英伟达 V100 GPU 实例，包括微软、谷歌、Oracle、国内 BAT，以及所有 OEM 厂商的合作推进都在稳步进行，Volta 的放量将会进一步对 Q4 和明年的数据中心业务增长产生贡献。

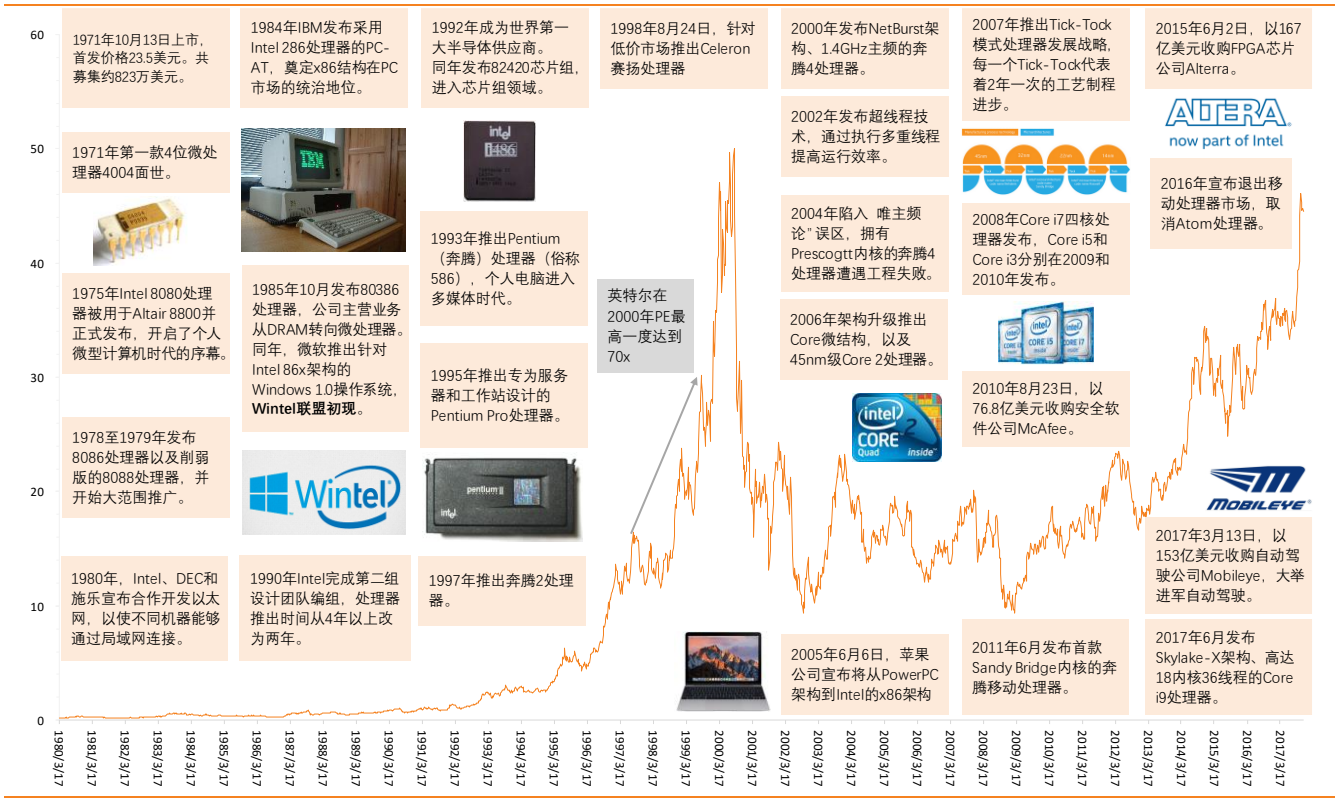
2016 年全球新增服务器 GPU 渗透率仅为 0.24%。我们预计英伟达数据中心业务在 2020 年前将达 40 亿美元，全球服务器 GPU 渗透率也将达 4 倍以上增长。CEO Jensen 强调超级数据中心和超级计算机 HPC 等市场的 GPU 加速对于包括物流行业 DHL、UPS、FedEx，交通行业滴滴、Uber、Lyft 等公司，都会从机器学习公司进化成 AI 公司，带动整个行业的革新。

自动驾驶广泛布局，3-5 年期长期驱动：当前汽车业务还处在合作布局阶段，自动驾驶领域正在向“车企+供应商+芯片巨头+打车软件”的组合格局发展，我们认为，随着针对 L5 完全自动驾驶的 Pegasus 平台的发布，英伟达自动驾驶计算平台已经拥有行业最强大的计算性能，明年开始应该可以看到英伟达的汽车业务逐渐提升无人驾驶业务的占比，并在无人驾驶出租车等的共享经济市场看到商业化雏形。广泛布局把握先发，虽然英伟达当前在软件层面产品有限，但我们认为随着无人驾驶产业普及，成本逐渐下降，完整的解决方案也会随之落地。

以游戏业务为现金马，三驾马车齐发力：游戏业务依托高端 PC 游戏、VR 以及电子竞技热情以及用户基数升级周期带来游戏业务的稳固增长。

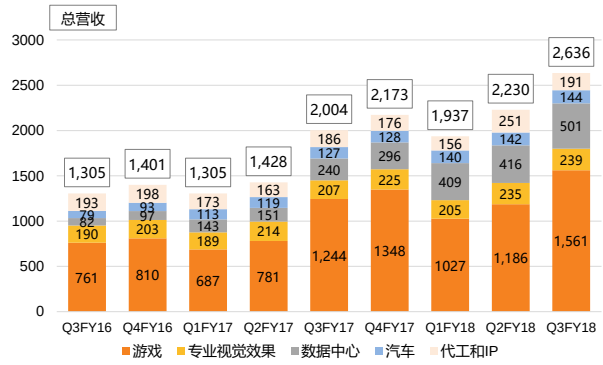
我们认为英伟达将持续巩固 GPU 市场龙头地位，保持现有业务充沛活力的同时自上而下推动 AI 浪潮。我们回顾英特尔的发展历程，在个人电脑 PC 兴起的九十年代，英特尔的股价也从 1992 年初的约 2 美元涨至 2000 年的约 50 美元，“十年金股”实现 25 倍的涨幅，对应 PE 在互联网泡沫之前都保持在 50x 以下的水平，但到 2000 年最高一度达到 70x。目前英伟达股价对应 PE 为 55x，2015 年以来的彭博预测下年 PE 也一直保持在 40x 以下。从 2014 年底约 20 美元涨至当前的 215 美元，三年涨幅约 10 倍。如果能够达到英特尔当时 70x 水平，市值会实现 2400 亿美元突破。若以“十年金股”为界，1300 亿美元市值还仅是 AI 立夏开端。**我们给予公司 2018/19 年 EPS 分别为 5.42、6.61 美元，对应 2018/19 年 PE 52/42x，目标价从 250 上调至 280 美元，重申“买入”评级。**

图 40：英特尔历史大事件，在 2000 年 PE 最高一度达到 70x



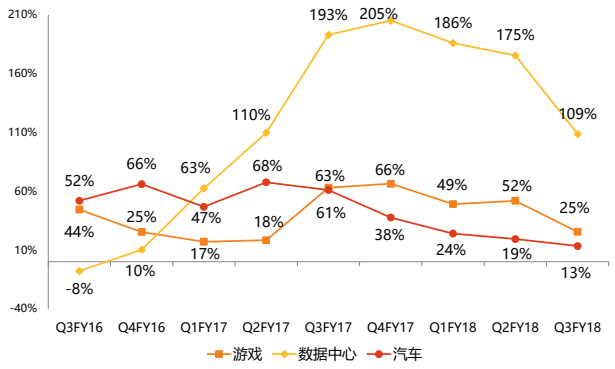
资料来源：Yahoo Finance，天风证券研究所整理，数据截止至 2017 年 11 月 27 日，采用调整后收盘价

图 41：英伟达各项业务营收比较（百万美元）



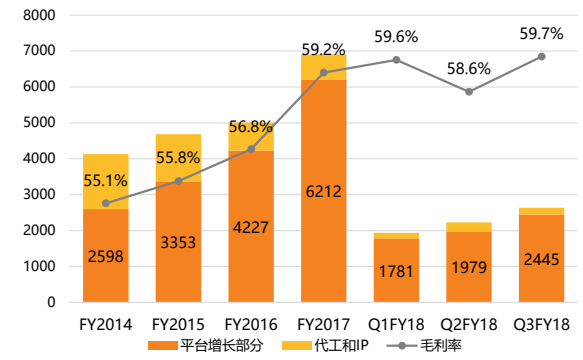
资料来源：公司财报，天风证券研究所整理

图 42：英伟达游戏、数据中心、汽车三块业务同比增速



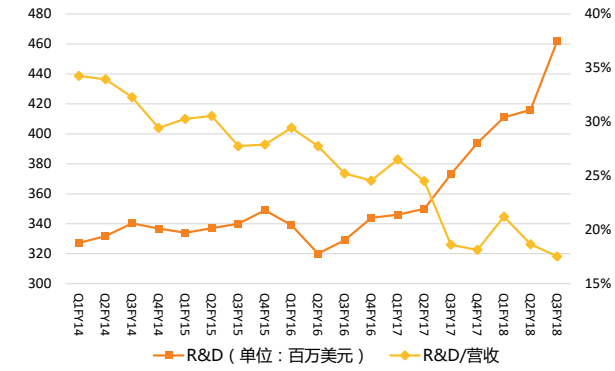
资料来源：公司财报，天风证券研究所整理

图 43：英伟达毛利率逐年增长



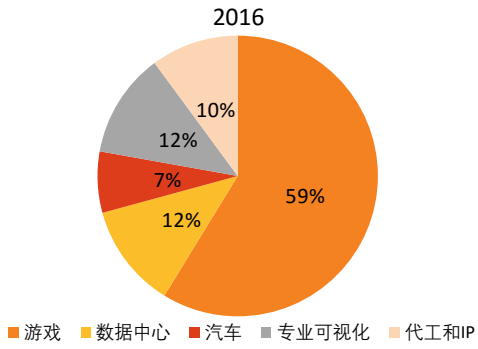
资料来源：公司财报，天风证券研究所整理

图 44：英伟达 R&D 投入以及 R&D/营收



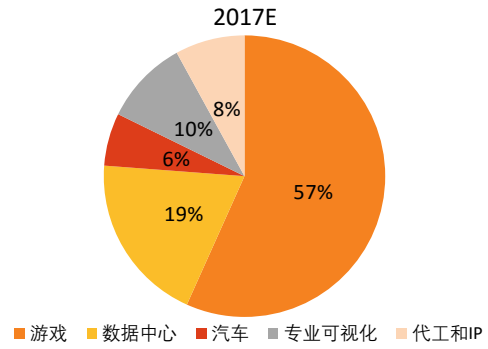
资料来源：公司财报，天风证券研究所整理

图 45：英伟达各项业务营收占比-2016 年



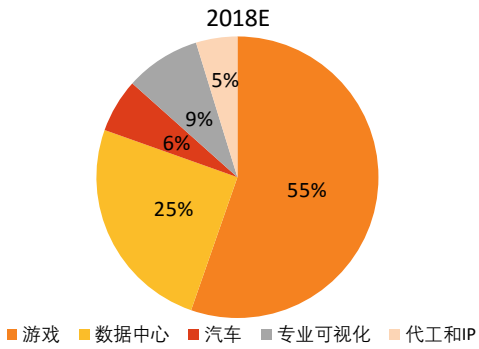
资料来源：公司财报，天风证券研究所整理

图 46：英伟达各项业务营收占比-2017 年 E



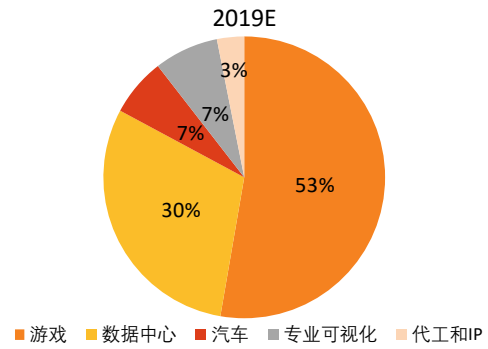
资料来源：公司财报，天风证券研究所预测

图 47：英伟达各项业务营收占比-2018 年 E



资料来源：公司财报，天风证券研究所预测

图 48：英伟达各项业务营收占比-2019 年 E



资料来源：公司财报，天风证券研究所预测

2.1.7. 英伟达整体盈利预测

图 49：英伟达整体盈利预测

百万美元	2014 FY2015	2015 FY2016	2016 FY2017	2017E FY2018	2018E FY2019	2019E FY2020
主营业务收入	**4,681.5**	**5,010.0**	**6,910.0**	**9,459.2**	**11,328.5**	**13,569.5**
同比增长%	77%	7%	38%	37%	20%	20%
游戏	2,058.0	2,818.0	4,060.0	5,364.6	6,330.3	7,153.2
同比增长%		37%	44%	32%	18%	13%
数据中心	317.0	339.0	829.8	1,843.0	2,800.3	4,088.3
同比增长%		7%	145%	122%	52%	46%
汽车	183.0	320.0	487.0	573.2	705.0	902.4
同比增长%	0.9	75%	52%	18%	23%	28%
专业可视化	795.0	750.0	835.0	922.0	963.5	1,002.0
同比增长%		-6%	11%	10%	5%	4%
代工和IP	1,329.0	783.0	698.0	756.4	529.5	423.6
同比增长%		-41%	-11%	8%	-30%	-20%
主营业务成本	2,082.0	2,199.0	2,847.0	3,849.9	4,576.7	5,400.7
GAAP毛利	**2,599.5**	**2,811.0**	**4,063.0**	**5,609.3**	**6,751.8**	**8,168.9**
毛利率%	55.5%	56.1%	58.8%	59.3%	59.6%	60.2%
GAAP营业支出	**1,840.5**	**2,064.0**	**2,129.0**	**2,743.2**	**3,115.3**	**3,528.1**
GAAP研发费用	1,359.7	1,331.0	1,463.0	1,891.8	2,152.4	2,442.5
as % of sales	29.0%	26.6%	21.2%	20.0%	19.0%	18.0%
GAAP管理费用	480.8	602.0	663.0	851.3	962.9	1,085.6
as % of sales	10.3%	12.0%	9.6%	9.0%	8.5%	8.0%
其他		131.0	3.0			
Non-GAAP毛利	**2,599.5**	**2,837.0**	**4,089.0**	**5,656.6**	**6,808.4**	**8,209.6**
毛利率%	**55.5%**	**56.6%**	**59.2%**	**59.8%**	**60.1%**	**60.5%**
Stock-Based Compensation等	183.5	343.0	262.0	473.0	419.2	339.2
Non-GAAP营业支出	1,657	1,721	1,867.0	2,270.2	2,696.2	3,188.8
as % of sales	35.4%	34.4%	27.0%	24.0%	23.8%	23.5%
Non-GAAP营业利润	954	1,125	2,221.0	3,386.4	4,112.3	5,020.7
as % of sales	20.4%	22.5%	32.1%	35.8%	36.3%	37.0%
所得税支出	153.0	196.0	371.0	575.7	699.1	853.5
Non-GAAP净利润	**801.0**	**929.0**	**1,850.0**	**2,810.7**	**3,413.2**	**4,167.2**
as % of sales	17.1%	18.5%	26.8%	29.7%	30.1%	30.7%
同比增长%		16%	99%	52%	21%	22%
Non-GAAP 摊薄加权平均股数	564	556	605	630	630	630
Non-GAAP摊薄加权平均每股收益	**1.42**	**1.67**	**3.06**	**4.46**	**5.42**	**6.61**
同比增长%		18%	83%	46%	21%	22%

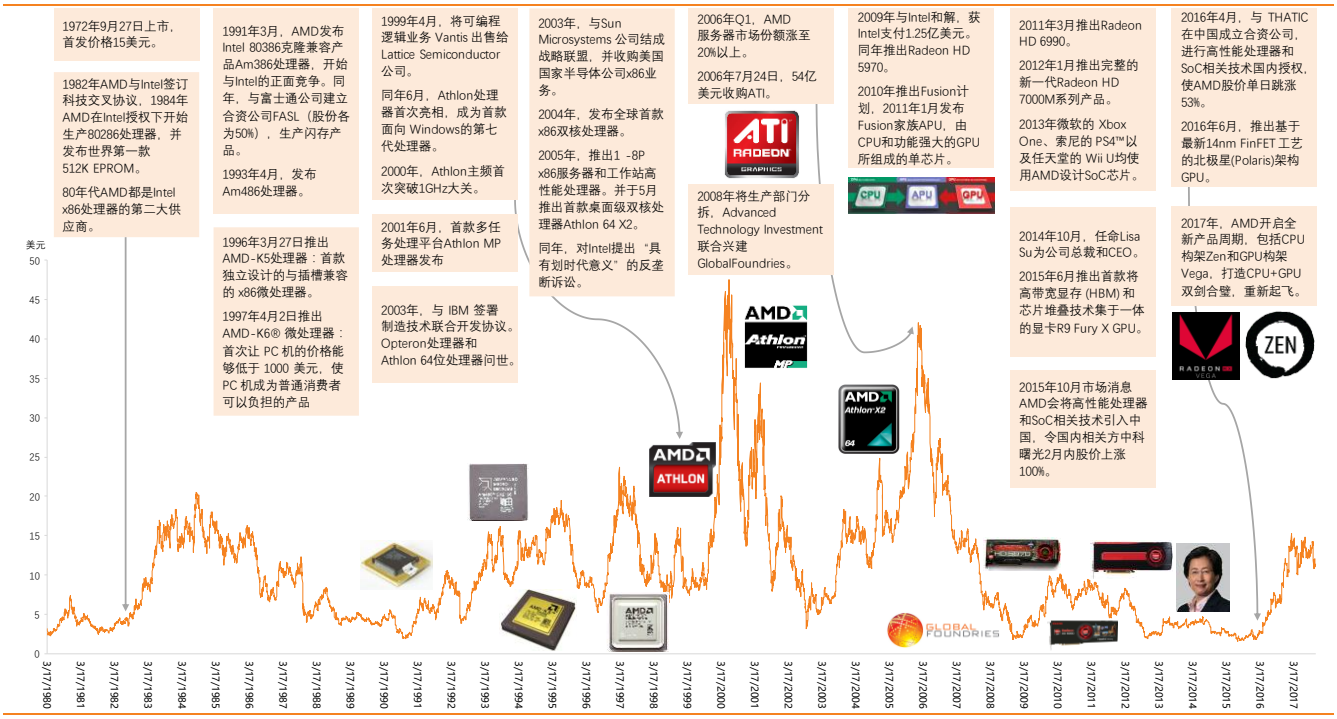
资料来源：公司财报，天风证券研究所预测

2.2. AMD：CPU+GPU 双剑合璧，不畏阻力起飞时

AMD 在 GPU 和 CPU 市场，都屈居行业老二的位置，当下在人工智能芯片布局上也慢英伟达一步。但是作为唯一拥有 GPU 和 x86 硅芯片技术的公司，我们认为随着公司产品线上移，重回高端市场，并且从零到一破局数据中心市场，利用 GPU+CPU 异构计算技术储备的协同效应，进一步修复利润率，在公司 CEO Lisa Su 的带领下，打开与英伟达、英特尔正面竞争之外的市场，200 亿美元市值亦可期。

但是作为唯一拥有 GPU 和 x86 硅芯片技术的公司，我们认为随着公司产品线上移，重回高端市场，并且从零到一破局数据中心市场，利用 GPU+CPU 异构计算技术储备的协同效应，进一步改善毛利，在公司 CEO Lisa Su 的带领下，打开与英伟达、英特尔正面竞争之外的市场，200 亿美元市值亦可期。

图 50：AMD 历史大事件



资料来源：Bloomberg，天风证券研究所整理，数据截止至 2017 年 11 月 27 日，采用调整后收盘价

AMD 一直强调在扩大计算性能的同时保持功耗稳定，尝试通过利用 GPU+CPU 在异构系统(heterogeneous system)中的协同，来提高每瓦特性能。而异构计算最为广泛的应用便是在深度学习中，从自动驾驶到高级机器人到医疗保健到诈骗监控。AMD 在早年就已经开发了加速处理单元(Accelerated Processing Unit, APU)来探索 CPU 和 GPU 更紧密的联结性。

AMD 的 CPU+GPU 双剑合璧，将会为公司的重新腾飞指明道路。

1、CPU 业务：Ryzen 闪耀 PC 端，EPYC 从零到一破局服务器市场

1）服务器产品 EPYC 正式推入市场，与所有主流 OEM 产品适配推进，并获得百度、微软 Azure、Bloomberg、Dropbox 等合作背书。从零到一回归服务器市场，带来业务增长最大弹性。我们认为 AMD 在 EPYC 发布的接下来 12 月周期内，市场份额能够从接近 0%提升到 1.5-3%，对应收入可以达到 2.7 亿美元至 5.4 亿美元。

2）PC 端完成 Ryzen 主流桌面产品布局，下半年产品包括 Ryzen Pro 和移动笔记本版本，我们看好 CPU 产品线全面布局实现量价齐升，在 PC 市场份额有机会返回 5 年前 25%的水平甚至 2010 年左右接近 30%的水平。

2、GPU 业务：进军云计算享 AI 之夏浪潮，Vega 新周期打开高端市场

1）发力云计算数据中心以及机器学习，已获得谷歌云和阿里云合作，我们看好 AMD 发挥 CPU+GPU 在异构系统(heterogeneous system)中的协同作用，实现从下游推理端(Inference)

和应用流程(deployment stage)向上游学习流程(training stage)的渗透，享受数据中心的巨大增长空间红利，3 年内看到 5%以上的份额突破。

2）Vega 构架 GPU 发布新周期，通过 Vega 发掘高端显卡市场机会，有望获得利润率的修复，我们将持续观察 Vega 系列正式放量之后的市场反馈。英伟达 Volta 构架面向游戏市场的 GPU 产品可能会到 2018 年才面市，因此 AMD Vega 会成为当前游戏市场上高端显卡性价比最高的选择。我们预测 AMD 凭借 Vega 显卡的正式放量，有赖传统下半年消费电子销售旺季推动，进一步提升下半年的市场份额。

3、中国云计算和 AI“春光乍现”，借力中科曙光拓宽市场

去年 4 月，AMD 将高性能处理器和 SoC 相关技术授权给与中科院背景的中科曙光参股子公司 THATIC 天津海光成立的 JV。我们认为 AMD 依托中科院/中科曙光等国内芯片最高生产力和资源，打开国内数据中心 CPU 服务器市场。目前 AMD 的 EPYC 已获百度合作支持。另外我们分析国内人工智能产业，当前仍处于“春光乍现”的萌芽阶段。中国拥有全球最多的用户和活跃数据生产主体，增长空间巨大，从上游芯片需求端来看，也适应于 AMD 方案的高性价比策略。我们认为 AMD 还可以依靠 GPU 的捆绑销售，加速切入国内数据中心和 AI 发展快车道。我们看好 AMD 在开拓中国人工智能上游芯片市场上时会拥有更为蓬勃的动力。

2.2.1. CPU 量价齐升：Ryzen 闪耀 PC 端，EPYC 从零到一破局服务器市场

蛰伏五年开发的全新 Zen 架构，作为从底层开始完全重新设计的 CPU 架构，为 AMD 带来 PC 端 Ryzen 以及数据中心服务器 EPYC。两款产品命名也颇具心思，EPYC 和 Ryzen 分别以 Y 指代 I，意指 Epic 和 Risen 的“传奇”与“崛起”。我们认为这两款处理器重磅新品，将会为 AMD 带来充沛营收的同时也修复利润：1）从零到一，重回服务器市场，带来破局性的单槽高性价比服务器；2）PC 端产品全布局，实现量价齐升。

2.2.1.1. EPYC——从零到一破局服务器市场

AMD 二季度发布的服务器处理器 EPYC，基于蛰伏五年开发的全新 Zen 架构，将数据中心市场的份额提升作为核心的战略规划。AMD 预计数据中心市场空间在 2020 年达到 210 亿美元，隐含约 160 亿 CPU 市场以及约 50 亿 GPU 市场，尤其在 GPU 部分预计 CAGR 可达 75%以上。

目前处于与所有主流 OEM 产品适配推进阶段，已获得百度、微软 Azure、Bloomberg、Dropbox 等合作背书，公司将数据中心市场份额提升作为核心的战略规划。我们看好 EPYC 以单插槽性价比作为最大竞争力，为双插槽服务器为主导的市场提供更多选择。结合公司目标和市场一致预期，我们预测 AMD 在 EPYC 发布的接下来 12 月周期内，数据中心处理器市场份额能够从接近 0%提升到 1.5-3%，对应收入可以达到 2.7 亿美元至 5.4 亿美元。

AMD 在 2004 年发布了全球首款 x86 双核处理器，并在 2006 年时服务器 CPU 市场份额一度达到接近 25%的高值。但随着在代号 Barcelona 的四核处理器上 bug 无法解决而导致延期，英特尔却带来了划时代的 Core 构架 Xeon 处理器，并抢先发布 x86 四核处理器。AMD 只能通过价格战疲于应对，产品降价措施也令 AMD 走入下坡路。此后的 Bulldozer 推土机架构也全面落后于英特尔，令 AMD 在服务器市场的份额丧失殆尽。

数据中心收入贡献弹性测试：毛利和营业利润双提升

我们认为进入数据中心处理器市场将会对公司的整体毛利和营业利润带来显著提升。我们对数据中心处理器业务给 AMD 的营收贡献进行测算。英特尔 2016 年数据中心业务平台收入为 159 亿美元，2015 年为 148.6 亿美元，同比增长 7%。2017 年 Q1 英特尔数据中心业务平台收入为 39.79 亿美元，同比增长 4%。公司指引未来 5 年的服务器 CPU 营收增速约为 CAGR 6%，营业利润率降低到 40-45%区间。

考虑到截止 2016 年，全球数据中心处理器市场的份额超过 99%为英特尔占据，我们以 159 亿美元作为 2016 年市场总空间，中性估计未来 4 年 CAGR 6%增速。则 2017-2020 年的数据中心处理器市场空间分别为 169 亿、179 亿、189 亿、200 亿美元。我们对 AMD 不同市

场份额下带来的营收贡献做弹性测试。结合公司目标和市场一致预期，我们预测 AMD 在 EPYC 新产品发布的接下来 12 月周期内，数据中心处理器市场份额能够从接近 0%提升到 1.5-3%，对应收入可以达到 2.7 亿美元至 5.4 亿美元。

图 51：AMD 服务器业务收入测算

		2016E	2017E	2018E	2019E	2020E
市场空间（百万美元） CAGR 6%		15,906	16,860	17,872	18,944	20,081
AMD 数据中心处理器业务收入 （百万美元）	悲观估计	0.8%	1.5%	2.8%	3.5%	
		135	268	530	703	
	中性估计	1.0%	2.0%	3.2%	4.5%	
		169	357	606	904	
	乐观估计	2.0%	3.0%	4.5%	6.0%	
		337	536	852	1205	

资料来源：AMD 官网，天风证券研究所

另外我们考察利润率的提升情况，回顾 2006 年，AMD 当年服务器 CPU 市场份额一度达到接近 25%的高值，2006Q1 的毛利率和营业利润率分别为 58.5%和 19.4%，当季实现净利润 1.85 亿美元，净利润率为 13.9%。在 2016Q4 的毛利率为 32%，营业利润率为 23%。

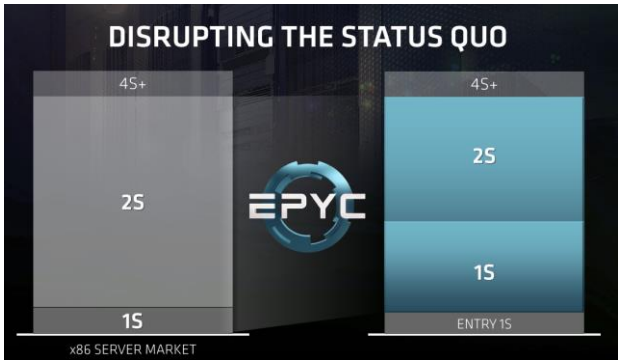
而英特尔方面 2006 年 Q1 Non-GAAP 毛利率和营业利润率分别为 44.9%和 23.2%。英特尔 2016Q4 在数据中心业务的营业利润为 18.81 亿美元，营业利润率达到了 40.3%。英特尔公司整体营业利润率为 27.6%，毛利率为 62%，我们看到数据中心业务的营业利润率显著高于英特尔其他业务。

EPYC：抢滩双插槽，破局单插槽

服务器产品 EPYC 原代号为 Naples，此次推出的 7000 系列一共包含 12 款服务器，包括 9 款双插槽和 3 款单插槽产品（产品标识为后缀 P）。EPYC 以 8 核心模块为基础，多个核心整合于单个芯片封装内，处理器内部和外部则主要通过全新的 Infinity Fabric 互联总线，作为 EPYC 系统的通信基础。

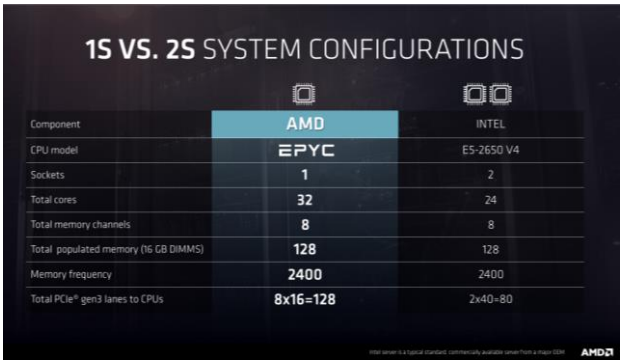
公司目前正与云计算服务商进行项目推进，将数据中心市场的份额提升作为核心的战略规划。目前包括 HPE、戴尔、华硕、技嘉、英业达、联想、曙光、超微、泰安和纬创已经推出基于 EPYC 处理器的产品，微软、红帽和 VMware 等主要虚拟机管理和服务器运行系统供应商也表示将优化支持 EPYC。在云计算服务商客户上，AMD 还宣布获得了百度、微软 Azure 以及 1&1、Bloomberg、Dropbox 和 LexisNexis 等公司的合作支持背书。

图 52：EPYC 目标成为单插槽服务器市场的破局者



资料来源：AMD 官网，天风证券研究所

图 53：EPYC 主打单插槽高性能，超过 50%英特尔双插槽产品



资料来源：AMD 官网，天风证券研究所

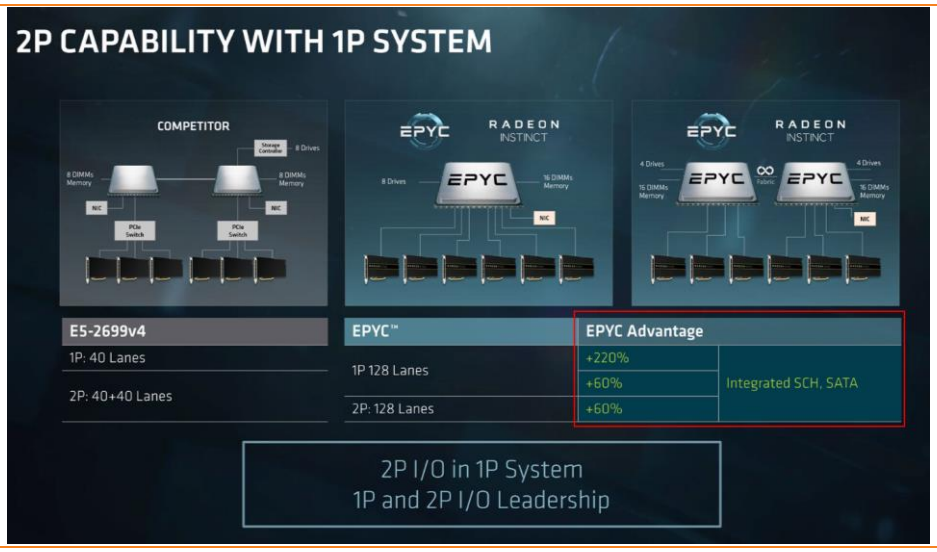
AMD 预计 EPYC 推出后对服务器市场的破局性重新划分，单插槽产品崛起将抢占近一半的原先由双插槽产品占据的市场份额。尤其在单插槽产品上可以对标英特尔 E5 双槽系统，在同等价位上可以提供比对手更多的核心与扩展、更高的性能，最大化性价比优势。根据 IDC 数据，在 2016 年数据中心市场 80%的处理器都是双插槽的，仅有 9%是单插槽。单槽服务器相对于双槽服务器的优势体现在所需要的主板更小且功耗更低。我们看好 EPYC 以单插槽性价比作为最大竞争力，为以双插槽服务器为主导的市场提供更多选择。

图 54：EPYC 系列产品基本情况（后缀 P 为单槽处理器）

型号	核心 / 线程	基准频率	最大超频	热设计功耗	较Intel同级产品性能提升	价格
EPYC 7601	32 / 64	2.2 GHz	3.2 GHz	180W	+47%	$ 4000起
EPYC 7551	32 / 64	2.0 GHz	3.0 GHz	180W	+44%	$ 3200起
EPYC 7501	32 / 64	2.0 GHz	3.0 GHz	155/170W	—	$ 3400
EPYC 7451	24 / 48	2.3 GHz	3.2 GHz	180W	+47%	$ 2400起
EPYC 7401	24 / 48	2.0 GHz	3.0 GHz	155/170W	+53%	$ 1850
EPYC 7351	16 / 32	2.4 GHz	2.9 GHz	155/170W	+63%	$ 1100起
EPYC 7301	16 / 32	2.2 GHz	2.7 GHz	155/170W	+70%	$ 800起
EPYC 7281	16 / 32	2.1 GHz	2.7 GHz	155/170W	+60%	$ 650
EPYC 7251	8 / 16	2.1 GHz	2.9 GHz	120W	+23%	$ 475
EPYC 7551P	32 / 64	2.0 GHz	3.0 GHz	180W	对比Intel双槽 +21%	$ 2100
EPYC 7401P	24 / 48	2.0 GHz	3.0 GHz	155/170W	对比Intel双槽 +22%	$ 1075
EPYC 7351P	16 / 32	2.4 GHz	2.9 GHz	155/170W	对比Intel双槽 +21%	$ 750

资料来源：AMD 官网，天风证券研究所

图 55：EPYC 与 Radeon Instinct 加速器协同精简架构



资料来源：AMD 官网，天风证券研究所

今年 7 月 11 日，英特尔正式发布了旗下全新一代 Xeon Scalable 处理器，架构代号为 Skylake-SP，划分为 Bronze、Sliver、Gold、Platinum 四个子型号，相比上代产品来说，Xeon Scalable 支持 Omni-Path 高速互连架构(100Gbps)万兆因特网接入、第三代 AVX-512 指令集和傲腾 SSD。英特尔表示，Xeon Scalable 处理器是业界十年来在平台技术上的最大进步，在内核、缓存、内存、I/O 等多项优化的辅助下，每个时钟周期浮点性能提升两倍，8K 数据块时压缩速度可达 100Gb/s。Platinum 系列支持 2/4/8 插槽，售价在 3000-13000 美元；Gold 系列支持 2/4 插槽，售价 1200-3600 美元；而在入门系列的 Silver/Bronze 支持双插槽，售价 213-1000 美元。

英特尔表示，Xeon Platinum 8180 相比于 AMD EPYC 7601 核性能高出 28%。考虑到英特尔当前在服务器芯片市场近乎垄断的市场地位，此次推出性能大幅提升的 Xeon Scalable 处理器也证明了英特尔并不会将服务器市场拱手相让。

图 56：英特尔 Xeon Skylake 系列和 AMD EPYC 系列参数对比

	Intel Xeon E5 Bronze / Silver	Intel Xeon E7 Gold / Platinum	AMD Naples Platform (2P)
系列代号	Skylake-SP	Skylake-SP	AMD EPYC
工艺节点	14纳米	14纳米	14纳米
PCH	Lewisburg PCH	Lewisburg PCH	SOC
插槽	Socket P (LGA 3647)	Socket P (LGA 3647)	SP3 LGA socket
最大核心数	26	28	32
最大线程数	52	56	64
最大三级缓存	16.5 MB L3	38.5 MB L3	64 MB L3
DDR4内存支持	6通道 DDR4	6通道 DDR4	8通道 DDR4
热设计功耗	70-85W	85-205W	120-180W

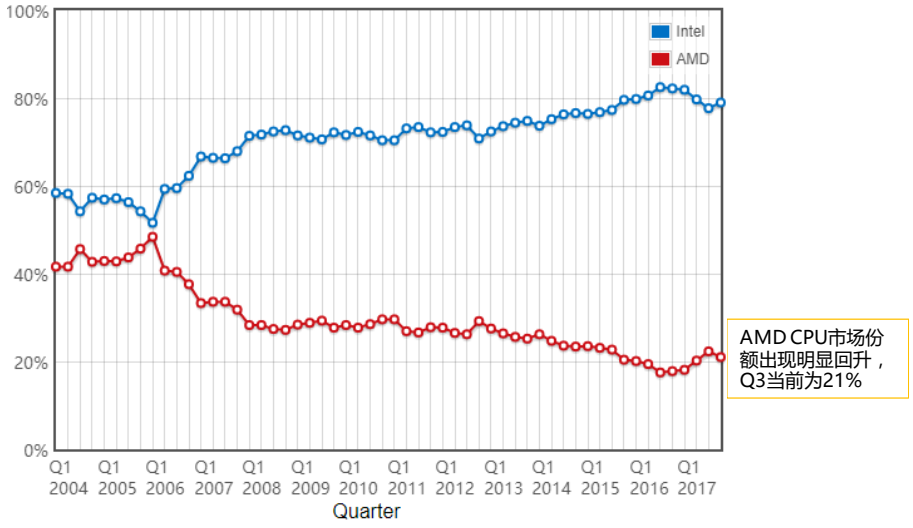
资料来源：英特尔官网，AMD 官网，天风证券研究所

2.2.1.2. Ryzen——蛰伏五年闪耀 PC 端

AMD 上半年推出的 Ryzen 7/5 系大受市场欢迎，Q2 的完整销售也为 AMD 的计算和图像业务（包含 PC 处理器和 GPU）带来同比 51%的收入增长。我们看到根据 PassMark 数据，AMD 在 PC 端 CPU 份额今年内已现明显回升至 20%以上。根据市场良好反馈，以及随着近日 Ryzen 3 发布完成 Ryzen 主流桌面产品布局，我们认为 AMD 的 PC 处理器市场份额有机会返回 5 年前 25%的水平甚至 2010 年左右接近 30%的水平。

公司通过 Ryzen 7/5/3 系列，下半年的 Ryzen Pro（台式机 CPU，下半年面向企业客户供货，明年上半年推出移动版），Ryzen Mobile（年末推出个人笔记本，明年推出商业笔记本），以及 Ryzen ThreadRipper 发烧级处理器（最高 16 核 32 线程，对标英特尔 i9 旗舰型号）等丰富整体 CPU 产品线配置。我们看好 CPU 新品迭出全面布局，实现量价齐升。

图 57：AMD 和英特尔的 PC 端 CPU 市场份额



资料来源：PassMark，天风证券研究所

今年 5 月英特尔发布了 Core i9 高端处理器系列，采用 Skylake-X 架构，核心数涵盖 10/12/14/16/18 核。目前，10 核 Core i9-7900X 已经发售，而 12 核 Core i9-7920X 预计将在 8 月 28 日发布，14 核 Core i9-7940X、16 核 Core i9-7960X 和 18 核的 Core i9-7980X 终极版将在 9 月 25 日推出，根据英特尔之前公布的发布日期，原定于 10 月份发布的 18 核 Core i9-7980XE 则有所提前，也是迫于 AMD Ryzen 大受市场好评所迫，英特尔不得不提前应对。

图 58：英特尔 Core i9-7920X 与 AMD Ryzen 1920X 基本参数对比

	Intel Core i9-7920X	AMD Ryzen 1920X
架构	Skylake-X	Zen
工艺节点	14nm	14nm
核心/线程	12核/24线程	12核/24线程
主频	2.90GHz	3.5GHz
超频	4.30GHz	4Ghz
PCIe通道	44	64
热功耗设计	140W	180W
高速缓存	16.5MB	38MB
定价	1199美元	799美元

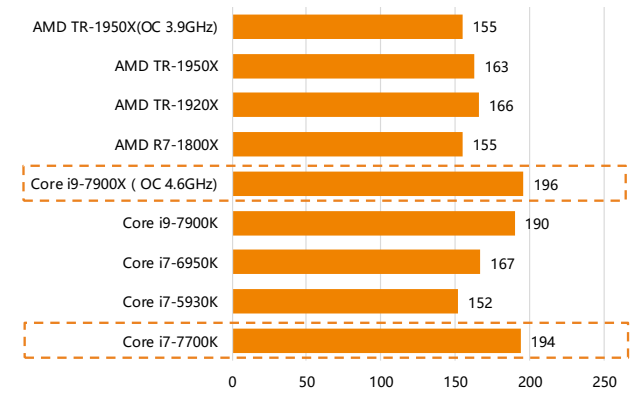
资料来源：英特尔官网、AMD 官网，天风证券研究所

英特尔一直是高端处理器市场的领导者，而在今年 7 月 9 日 AMD 发布了高端桌面 Threadripper 系列 CPU 之后，AMD 正在努力摆脱近年来低端产品代名词的名号。

通过 Threadripper 与英特尔 Skylake-X 系列产品基础数据的对比看出，AMD 将很有机会凭借单核价格(core/price)优势获取高端市场份额。AMD 在其产品中通过不断增加内核来提升整体性能，以对抗英特尔卓越的 IPC 性能（每个时钟周期执行的平均指令数），而目前来看 AMD 在保持较低价格的同时，性能则有了显著提升。面对核心、线程数迅速增加的 AMD，英特尔也在不断提高 CPU 核心及线程数，应对 AMD 提高不断施加的参数及性能压力。

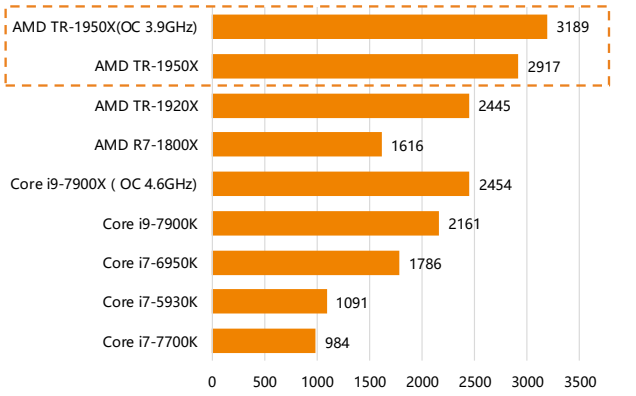
英特尔与 AMD 处理器单线程与多线程性能对比实验中，结果差异巨大。在单线程处理测试中，10 核的 Core i9-7900X 在对比 16 核 AMD 1950X 以及 12 核 AMD 1920X 上优势明显，体现了英特尔优秀的单核 IPC 性能。但在多线程测试中，英特尔 i9-7900X 被 AMD 1950X 超越，甚至与 AMD 1920X 处于同一水平。更多线程的设计使 AMD 处理器在进行内容创作、视频解码、3D 渲染和软件编译等任务时拥有实质性的优势，能够快速同时处理多个任务。我们也认为，购买 700-1000 美元价位高端 CPU 处理器的客户，包括游戏发烧玩家、视频处理工作者等，更加关注多线程下多任务处理性能，AMD Threadripper 的表现有望对高端 CPU 处理器市场带来相当的冲击。

图 59：处理器 Cinebench 单线程测试（越高越好）



资料来源：arstechnica，天风证券研究所

图 60：处理器 Cinebench 多线程测试（越高越好）



资料来源：arstechnica，天风证券研究所

2.2.1.3. 中国云计算和 AI“春光乍现”，借力中科曙光拓宽市场

去年 4 月，AMD 宣布将高性能处理器和 SoC 相关技术授权给其与 THATIC（天津海光先进技术投资有限公司，是中科曙光的控股子公司，后者又是依托于中科院）新成立的合资公司，开发只在中国市场进行销售的服务器芯片。AMD 预计会获得总额 2.93 亿美元的专利授权费，未来可能会从服务器销售中获得版税提成。消息公布后，AMD 的股价单日大涨 52.29%。AMD 通过与天津海光合资采用了双层架构设计，规避英特尔的专利限制，以及满足合资企业以国产身份销售境外 CPU 的要求。

AMD 与中科院背景的中科曙光子公司合作，代表了国内芯片生产技术的最高水平。我们认为现在仍然处于技术消化期，有消息称天津海光会在年内推出首款产品。我们认为 AMD 依托中科院/中科曙光等国内芯片最高生产力和资源，打开国内数据中心 CPU 服务器市场。目前 AMD 的 EPYC 已获百度合作支持。另外我们分析国内人工智能产业，当前仍处于“春光乍现”的萌芽阶段，中国拥有全球最多的用户和活跃数据生产主体，增长空间巨大，从上游芯片需求端来看，也适应于 AMD 方案的高性价比策略。

我们认为 AMD 还可以依靠 GPU 的捆绑销售，加速切入国内数据中心和 AI 发展快车道。我们看好 AMD 在开拓中国人工智能上游芯片市场上时会拥有更为蓬勃的动力。

2.2.2. GPU 上下游布局：Vega 打开高端市场，进军云计算享 AI 之夏浪潮

AMD 此前由于 Polaris 构架没有把握制程进步带来的功耗降低优势，而逐步失去了高端显卡市场的话语权。在人工智能机器学习最上游训练部分的布局渗透也慢英伟达一步。

不过 AMD 从去年 Q4 开始正式进军云计算市场，发力在云计算服务商超大规模(Hyperscale)数据中心的 GPU 部署，并收获谷歌云和阿里云的合作。AMD 去年 12 月发布针对机器学习的 GPU 加速器，Vega 也让 AMD 有能力直面上游训练端 EB 级数据处理及并行工作负载，我们看好 AMD 接下来进一步发挥 EPYC+Vega 的协同效应，通过交叉销售等方式，在行业兴起迅速扩张之时最大程度的享受人工智能芯片浪潮。

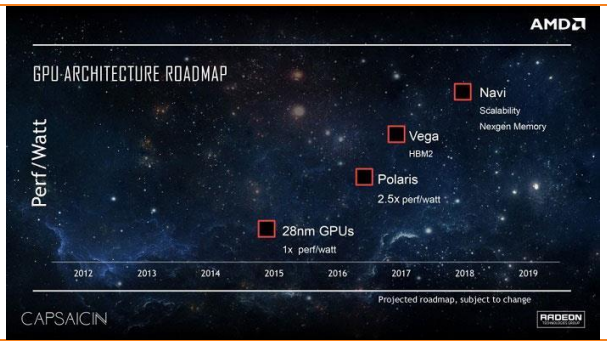
同时 AMD 规划在 14nm 工艺的 Vega 系列后，可能推出 7nm 的 Vega 升级构架，并确认了下一代显卡 Navi 构架，会采用 7nm FinFET 工艺，打造为首款专门针对人工智能运算优化的显卡，追赶英伟达 Volta 构架显卡的步伐。

今年 AMD 发布全新设计的 Vega 构架，进入 Vega 系列发布周期直击 GPU 高端市场，希望通过 Vega 发掘高端显卡市场机会，进一步提升高端市场份额。AMD 表示，高端 GPU 销量只占 GPU 市场的不到 15%，但能贡献超过 66%的利润。进入高端市场有望获得利润率的修复。

不过从下半年的 Vega 系列市场反馈来看，暂时没有达到 AMD 此前预期，高端显卡性能与英伟达 Pascal 架构相持平，未体现性价比优势。同时，11 月初 AMD 和英特尔联合宣布，将共同开发一款针对高端笔记本电脑的处理器产品，由英特尔负责 CPU 架构、AMD 负责 GPU 架构。双方的合作消息最早在今年 2 月传出，当时由于英特尔和英伟达的 IP 交叉授权协议行将到期，市场预期英特尔将转向 AMD 进行 GPU 技术合作。我们认为，英特尔和 AMD 的合作将有效拉低游戏笔记本的门槛，扩大游戏本的适用客户人群，新产品预计将于 2018Q1 发布。另外 AMD 于 Q4 发布的 Ryzen Mobile 笔记本处理器，主要面向商务笔记本市场，我们认为两者不会产生相互蚕食。

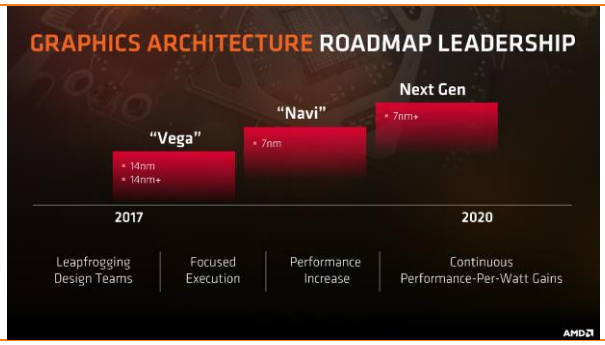
另外 AMD 高级副总裁、图像芯片设计核心部门 Radeon Technology Group 负责人、首席架构师 Raja Koduri 最近宣布离任，后火速加入英特尔，市场认为会对 AMD 的 GPU 业务产生较大影响，但我们认为由于 Vega 的开发低于预期，Raja 离任后部门归由 CEO Lisa Su 掌管，反而能进一步提高执行力。我们认为，高层变动叠加恰逢 AMD 与英特尔合作宣布，可能在未来产品线开发上会有更深入的合作推进。

图 61：AMD GPU 规划路径



资料来源：WCCFTech，天风证券研究所

图 62：AMD GPU 规划路径



资料来源：AMD 官网，天风证券研究所

2.2.2.1. 进军云计算，享人工智能芯片浪潮

AMD 从 16Q4 正式进军云计算市场，包括与谷歌云、阿里云达成合作，提供 GPU 加速方案。我们认为 AMD 发力云计算数据中心以及机器学习，尤其是在云计算服务商超大规模(Hyperscale)数据中心的 GPU 部署，将能让 AMD 最大程度的享受人工智能芯片浪潮。

我们看好人工智能立夏将至数据中心的巨大增长空间，当前全球服务器中 GPU 的渗透率仅为 0.24%的兴起之时，我们测算到 2020 年全球服务器 GPU 渗透率将达 4 倍以上增长。

AMD 将会发挥 GPU+CPU 在异构系统(heterogeneous system)中的协同作用，实现从下游推理端(Inference)和应用流程(deployment stage)向上游学习流程(training stage)的渗透，享受足够的市场份额红利，并借机缩小与英伟达的差距。

我们在前文对数据中心 GPU 市场的空间以及英伟达和 AMD 的收入贡献进行了测算，我们算得 2017 年 AMD 数据中心 GPU 收入约为 2000 万美元，2020 年增长至 2.5 亿美元，CAGR 达 132%。虽然总体体量并不大，但翻倍以上的增速将让 AMD 在行业兴起迅速扩张之时最大程度的享受人工智能芯片浪潮。

图 63：全球服务器 GPU 市场估计

	2013	2014	2015	2016	2017E	2018E	2019E	2020E
全球服务器出货量（千）	9,887	10,091	11,091	11,104	11,881	12,594	13,123	13,648
增长率%		*2%*	*10%*	*0%*	*7%*	*6%*	*4%*	*4%*
服务器AI使用率%				7.0%	10.0%	12.5%	14.5%	16.0%
服务器AI使用量（千台）				777	1,188	1,574	1,903	2,184
AI服务器中GPU使用率%				3.4%	5.1%	6.3%	8.0%	9.3%
AI服务器中GPU使用量（千台）				26	60	99	152	203
GPU占全球服务器使用率%				**0.24%**	**0.51%**	**0.85%**	**1.16%**	**1.49%**
英伟达数据中心深度学习使用率%				59%	64%	68%	70%	72%
数据中心GPU使用量（千台）				44.6	93.9	145.8	217.5	282.0
英伟达市场份额%				98%	98%	96%	94%	92%
数据中心英伟达GPU使用量（千台）				43.7	92.1	140.0	204.4	259.5
每台英伟达GPU服务器ASP（千美元）				19	20	20	20	20
英伟达数据中心收入（百万美元）				**830**	**1,843**	**2,800**	**4,088**	**5,190**
增长率%					*122%*	*52%*	*46%*	*27%*
AMD市场份额%				2%	2%	4%	6%	8%
数据中心AMD GPU使用量（千台）				.9	1.8	5.8	13.0	22.6
每台AMD GPU服务器ASP（千美元）				12	12	12	12	12
AMD数据中心GPU收入（百万美元）				**11**	**21**	**70**	**157**	**271**
增长率%					*100%*	*227%*	*124%*	*73%*
GPU数据中心市场空间（百万美元）				**840**	**1,864**	**2,870**	**4,245**	**5,460**
增长率%					*122%*	*54%*	*48%*	*29%*

资料来源：公司财报，Gartner，英特尔，天风证券研究所预测

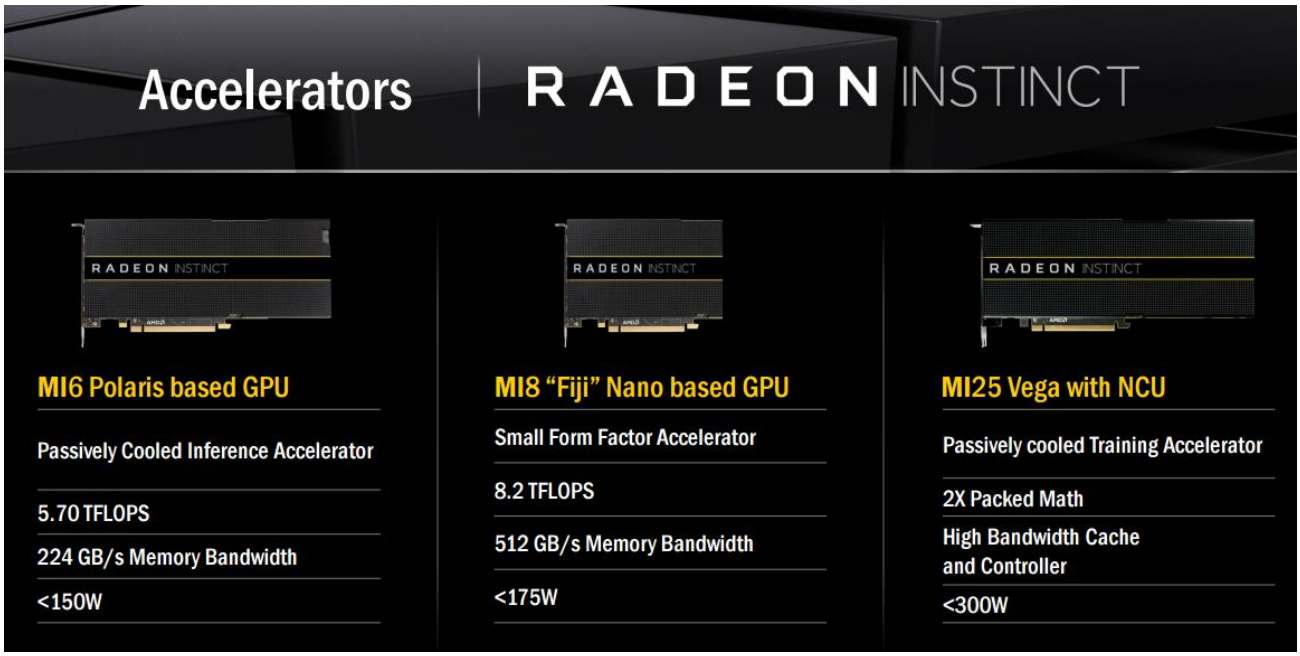
在与谷歌云的合作中，谷歌云计算平台将选择 AMD 最先进的单精度双 GPU 加速器 FirePro S9300 X2 服务器，用于包括复杂的医疗和财务模拟、地震和地下勘探、机器学习、视频渲染和代码转换以及科学分析等方面。对比英伟达的 GPU 性能，AMD FirePro 服务器的最大单精度性能为 13.9 TFLOPS，而英伟达的 Tesla P100 GPU 为 9.3 TFLOPS；FirePro 拥有 1TB/s 的内存带宽，对比英伟达 P100 的 732GB/s。此外 FirePro 还拥有 8GB HBM 内存以及 300 瓦能耗。AMD 与阿里云的合作则将提升阿里云远程工作站、云游戏、云计算以及虚拟桌面架构(VDI)运行的安全性。

我们认为，AMD 将在 CPU+GPU 混合型服务器市场较英特尔更强竞争力。我们看好 AMD 接下来进一步发挥 EPYC+Vega 的协同效应，通过交叉销售等方式，在行业兴起迅速扩张之时最大程度的享受人工智能芯片浪潮。去年 12 月，AMD 发布 Radeon Instinct 加速器以及配套开源软件，是基于 GPU 的深度学习推理和训练加速解决方案。Radeon Instinct 加速器拥有三种型号，分别基于 Polaris、Fiji 以及最新的 Vega GPU 架构，前两者主要针对推理端加速，而基于 Vega 的 Radeon Instinct MI25 加速器专为深度学习训练端而设计，将与 ROCm 软件平台和 MIOpen 深度学习库协作，针对求解整体时间进行优化。

数据中心加速市场空间呈现不断扩大发展趋势，CPU+GPU 的混合处理器组合逐渐受到数据中心和云计算供应商的欢迎，这种组合将计算负担从 CPU 转移到 GPU，通过发挥 GPU 的运算速度和能力来提高工作效率。

随着 EPYC 和 Vega 产品的逐步面市，公司目前正与云计算服务商进行项目推进，将数据中心市场的份额提升作为核心的战略规划。我们认为发力云计算数据中心以及机器学习，尤其是云计算服务商超大规模(Hyperscale)数据中心的 GPU 部署，将在 GPU 市场进一步缩小与英伟达的差距，从下游应用流程(deployment stage)向上游的机器学习流程的渗透。

图 64：AMD 去年发布的 Radeon Instinct 加速器



资料来源：AMD 官网，天风证券研究所

2.2.2.2. 进入 Vega 发布新周期，重回显卡高端市场

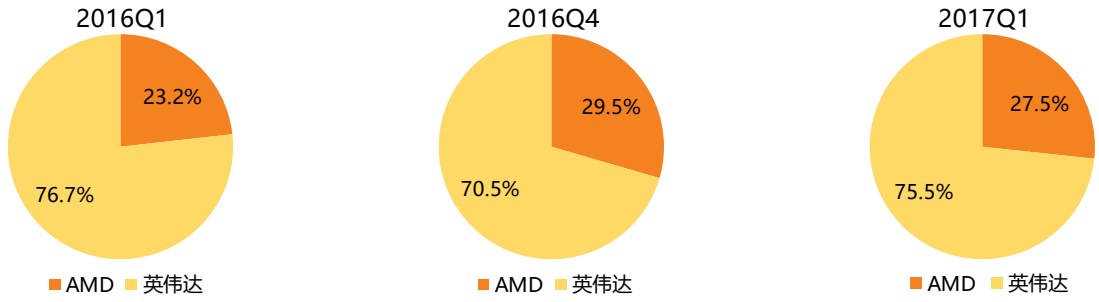
公司表示，高端 GPU 销量只占 GPU 市场的不到 15%，但能贡献超过 66%的利润。接下来 AMD 进入 Vega 构架 GPU 发布周期，希望通过 Vega 发掘高端显卡市场机会，进一步提升高端产品市场份额，对比上一代 Polaris 构架显卡面对 200 美元以下市场。我们将持续观察 Vega 系列正式放量之后的市场反馈。

AMD 在 2016 年发布了 Polaris 构架显卡，对标英伟达上上代 Maxwell 构架显卡，但售价比英伟达低。Polaris 帮助 AMD 在 2016Q4 的 299 美元以下 GPU 份额达到 31%（对比 2015Q4

的 21%），2016 全年整体 GPU 市场份额提升至 28%（对比 2015 年 20%）。

根据 Jon Peddie 市场预测，AMD 在今年 Q1 的市占率在 27.5%，较 16Q4 下跌了 2%，主因英伟达调整了旧款显卡售价，并发布了高端型号 GeForce GTX 1080 Ti 和 GTX 1080 Titan XP。而 AMD 在上半年推迟 Vega 发布的情况下推出了 Polaris 构架的 RX 500 系列显卡，并在数字货币挖矿热潮驱动下，受到了矿工的大肆抢购。因此我们预测今年 Q2 的市占率 AMD 下降的因素会被挖矿抢购所抵消。

图 65：AMD 与英伟达显卡市占率对比



资料来源：Jon Peddie，天风证券研究所

6 月 27 日，AMD 正式发布 Vega 构架的第一款显卡——Radeon Vega Frontier Edition 显卡，针对机器学习开发和高级可视化工作负载。风冷版定价 999 美元，水冷版定价 1499 美元。Radeon Vege Frontier 显卡内部为 Vega 10 架构，搭配 16GB HBM2 显存，使用双 8pin 接口供电，单精度浮点 13 TFLOPS，半精度浮点 25 TFLOPS，拥有 64 个计算单元，4096 个流处理器，显存带宽为 480 GB/s。

图 66：AMD Radeon Vega Frontier 显卡

资料来源：AMD 官网，天风证券研究所

图 67：AMD Radeon Vega Frontier 基本信息

Radeon™ Vega Frontier Edition

Compute units	64
Single precision compute performance (FP32)	~13 TFLOPS
Half precision compute performance (FP16)	~25 TFLOPS
Pixel Fillrate	~90 Gpixels/sec
Memory capacity	16 GBs of High Bandwidth Cache
Memory bandwidth	~480 GBs/sec

资料来源：AMD 官网，天风证券研究所

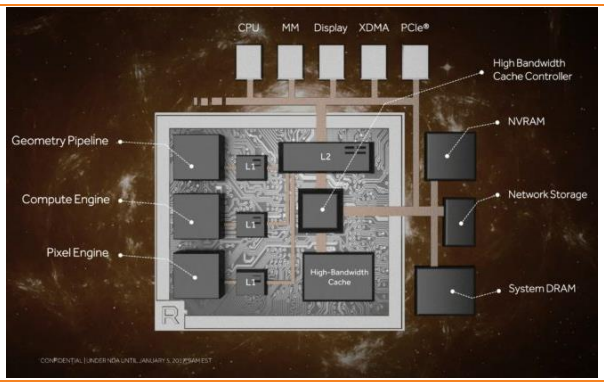
图 68：Vega Frontier 性能比较

AMD Vega specs

	Radeon Vega Frontier Edition	Radeon RX 580	GTX 1080 Ti
GPU	AMD Vega 10	AMD Polaris 20	Nvidia GP102
Architecture	GCN 4.0	GCN 4.0	Pascal
Lithography	14nm FinFET	14nm FinFET	16nm FinFET
Stream Processors	4,096	2,304	3,584
Texture units	256	144	224
Render output units	64	32	88
Memory Capacity	16GB HBM2	8GB GDDR5	11GB GDDR5X
Memory bus	2,048-bit	256-bit	352-bit
Performance	12.5 TFLOPs	5.8 TFLOPs	11.8 TFLOPs
TDP	< 300W	185W	250W

资料来源：PCgames，天风证券研究所

图 69：Vega 架构设计

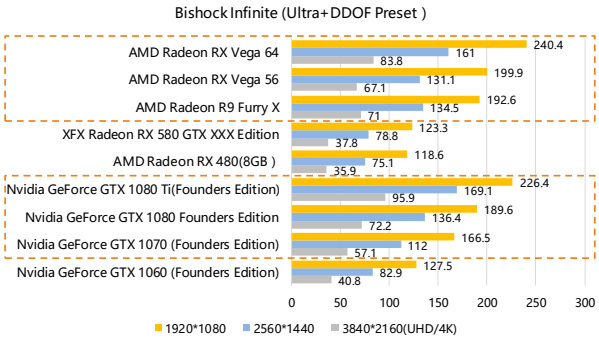


资料来源：Wccftech，天风证券研究所

最近发布的旗舰游戏显卡 Radeon RX Vega 64/56 两款，市场提前评测褒贬不一，综合性能媲美英伟达 Pascal 系列，Vega 64 对标 GTX 1080，Vega 56 对标 GTX 1070。不过对比已经发售了一年的 1070/1080 并没有体现出明显优势，同时在功耗上明显高于英伟达显卡。

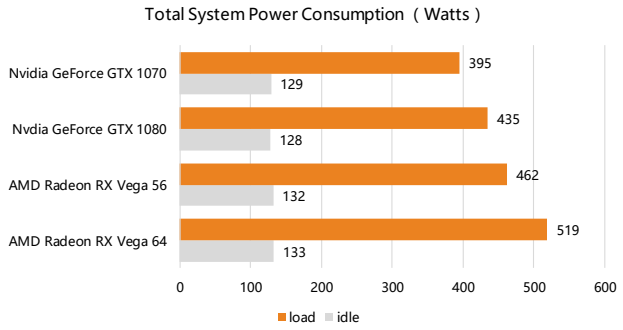
而展望下半年，英伟达 Volta 构架面向游戏市场的 GPU 产品可能会到 2018 年才面市，因此 AMD Vega 会成为当前游戏市场上高端显卡性价比最高的选择。另外苹果在更新 iMac 桌面电脑产品线后拥抱选择 AMD 显卡，也会一定程度拉高消费者喜好程度。我们预测 AMD 凭借 Vega 显卡的正式放量，有赖传统销售旺季推动，会进一步提升下半年的市场份额。

图 70：Vega RX 和英伟达 GTX 性能比较（越高越好）



资料来源：PCgames，天风证券研究所

图 71：系统总功耗（越低越好）



资料来源：Forbes，天风证券研究所

2.2.2.3. 苹果全新 iMac Pro“钦点”Vega

苹果在 6 月举办的 WWDC 开发者大会上，丰富的硬件产品线更新亦让 AMD 作为显卡供应商收获市场关注。最新的 21.5 英寸和 27 英寸 iMac 产品使用 Polaris 构架的 Radeon Pro 500 系列，包括 555/560/575/580 等型号，以 Radeon Pro 580 为例，该显卡有 2304 个流处理器，8GB 显存，浮点运算性能为 5.8 TFLOPS，旨在为用户提供沉浸式 VR 体验和更流畅的媒体设计创作平台。用户可以在 Mac 上运行多媒体程序时体验 GPU 加速功能。

苹果还发布了旗舰版 iMac Pro，采用最顶级的 Radeon Vega 显卡，单精度浮点运算能力达 11 TFLOPS，半精度浮点运算达到 22 TFLPS，而在五月 AMD 发布的 Vega 版本中，单精度性能是 13TFLOPS，我们认为 iMac Pro 一体机设计以及定制化的 Vega 设置会稍微降低 TFLOPS 精度或降低 clock rate。iMac Pro 有两款 Vega GPU 可供选择，Vega 56 和 Vega 64，分别有 3584 个和 4096 个流处理器，对应 8GB 和 16GB 的 HBM2 显存。新版显卡每时钟周期可处理的多边形建模数量是之前版本的两倍多，同时，高带宽缓存及控制器工艺，采用 HBM2 技术，层叠式封装的 HBM2 显存取代了外置显存，使图形处理器能以最高达 400GB/s 的速度提取数据。

图 72：苹果更新 iMac 产品线，高端机型将搭载 AMD 显卡

资料来源：Apple 官网，天风证券研究所

图 73：苹果发布的 iMac Pro 将搭载 Vega 显卡

资料来源：Theverge，天风证券研究所

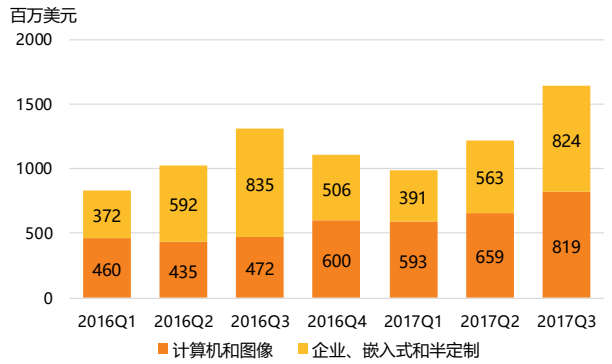
2.2.3. 估值：行业老二起飞时，重申“买入”，TP 16 美元

AMD 3Q17 实现 non-GAAP EPS 0.10 美元，营收 16.4 亿美元，均超过华尔街预期的 EPS 0.08 美元和营收 15 亿美元，主要鉴于 Ryzen CPU 销售火爆在 DIY 电脑市场的持续受捧提振，Vega GPU 开始逐渐放量，另外以太坊挖矿也带来一定业绩持续。计算机和图像业务营收为 8.2 亿美元，同比大涨 74%，继续保持营业净利润。企业、嵌入式和半定制业务营收为 8.2 亿美元，同比持平，环比大涨 46%。公司整体毛利上升到 35%。

Q4 指引营收环比跌 15%，同比增长 26%，环比下跌主要因为半定制游戏主机业务的正常季节因素，指引仍高于市场预期，毛利率指引维持 35%。全年营收指引进一步上调至超过 20%，毛利率指引 34%，并实现全年 Non-GAAP 扭亏为盈。

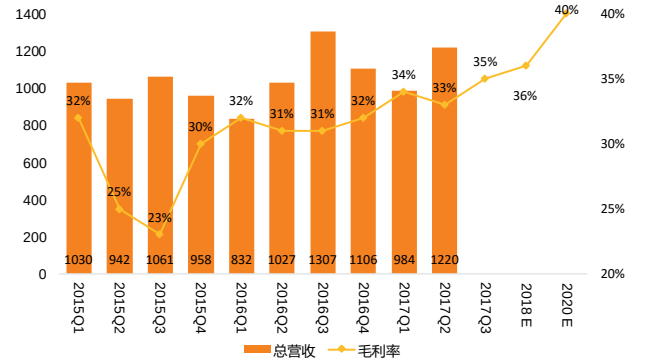
AMD 此前在 5 月中旬的分析师开放日上给出公司长期财务规划，预计毛利率从 16 年 31%提升至 17 年 34%，18/20 年指引分别为 36%和 40%；长期目标毛利率进入 40-44%区间，经营费用降至 26-30%，Non-GAAP EPS 达到 0.75 美元以上。公司预计总体市场空间达 640 亿美元，包括 280 亿美元 PC 市场，150 亿游戏市场以及 210 亿数据中心市场，数据中心**隐含约 160 亿 CPU 以及约 50 亿 GPU 市场，尤其在 GPU 部分预计 CAGR 可达 75%以上。**

图 74：AMD 各项业务营收比较（百万美元）



资料来源：公司财报，天风证券研究所整理

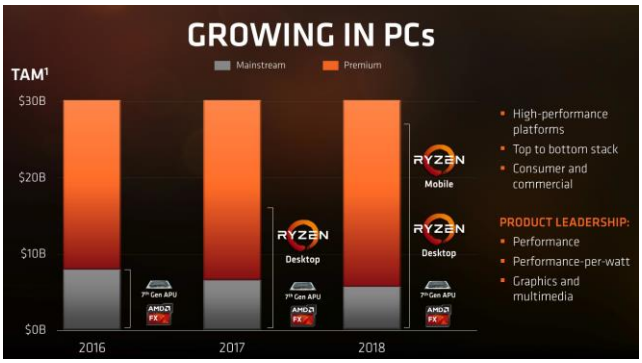
图 75：AMD 各季营收（百万美元）及毛利率指引



资料来源：公司官网，天风证券研究所整理

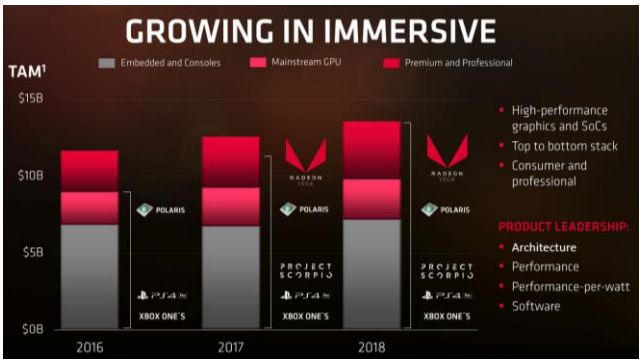
我们认为，公司新品迭出将拉动营收进入双位数增长，毛利的提升在于产品线持续上移。公司表示 PC 市场 70%以上份额属于高端产品，正在通过 Ryzen 7/5 系列（已获主流 PC 厂适配），下半年的 Ryzen Pro（台式机 CPU，下半年面向企业客户供货，明年上半年推出移动版），Ryzen Mobile（集成 Zen 架构 CPU 内核以及 Vega 构架 GPU，相对于第 7 代 APU 性能 CPU/GPU 性能提高 50%/40%，功耗降低 50%，Q4 推出个人本，明年推出商业本），以及 Ryzen ThreadRipper 处理器（最高 16 核 32 线程，64 条 PCIe 3.0 总线，16MB 三级缓存，四通道 DDR4 内存，对标英特尔 i9 旗舰型号）等丰富产品线配置。

图 76：AMD PC 市场将依赖 Ryzen 系列扩张高端市场



资料来源：AMD 官网，天风证券研究所

图 77：AMD 游戏市场将靠 Vega 系列重回高端



资料来源：AMD 官网，天风证券研究所

AMD 在 GPU 和 CPU 市场，都屈居行业老二的位置，当下在人工智能芯片布局上也慢英伟达一步。但是作为唯一拥有 GPU 和 x86 硅芯片技术的公司，我们认为随着公司产品线上移，重回高端市场，并且从零到一破局数据中心市场，利用 GPU+CPU 异构计算技术储备的协

同效应，进一步改善毛利，在公司 CEO Lisa Su 的带领下，打开与英伟达、英特尔正面竞争之外的市场，200 亿美元市值亦可期。

在游戏主机业务方面，AMD 将这块的销售收入归入企业、嵌入式和半定制(EESC)业务中。半定制业务就是为包括索尼 PS4 和微软 Xbox One 等游戏主机提供处理器芯片。去年索尼发布 PS4 Pro，带动 AMD 的 EESC 业务在 2016Q3 达到 8.35 亿美元。今年上半年 EESC 业务收入 9.54 亿美元，同比跌 1%，属于淡季正常业绩。今年下半年微软可能会发布 Xbox Scorpio 游戏主机，也有可能会对 AMD EESC 下半年业务带来刺激。

另外针对 AMD 将高性能处理器和 SoC 相关技术授权给与中科院背景的中科曙光子公司天津海光的 JV，AMD 预计会获得总额 2.93 亿美元的专利授权费，未来可能会从服务器销售中获得版税提成。我们分析国内人工智能产业，当前仍处于“春光乍现”的萌芽阶段，对比美国产业已经积累了大规模技术创新优势，中国在基础算法和理论研究方面还有相当差距，也决定了国内的数据服务、底层构架方面的发展仍然比较初期。因此国内人工智能产业在商业落地的上游需求方面，更加适应于 AMD 方案的高性价比策略。

AMD 的 EPYC 处理器已获百度合作支持，未来可以通过 GPU 的捆绑销售，加速切入国内 AI 发展快车道。我们看好依托中科院/中科曙光等国内芯片最高生产力和资源，AMD 在开拓中国人工智能上游芯片市场上时会拥有更为蓬勃的动力。

图 78：中科曙光股价变动

资料来源：Wind，天风证券研究所

公司 2018 年 PS 为 2.18x，对比英伟达 10x，我们预计公司今明年毛利率将进入 34-36%区间，并实现 Non-GAAP EPS 转正，重新走上盈利正轨。利用 GPU+CPU 异构计算技术储备的协同效应，自上而下的产品线羽翼渐丰，云计算和人工智能布局为 AMD 带来更高估值弹性。我们预测公司 2018 年营收/EPS 分别为 56.85 亿美元/0.35 美元，我们认为 2.65x PS 和 45x PE 较合理，重申“买入”评级，目标价维持 16 美元。

2.2.4. AMD 整体盈利预测

图 79：AMD 整体盈利预测

百万美元	2015 12/31/2015	2016 12/31/2016	2017E 12/31/2017	2018E 12/31/2018	2019E 12/31/2019
主营业务收入	**3,991.0**	**4,272.0**	**5,047.1**	**5,685.9**	**6,282.3**
同比%	-27.5%	7.0%	18.1%	12.7%	10.5%
计算与图形	1,805.0	1,967.0	2,419.4	2,782.3	3,088.4
同比%	-42.4%	9.0%	23.0%	15.0%	11.0%
企业、嵌入和半定制	2,186.0	2,305.0	2,627.7	2,903.6	3,194.0
同比%	-7.9%	5.4%	14.0%	10.5%	10.0%
Non-GAAP主营业务成本	2,875.0	2,932.0	3,331.1	3,599.2	3,863.6
GAAP毛利	**1,080.0**	**998.0**	**1,312.2**	**1,478.3**	**1,696.2**
毛利率%	27.1%	23.4%	26.0%	26.0%	27.0%
GAAP营业支出	**1,561.0**	**1,458.0**	**1,504.0**	**1,637.5**	**1,746.5**
GAAP研发费用	947.0	1008.0	1049.8	1154.2	1243.9
as % of sales	23.7%	23.6%	20.8%	20.3%	19.8%
GAAP管理费用	482.0	460.0	454.2	483.3	502.6
as % of sales	12.1%	10.8%	9.0%	8.5%	8.0%
无形资产摊销	3.0				
重组及其他费用	129.0	(10.0)			
许可权收入		(88.0)			
Non-GAAP毛利	**1,116.0**	**1,340.0**	**1,716.0**	**2,086.7**	**2,418.7**
毛利率%	28.0%	31.4%	34.0%	36.7%	38.5%
Non-GAAP营业支出	**1,369.0**	**1296.0**	**1463.7**	**1580.7**	**1696.2**
as % of sales	34.3%	30.3%	29.0%	27.8%	27.0%
	63.0	86.0	120.4	108.4	113.8
Non-GAAP营业利润（亏损）	**(253.0)**	**44.0**	**252.4**	**506.0**	**722.5**
营业利润率%	-6.3%	1.0%	5.0%	8.9%	11.5%
Stock-based compensation及其他	(241.0)	(380.0)			
所得税支出	-	-	49.3	98.9	159.7
			15.0%	17.0%	20.0%
Non-GAAP净利润	**-419.0**	**-117.0**	**127.1**	**331.1**	**486.8**
净利润率%	-10.5%	-2.7%	3.0%	5.8%	7.7%
同比%	-417.4%	-72.1%	208.6%	160.5%	47.0%
Non-GAAP 摊薄加权平均股数	783	835	940	940	940
Non-GAAP摊薄加权平均每股收益	**-0.54**	**-0.14**	**0.14**	**0.35**	**0.52**

资料来源：公司财报，天风证券研究所预测

2.3. GPU 需求和虚拟货币的关系：“微小但不是零的”

2017 年以来，数字虚拟货币连创新高，以太坊(Ethereum)技术下的以太币(ETH)涨逾 30 倍，比特币(BTC)也涨逾 7 倍突破 8000 美元。全球数字货币市值也从 180 亿美元增长至逾 2300 亿美元。8 月 1 日通过比特币硬分叉正式上线的比特币现金(BCH)随着生态化进程加速，迅速成为比特币、以太坊之后市值第三的虚拟数字货币。受益于数字货币的持续高度关注，通过显卡“挖矿”而获取货币的热潮，也发掘了对 AMD 和英伟达显卡的需求。

数字货币挖矿对 GPU 巨头的影响整体空间有限，Jensen 强调 5 次“微小但不是零的”也不为过

我们从 7 月以来便强调：数字货币挖矿的热潮虽出现行情延续，但整体空间有限。目前挖矿对显卡需求的驱动虽会持续存在但将进一步趋平。原因包括，1）遵循比特币挖矿路径，挖矿需求会向专门芯片矿机转移；2）以太币正在进行“工作量证明”向“权益证明”的升级，算力需求将会下降；3）挖矿市场的狂热需求也会影响正常游戏显卡市场的需求并带来二手卡问题，也不是英伟达和 AMD 所想见。

英伟达 CEO Jensen 在 Q3 季报会议上屡次被问及数字货币挖矿对公司业务的影响，他 5 次强调：挖矿市场对英伟达长期来说将会是“微小但不是零的”。Jensen 表示，随着旧有货币（例如比特币、莱特币）在体量上的不断扩大和挖矿算法的不断优化，挖矿会逐步转移至专业 ASIC 矿机，让 GPU 挖矿不再具备经济效益。然而，新生数字货币也不断的出现。在吸引矿工去挖掘的前提下，挖矿算法的难度也会不断增加。在此基础上 GPU 的需求可将继续。挖矿市场不断从旧有货币转移到新生货币上，类似今年的以太坊挖矿热潮替代了 2013 年的比特币、莱特币挖矿热潮，由此形成数字货币的挖矿生态循环。

挖矿虽然利好英伟达和 AMD，但我们认为不会改变他们长期业绩增长结构。挖矿对英伟达 Q3 贡献约 7000 万美元（对比 Q2 的 1.5 亿美元）。而 AMD 方面在 Q2 时由于显卡挖矿明显的经济效益一度出现“一卡难求”的局面，但并未公布具体数据（Q2 整体营收超预期 6000 万美元，Q3 超预期 1.4 亿美元），公司在 Q3 业绩发布会上表示 Q4 挖矿需求会进一步趋平。我们针对以太坊的 GPU 矿机需求进行测算，整体市场规模不足 10 亿美元。

区块链技术将长远影响商业行为，但数字货币价值的波动应不会对英伟达/AMD 带来持续业绩影响。挖矿虽利好英伟达/AMD，但鉴于空间较小，我们应该关注公司的经营基本面。

首先需要明确：目前数字货币市值第一大货币还是比特币，但比特币挖矿市场中，主要使用专业 ASIC 矿机（比特大陆蚂蚁矿机占据 70%的比特币矿机市场），GPU 挖矿已不具备经济效益。而以太币作为第二大市值货币，是当前 GPU 挖矿主要需求。我们针对以太坊的 GPU 矿机需求进行测算，整体市场规模不足 10 亿美元。

图 80：比特大陆 ASIC 蚂蚁矿机

资料来源：公司官网，天风证券研究所预测

数字货币价格的高波动性会为挖矿收益带来不确定性。目前以太币价格约在 320 美元附近，我们以 1 ETH=320 美元计，对以太币市值空间和 GPU 矿机的盈利情况进行详细拆分。(ETH 于 10 月完成应分叉，每 15 秒发行货币数从 5 ETH 减少为 3 ETH)。

图 81：以太币市值估计

以太坊市值TAM	
每天秒数	84,600
15秒发行量	5,760
单次发行货币数	3
以太坊发行总量/天	17,280
以太坊发行总量/年	6,307,200
以太坊价格/币（假设）	320
以太坊挖矿潜在价值（百万美元）	2,018
以100%投资回报计挖矿机市场（百万美元）	1,009
GPU成本占比	60%
GPU挖矿机潜在市场（百万美元）	605

资料来源：Cryptocompare，天风证券研究所预测

图 82：GPU 矿机盈利估计

GPU矿机盈利分拆（美元）	AMD RX 470	英伟达 GTX 970
Hash Rate（挖矿算力）	132 MH/s	
Blocktime（区块产生间隔时间）	15 s	
每天挖出ETH数量	0.058653	0.039102
每天挖出ETH价值	17.60	11.73
每年挖出ETH价值	6422.50	4281.67
GPU矿机成本分拆（美元）		
GPU单价	219	360
GPU总价（每台6块计）	1314	2160
成本占比%	54%	66%
内存、硬盘、主板、冷却系统等	906	906
成本占比%	37%	28%
供电	200	200
成本占比%	8%	6%
总成本	2420	3266
年化盈利	4002.50	1015.67
利润率	62%	24%

资料来源：Cryptocompare，天风证券研究所预测

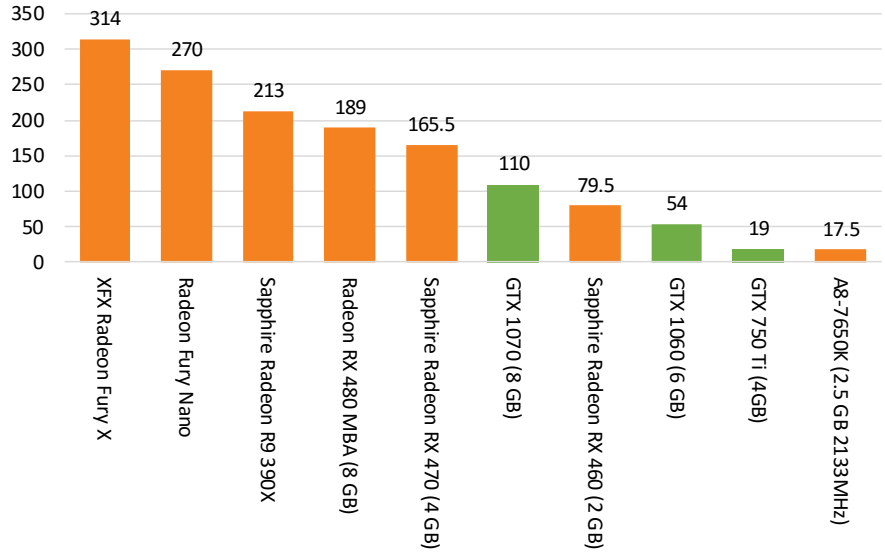
根据 cryptocompare 网站数据，我们详细对比 AMD 和英伟达 GPU 挖矿经济效益。可以看到 AMD RX 470 GPU 的矿机有明显优于英伟达 GTX 970 GPU 的经济回报，为了有效消弭挖矿和游戏需求冲突，并避免二手卡问题，英伟达针对虚拟数字货币挖矿热潮推出专门挖矿显卡（基于 GTX 1060 6GB 产品，完全取消显示输出接口，仅提供 90 天的质保）；AMD 则发布了专门的挖矿驱动 Radeon Software Crimson ReLive Edition Beta for Blockchain Compute，为区块链计算工作负荷优化性能。

2.3.1. 浅谈挖矿

比特币挖矿速率使用 Mhash/s 为单位，即每秒能够获得的哈希值量。每秒能够获得的哈希值量受算法（OpenCL 优于 CUDA）、显卡使用的图形核心种类、流处理器数量和频率等方

面影响。挖矿采用的是 SHA256 Hash 密码算法，几乎都是独立并发的整数计算，与 GPU 的大规模并发计算非常契合。其次 OpenCL 可以利用 GPU 片上大量的统一渲染架构(unified shader)来作为整数计算的资源。而 AMD 显卡相比 CPU 以及英伟达显卡有明显的性能优势，主要是同级别 A 卡流处理器资源数倍于 N 卡。

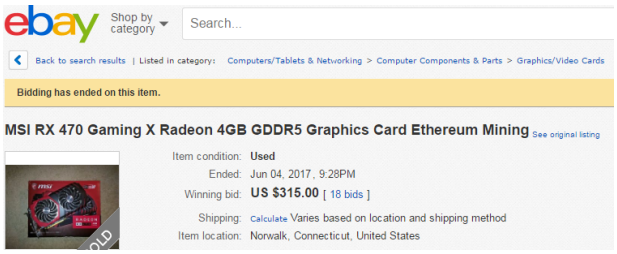
图 83：AMD 与英伟达显卡性能比较（单位：sol/s, solution per second）



资料来源：公司官网，天风证券研究所

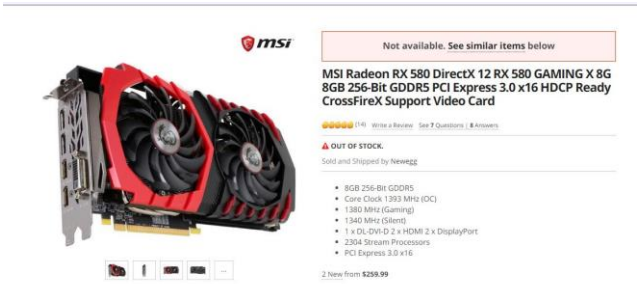
在今年二季度，挖矿热潮也让 AMD 新发布的 RX 500 系显卡供不应求。AMD RX 570 和 580 当时在多个国家市场断货，电商平台包括亚马逊、百思买、新蛋也出现大面积售罄。

图 84：Ebay 上二手微星 RX 470 曾被标价 315 美元高价



资料来源：Ebay，天风证券研究所

图 85：Newegg 上微星 RX 580 在 Q2 一度售罄



资料来源：Newegg，天风证券研究所

所以除了上文提到的数字货币共有的通用风险，随着以太坊代码和平台版本的升级转变，以太币逐步向 PoS 转型，也不会给 GPU 市场带来持续的需求增量。我们此前就强调，二季度对 AMD 显卡突然需求扩张，很大程度也是 RX 500 系列显卡刚刚发布，AMD 短期备货和库存不足所致，我们不看好真实需求的持续。

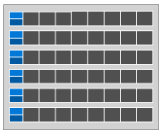
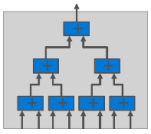
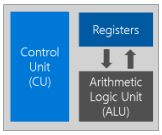
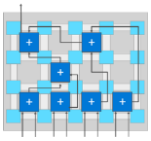
3. Google：TPU 以时间换吞吐量，破局者加速云端 AI 帝国

AI 芯片市场蛋糕越做越大，足以让拥有不同功能和定位的芯片和平共存，百家争鸣非零和博弈。“通用性和功耗的平衡”——在深度学习上游训练端（主要用在云计算数据中心里），GPU 是当仁不让的第一选择，ASIC 包括谷歌 TPU、寒武纪 MLU 等也如雨后春笋。而下游推理端更接近终端应用，需求更加细分，GPU 主流芯片之外，包括 CPU/FPGA/ASIC 也会在这个领域发挥各自的优势特点。

但我们需要强调，包括 TPU 在内的 ASIC 仍然面临通用性较弱，以及开发成本高企等局限。TPU 虽然理论上支持所有深度学习开发框架，但目前只针对 TensorFlow 进行了深度优化。另外 ASIC 芯片开发周期长和成本非常高，在开发调试过程中复杂的设计花费有时甚至会超过亿美元，因此需要谷歌这样的计算需求部署量才能将成本分摊到大量使用中。同时 ASIC 开发周期长，也可能会出现硬件开发无法匹配软件更新换代而失效的情况。

ASIC（Application Specific Integrated Circuit，专用集成电路）：细分市场需求确定后，以 TPU 为代表的 ASIC 定制化芯片（或者说针对特定算法深度优化和加速的 DSA, Domain-Specific-Architecture），在确定性执行模型(deterministic execution model)的应用需求中发挥作用。我们认为深度学习 ASIC 包括英特尔的 Nervana Engine、Wave Computing 的数据流处理单元、英伟达的 DLA、寒武纪的 NPU 等逐步面市，将依靠特定优化和效能优势，未来在深度学习领域分一杯羹。

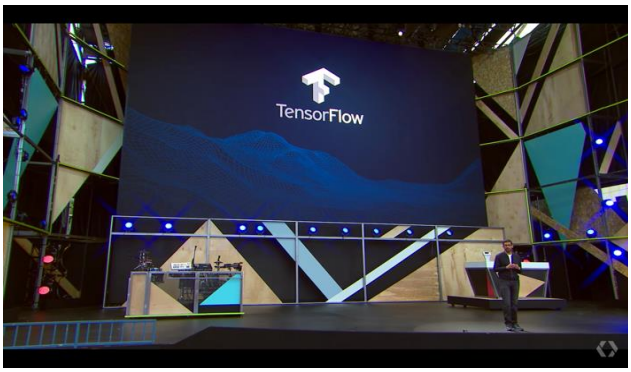
图 86：目前深度学习领域常用的四大芯片类型，“通用性和功耗的平衡”

	训练端	推理端	
	**GPU**：以英伟达为主，AMD为辅标榜通用性，多维计算及大规模并行计算架构契合深度学习的需要。在深度学习上游训练端（主要用在云计算数据中心里），GPU是当仁不让的第一选择。	**GPU**：英伟达Volta GPU也开始布局推理端。深度学习下游推理端虽可容纳CPU/FPGA/ASIC等芯片，但竞争态势中英伟达依然占主导。	
	**ASIC**：以谷歌的TPU、英特尔的Nervana Engine为代表，针对特定框架进行深度优化定制。但开发周期较长，通用性较低。比特币挖矿目前使用ASIC专门定制化矿机。	**ASIC**：下游推理端更接近终端应用，需求也更加细分，英伟达的DLA，寒武纪的NPU等逐步面市，将依靠特定优化和效能优势，未来在深度学习领域分一杯羹。	
	**CPU**：通用性强，但难以适应于人工智能时代大数据并行计算工作。	**FPGA**：依靠可编程性及电路级别的通用性，适用于开发周期较短的IoT产品、传感器数据预处理工作以及小型开发试错升级迭代阶段等。但较成熟的量产设备多采用ASIC。	

资料来源：微软 Build，天风证券研究所整理

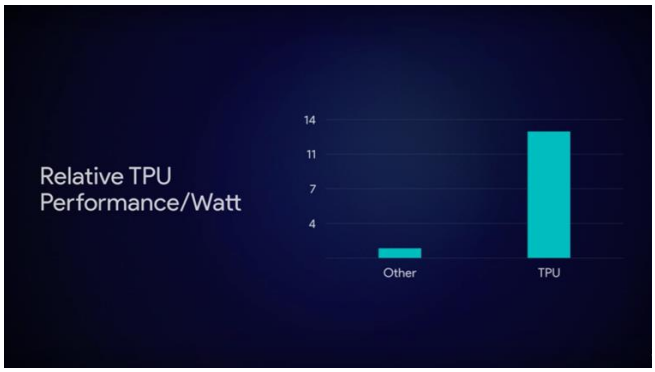
在推理阶段，由于更接近终端应用需求，更关注响应时间而不是吞吐率。由于 CPU 和 GPU 结构设计更注重平均吞吐量(throughout)的 time-varying 优化方式，而非确保延迟性能。谷歌设计了一款为人工智能运算定制的硬件设备，张量处理单元(Tensor Processing Unit, TPU)芯片，并在 2016 年 5 月的 I/O 大会上正式展示。

图 87：皮查伊在 2016 I/O 大会上介绍 TensorFlow



资料来源：谷歌 2016 I/O 大会现场照片，天风证券研究所

图 88：皮查伊介绍 TPU 性能对比



资料来源：谷歌 2016 I/O 大会现场照片，天风证券研究所

第一代 TPU 的确定性执行模型(deterministic execution model)针对特定推理应用工作，更好的匹配了谷歌神经网络在推理应用 99%的响应时间需求。第一代 TPU 是在一颗 ASIC 芯片上建立的专门为机器学习和 TensorFlow 量身打造的集成芯片。该芯片从 2015 年开始就已经在谷歌云平台数据中心使用，谷歌表示 TPU 能让机器学习每瓦特性能提高一个数量级，相当于摩尔定律中芯片效能往前推进了七年或者三代。

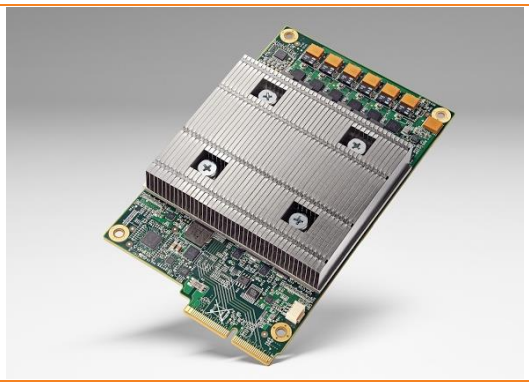
谷歌表示，这款芯片目前不会开放给其他公司使用，而是专门为 TensorFlow 所准备。TPU 的主要特点是：

1、从硬件层面适配 TensorFlow 深度学习系统，是一款定制的 ASIC 芯片，谷歌将 TPU 插放入其数据中心机柜的硬盘驱动器插槽里来使用；

2、数据的本地化，减少了从存储器中读取指令与数据耗费的大量时间；

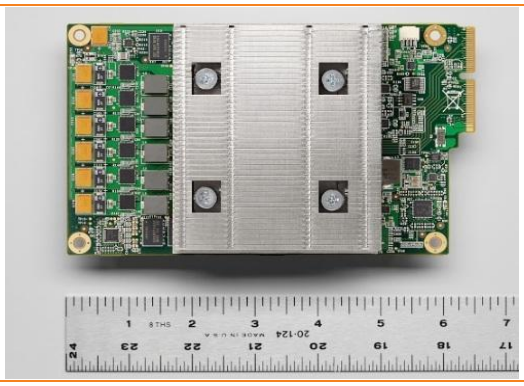
3、芯片针对机器学习专门优化，尤其对低运算精度的容忍度较高，这就使得每次运算所动用的晶体管数量更少，在同时间内通过芯片完成的运算操作也会更多。研究人员就可以使用更为强大的机器学习模型来完成快速计算。

图 89：谷歌第一代 TPU 电路板



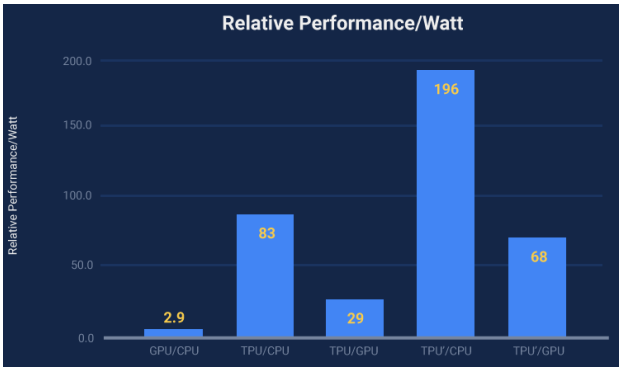
资料来源：谷歌研究所官方博客，天风证券研究所

图 90：谷歌第一代 TPU 尺寸示意图



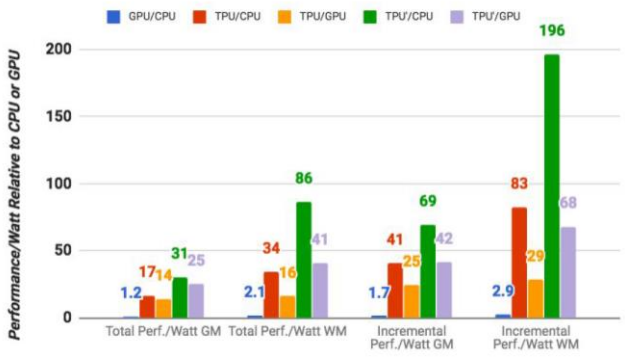
资料来源：谷歌研究所官方博客，天风证券研究所

图 91：TPU 的性能/功耗比较优势



资料来源：公司官网，天风证券研究所

图 92：TPU 的性能/功耗比较优势



资料来源：公司官网，天风证券研究所

自 2016 年以来，TPU 运用在人工智能搜索算法 RankBrain、搜索结果相关性的提高、街景 Street View 地图导航准确度提高等方面。在 I/O 大会上，皮查伊顺带提到了 16 年 3 月份行的举世瞩目人机大战里，在最终以 4:1 击败围棋世界冠军李世石的 AlphaGo 身上，谷歌也使用了 TPU 芯片。

谷歌把：

1、2015 年击败初代击败樊麾的版本命名为 AlphaGo Fan，这个版本的 AlphaGo 运行于谷歌云，分布式机器使用了 1202 个 CPU 和 176 个 GPU。

2、去年击败李世石的版本 AlphaGo Lee 则同样运行于云端，但处理芯片已经简化为 48 个

第一代 TPU。

3、今年击败柯洁的 Master 以及最新版本 Zero 则通过单机运行，只在一个物理服务器上部署了 4 个第一代 TPU。（AlphaGo 的背后算法详解，可参见我们此前的深度报告《谷歌人工智能：从 HAL 的太空漫游到 AlphaGo，AI 的春天来了》）

图 93：AlphaGo 版本进化

版本	AlphaGo Fan	AlphaGo Lee	AlphaGo Master	AlphaGo Zero
时间	2015年10月	2016年3月	2017年5月	2017年10月
使用芯片	运行于谷歌云，分布式机器使用1202个CPU和176个GPU。	运行于谷歌云，使用48个第一代TPU。	单机运行，只在一个物理服务器上部署4个第一代TPU。	单机运行，只在一个物理服务器上部署4个第一代TPU。

资料来源：DeepMind 官网，天风证券研究所

3.1. 谷歌以 TPU 为破局者，软硬兼施，加速云端 AI 帝国

AI 芯片领域数据中心市场空间巨大，我们看到市场主流 GPU 之外，谷歌破局者之态依靠 TPU 2.0 的浮点运算升级自下而上进入云计算服务。谷歌当下不直接销售硬件，但将 TPU 部署在云计算中以云服务形式进行销售共享，在为数据中心加速市场带来全新的需求体验的同时，可进一步激活中小企业的云计算需求市场，另辟 AWS、Azure 之外蹊径。我们长期看好谷歌基于公司 AI First 战略规划打造 AI 开发软硬件一体化开发帝国。

不过 TPU 虽然理论上支持所有深度学习开发框架，但目前只针对 TensorFlow 进行了深度优化。而英伟达 GPU 支持包括 TensorFlow、Caffe 等在内所有主流 AI 框架。因此谷歌还在云计算平台上提供基于英伟达 Tesla V100 GPU 加速的云服务。在开发生态方面，TensorFlow 团队公布了 TensorFlow Research Cloud 云开发平台，向研究人员提供一个具有 1000 个云 TPU 的服务器集群，用来服务各种计算密集的研究项目。

图 94：TPU Pod 由 64 台第二代 TPU 组成

资料来源：公司官网，天风证券研究所

图 95：TensorFlow Research Cloud 云开发平台

TensorFlow
RESEARCH CLOUD

资料来源：公司官网，天风证券研究所

3.2. 第一代 TPU：脉动阵列“获新生”，以时间换吞吐量

第一代 TPU 面向的推理阶段，由于更接近终端应用需求，更关注响应时间而不是吞吐率。相对于 CPU 和 GPU 结构设计更注重平均吞吐量(throughout)的 time-varying 优化方式，而非确保延迟性能。第一代 TPU 的确定性执行模型(deterministic execution model)针对特定

推理应用工作，更好的匹配了谷歌神经网络在推理应用上 99%的响应时间需求。由于 TPU 没有任何存储程序，仅执行从主机发送的指令，这些功能的精简让 TPU 有效减小芯片面积并降低功耗。

谷歌在今年 4 月的体系结构顶会 ISCA 2017 上面，发布了一篇介绍 TPU 相关技术以及与其它硬件比较的论文，并被评为最佳论文。我们通过论文得以看到第一代 TPU 的设计思路以及性能比较。

第一代 TPU 从 2015 年开始就被使用在谷歌云计算数据中心的机器学习应用中，面向的是推理阶段。首先看性能比较（鉴于 2016 年以前大部分机器学习公司主要使用 CPU 进行推理，谷歌在论文中 TPU 的比较对象产品为英特尔服务器级 Haswell CPU 和英伟达 Tesla K80 GPU），谷歌表示：

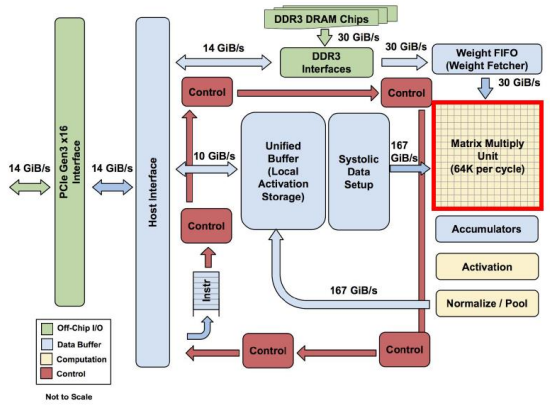
1、 针对自身产品的人工智能负载，推理阶段，TPU 处理速度比 CPU 和 GPU 快 15-30 倍；

2、 TPU 的功耗效率（TOPS/Watt，万亿次运算/瓦特）也较传统芯片提升了 30-80 倍；

3、 基于 TPU 和 TensorFlow 框架的神经网络应用代码仅需 100-1500 行。

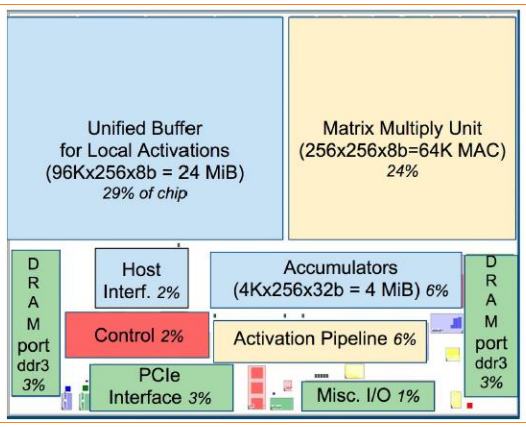
基于在成本-能耗-性能(cost-energy-performance)上的提升目标，TPU 的设计核心是一个 65,536(256x256)个 8 位 MAC 组成的矩阵乘法单元(MAC matrix multiply unit)，可提供峰值达到 92 TOPS 的运算性能和一个高达 28 MiB 的软件管理片上内存。TPU 的主要设计者 Norman Jouppi 表示，谷歌硬件工程团队最开始考虑过 FPGA 的方案，实现廉价、高效和高性能的推理解决方案。但是 FPGA 的可编程性带来的是与 ASIC 相比在性能和每瓦特性能的巨大差异。

图 96：第一代 TPU 各模块的框图，红框为核心矩阵乘法单元



资料来源：公司官网，天风证券研究所

图 97：第一代 TPU 的芯片布局图



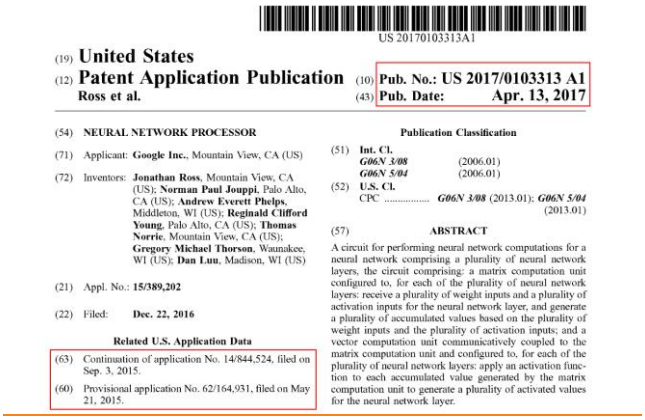
资料来源：公司官网，天风证券研究所

从上图我们看到，TPU 的核心计算部分是右上方的黄色矩阵乘法单元(Matrix Multiply unit)，输入部分是蓝色的加权 FIFO 和一致缓冲区(Unified Buffer，输出部分是蓝色的累加器(Accumulators)。在芯片布局图中我们看到，蓝色的缓存的面积占 37%，黄色的计算部分占 30%，红色的控制区域只占 2%，一般 CPU、GPU 的控制部分会更大而且难以设计。

我们深挖谷歌 TPU 论文，在参考文献中提及了谷歌申请的专利，核心的专利 Neural Network Processor 作为总构架在 2015 年就已提交，并在 2016 年公开（后续专利在 2017 年 4 月公开，专利号：US 2017/0103313，即下图 96 所示），同时还包括了几个后续专利：如何在该构架上进行卷积运算、矢量处理单元的实现、权重的处理、数据旋转方法以及 Batch 处理等。

专利摘要概述：一种可以在多网络层神经网络中执行神经网络计算的电路，包括一个矩阵运算单元(matrix computation unit)：对多个神经网络层中的每一层，可以被配置为接收多个 weights 输入和多个 activation 输入，并对应生成多个累积值；以及矢量运算单元(vector computation unit)，其通信耦合到所述矩阵运算单元。

图 98：TPU 论文核心专利：Neural Network Processor



US 20170103313A1

(19) United States

(12) Patent Application Publication

Ross et al.

(10) Pub. No.: US 2017/0103313 A1

(43) Pub. Date: Apr. 13, 2017

(54) NEURAL NETWORK PROCESSOR

(71) Applicant: Google Inc., Mountain View, CA (US)

(72) Inventors: Jonathan Ross, Mountain View, CA (US); Norman Paul Jouppi, Palo Alto, CA (US); Andrew Everett Phelps, Middleton, WI (US); Reginald Clifford Young, Palo Alto, CA (US); Thomas Norrie, Mountain View, CA (US); Gregory Michael Thorson, Waunakee, WI (US); Dan Luu, Madison, WI (US)

(21) Appl. No.: 15/389,202

(22) Filed: Dec. 22, 2016

Related U.S. Application Data

(63) Continuation of application No. 14/844,524, filed on Sep. 3, 2015.

(60) Provisional application No. 62/164,931, filed on May 21, 2015.

Publication Classification

(51) Int. Cl.
G06N 3/08 (2006.01)
G06N 5/04 (2006.01)

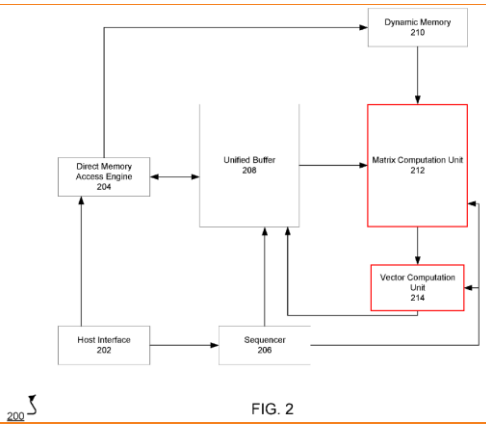
(52) U.S. Cl.
CPC G06N 3/08 (2013.01); G06N 5/04 (2013.01)

(57) ABSTRACT

A circuit for performing neural network computations for a neural network comprising a plurality of neural network layers, the circuit comprising: a matrix computation unit configured to, for each of the plurality of neural network layers: receive a plurality of weight inputs and a plurality of activation inputs for the neural network layer, and generate a plurality of accumulated values based on the plurality of weight inputs and the plurality of activation inputs; and a vector computation unit communicatively coupled to the matrix computation unit and configured to, for each of the plurality of neural network layers: apply an activation function to each accumulated value generated by the matrix computation unit to generate a plurality of activated values for the neural network layer.

资料来源：Google TPU 专利，天风证券研究所

图 99：第一代 TPU 各模块设计原理专利，核心为矩阵运算单元和矢量运算单元

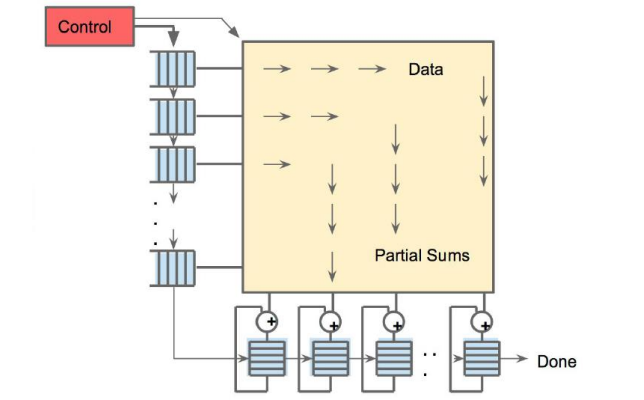


资料来源：Google TPU 专利，天风证券研究所

TPU 的设计思路比 GPU 更接近一个浮点运算单元，是一个直接连接到服务器主板的简单矩阵乘法协处理器。TPU 上的 DRAM 是作为一个独立的并行单元，TPU 类似 CPU、GPU 一样是可编程的，并不针对某一特定神经网络设计的，而能在包括 CNN、LSTM 和大规模全连接网络(large, fully connected models)上都执行 CISC 指令。只是在编程性上 TPU 使用矩阵作为 primitive 对象，而不是向量或标量。TPU 通过两个 PCI-E 3.0 x8 边缘连接器连接协处理器，总共有 16 GB/s 的双向带宽。

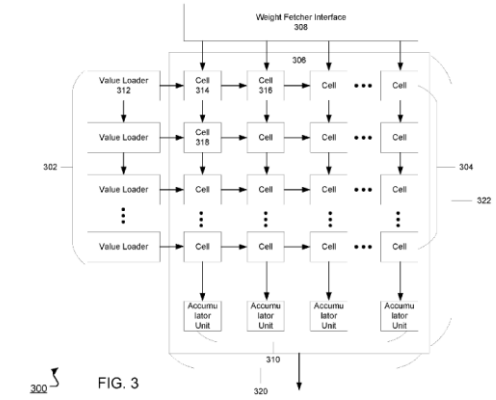
我们看到，TPU 的 matrix 单元就是一个典型的脉动阵列架构(systolic array computers)。weight 由上向下流动，activation 数据从左向右流动。控制单元实际上就是把指令翻译成控制信号，控制 weight 和 activation 如何传入脉动阵列以及如何在脉动阵列中进行处理和流动。由于指令比较简单，相应的控制也是比较简单的。

图 100：矩阵乘法单元的脉动数据流(Systolic data flow)



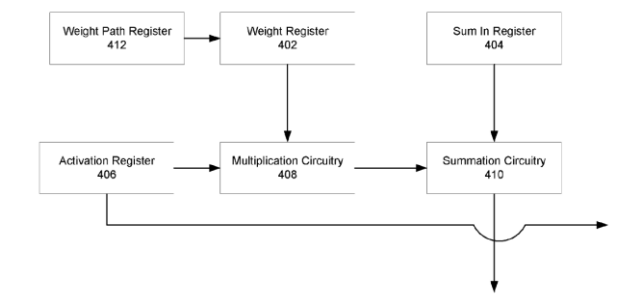
资料来源：公司官网，天风证券研究所

图 101：矩阵运算单元的架构原理图



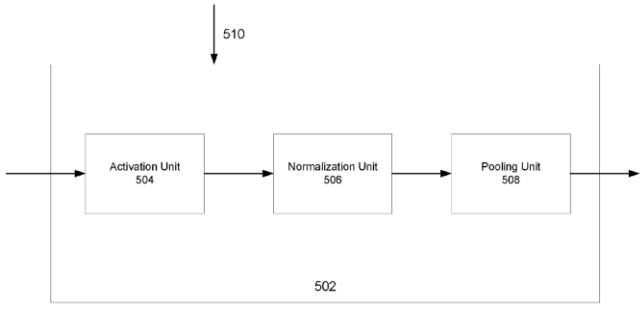
资料来源：Google TPU 专利，天风证券研究所

图 102：矩阵运算单元中一个 Cell 的架构



资料来源：Google TPU 专利，天风证券研究所

图 103：矢量运算单元的架构原理图



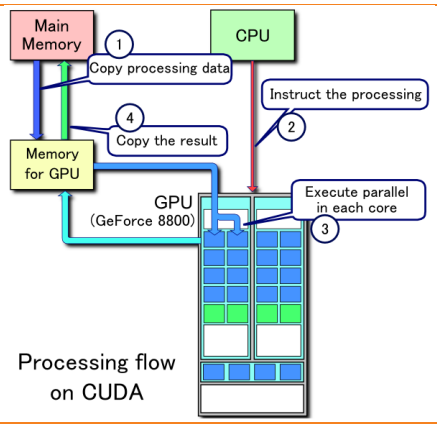
资料来源：Google TPU 专利，天风证券研究所

从性能上，脉动阵列架构在大多数 CNN 卷积操作上效率很好，但在部分其他类型的神经网络操作上，效率不是太高。另外脉动阵列架构在上世纪 80 年代就已经被提出，Simple and regular design 是脉动阵列的一个重要原则，通过简单而规则的硬件架构，提高芯片的设计和实现的能力，从而尽量发挥软件的能力，并平衡运算和 I/O 的速度。脉动阵列解决了传统计算系统：数据存取速度往往大大低于数据处理速度的问题，通过让一系列在网格中规律布置的处理单元(Processing Elements，PE)，进行多次重用输入数据来在消耗较小的带宽的情况下实现较高的运算吞吐率。但是脉动阵列需要带宽的成比例的增加来维持所需的加速倍数，所以可扩展性问题仍待解决。

图 104：英伟达 GeForce GTX 1070 Ti 模块框图

资料来源：公司官网，天风证券研究所

图 105：CUDA 核心计算处理流程图



资料来源：Wikipedia，天风证券研究所

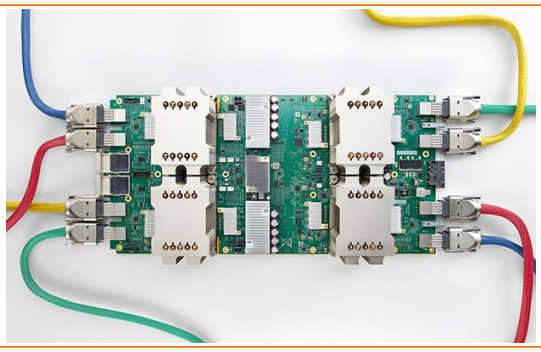
对比 GPU 的硬件架构，英伟达的游戏显卡 GeForce GTX 1070 Ti 使用的是 Pascal 架构 16 纳米制程，主频 1,607MHz，拥有 2,432 个 CUDA 核心和 152 个纹理单元，2 MB L2 cache，功耗 180 W，8GB GDDR5 内存。英伟达 GPU 的核心计算单元 CUDA 核心专为同时处理多重任务而设计，数千个 CUDA 核心组成了 GPU 的大规模并行计算架构。而在计算过程中，主要计算流程为：1）从主机内存将需要处理的数据 read 到 GPU 的内存；2）CPU 发送数据处理执行给 GPU；3）GPU 执行并行数据处理；4）将结果从 GPU 内存 write 到主机内存。通过编译优化把计算并行化分配到 GPU 的多个 core 里面，大大提高了针对一般性通用需求的大规模并发编程模型的计算并行度。

3.3. 第二代 TPU：可进行深度学习上游训练计算

第二代 TPU，又名 Cloud TPU，能够同时应用于高性能计算和浮点计算，峰值性能达到 180 TFLOPS/s。与第一代 TPU 只能应用于推理不同，**第二代 TPU 还可以进行深度学习上游训练环节**。随着第二代 TPU 部署在 Google Compute Engine 云计算引擎平台上，谷歌将 TPU 真正带入云端。

谷歌在今年 5 月 17 日举办了 2017 年度 I/O 开发者大会。一场并未有太多亮点的大会上，谷歌 CEO 皮查伊继续强调公司 AI First 的传略规划。最为振奋人心的当属第二代 TPU——Cloud TPU 的发布。

图 106：第二代 TPU 包含 4 个芯片



资料来源：siliconangle，天风证券研究所

图 107：第二代 TPU 包含 4 个芯片

资料来源：siliconangle，天风证券研究所

谷歌同时发布了 TPU Pod，由 64 台第二代 TPU 组成，算力达 11.5 petaflops。谷歌表示 1/8 个 TPU Pod 在对一个大型机器翻译模型训练的只需要 6 个小时，训练速度是市面上 32 块性能最好的 GPU 的 4 倍。

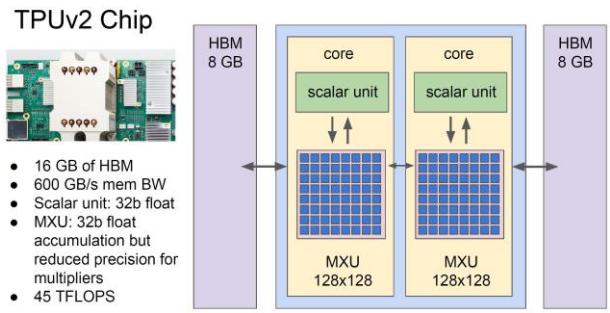
谷歌此前强调，第一代 TPU 是一款推理芯片，并不用作神经网络模型训练阶段，训练学习阶段的工作仍需交由 GPU 完成。早在去年 I/O 大会上公布 TPU 之前，谷歌就已经将 TPU 应用在各领域任务中，包括：图像搜索、街景、谷歌云视觉 API、谷歌翻译、搜索结果优化以及 AlphaGo 的围棋系统中。

而这次第二代 TPU 的升级，自下而上的进入深度学习上游，应用在图像和语音识别，机器翻译和机器人等领域，加速对单个大型机器学习模型的训练。第二代 TPU 在左右两侧各有四个对外接口，左侧还有两个额外接口，未来可能允许 TPU 芯片直接连接存储器，或者是高速网络，实现更加复杂的运算以及更多的扩展功能。在半精度浮点数(FP16)情况下，第二代 TPU 的单芯片可以达到 45 Teraflops（每秒万亿次的浮点运算），4 芯片的设计能达到 180 Teraflops。(对比第一代 TPU 算力：8 位整数运算达 92 TOPS，16 位整数运算达 23 TOPS)

图 108：TPU Pod，由 64 台 TPU 组成，算力达 11.5 petaflops

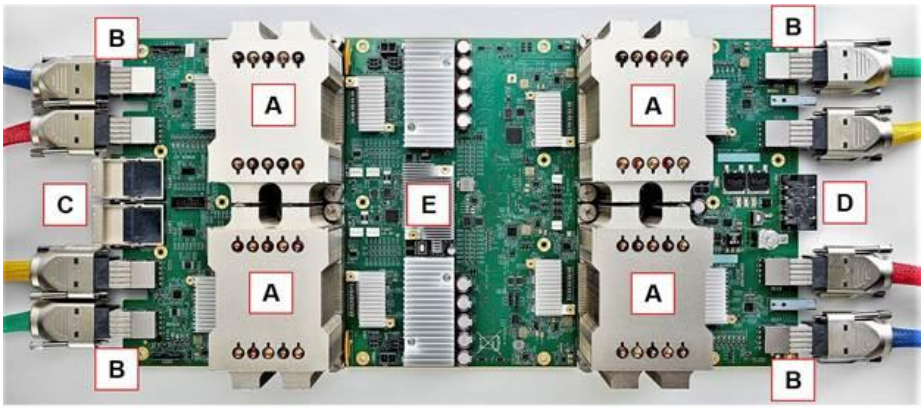
资料来源：公司官网，天风证券研究所

图 109：第二代 TPU 使用了 16 GB HBM 内存



资料来源：servethehome，天风证券研究所

图 110：A 是第二代 TPU 及散热片，B 是每块 TPU 的 2 根 BlueLink 25GB/s 电缆，C 是 Omni-Path 架构(OPA)电缆接口，D 是电源连接器背面，E 可能为网络交换机

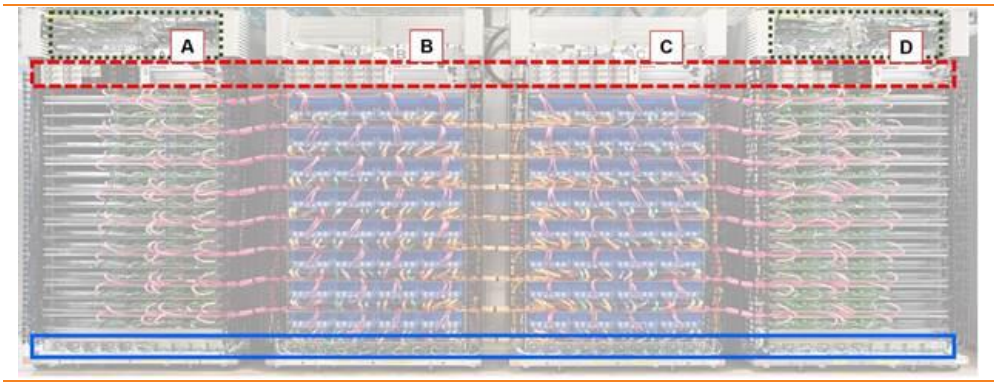


资料来源：The Next Platform，天风证券研究所

对 TPU Pod 的结构进行简要分析，四机架的镜像结构包含 64 个 CPU 板和 64 个第二代 TPU 板，The Next Platform 推测 CPU 板是标配英特尔 Xeon 双插槽主板，因此整个 Pod 机柜包括 128 个 CPU 芯片和 256 个 TPU 芯片。

The Next Platform 认为，谷歌使用两条 OPA 线缆将每块 CPU 板一一对应连接至 TPU 板，使得 TPU 与 CPU 的使用比例为 2:1，这种 TPU 加速器与处理器之间高度耦合的结构，与典型的深度学习加速结构中 GPU 加速器 4:1 或 6:1 的比例不太一样，更强调了 TPU 作为协处理器的设计理念——CPU 处理器还是需要完成大量的计算工作，只是把矩阵计算的的任务卸载到 TPU 中完成。

图 111：A 和 D 是 CPU 机架，B 和 C 是 TPU 机架，蓝色方框为不间断电源(UPS)，红色方框为电源，右上角绿色方框为网络交换机顶部



资料来源：The Next Platform，天风证券研究所

3.4. 谷歌重申买入：人工智能巨头新征途——云+YouTube+硬件

我们早在年初已经开始强调，人工智能巨头新征途——云+YouTube+硬件。YouTube & 云计算的巨大增长动力将是谷歌持续转型的助推器，长期看好 AI 和 Other Bets 创新业务厚积薄发。

3Q17 营收 277.7 亿美元，同比涨 24%，高于华尔街预期 219 亿美元，主要鉴于移动端广告搜索业务和 YouTube 的增长。EPS 9.57 美元，高于预期 8.31 美元。广告业务营收 240.7 亿美元，同比涨 21%，其他业务包括云计算和硬件销售达 34.1 亿美元，同比大涨 40%（尚未囊括 10 月发布的 Pixel 2 等新产品销售收入）。新兴业务 Other Bets 营收同比涨 53%至 3.02 亿美元，但亏损环比略涨至 8.12 亿美元。

核心广告指标 Cost per click 实现环比转正，移动端转型之势给予市场极大信心。谷歌股价 3 季度跑输大盘，外部压力包括欧盟审查、美国选举操控等舆论监管压力。我们认为虽然在情绪面上承压，但对公司业绩基本面影响有限。根据彭博一致预期 2018 年 EPS 41.46 美元，给予 31x PE，目标价从 1200 上调至 1300 美元，重申“买入”评级。

YouTube 百般武艺冲劲十足，移动端积极转型执行力坚决

YouTube 成长继续保持蓬勃动力，Pichai 表示用户通过电视观看 YouTube 的总时长达到 1 亿小时/日，同比剧增 70%。YouTube TV 网络电视服务超过 30 个城市，包括 40 个电视台节目的打包订阅费 35 美元/月，仅为有线电视订阅均价的一半。根据 eMarketer 预测，2017 年美国视频广告市场增速强劲，整体规模预计增长 23.7%至 132.3 亿美元，YouTube 作为龙头将贡献 21.7%约 28.7 亿美元。

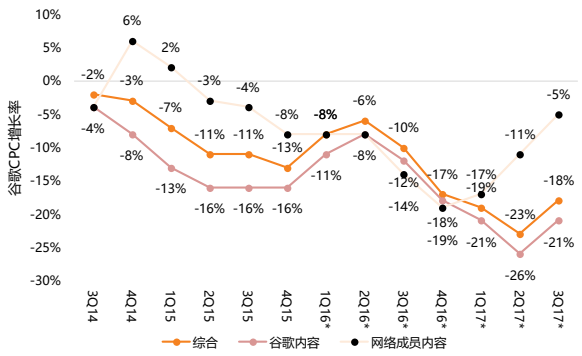
广告营收向移动设备转移步伐扎实，广告业务净营收增速回升至 21%，广告业务指标 Cost per click 同比降-18%，对比 Q2 的-23%和 Q1 的-19%，但 16 年以来环比首现转正。Paid clicks 同比涨 47%，对比 Q2 的 52%和 Q1 的 44%，自由网站尤其是 YouTube 极大拉动用户点击意愿。我们强调，在移动端获取搜索流量的成本会高于 PC 端，谷歌需要向包括 iPhone 在内的合作伙伴支付更多的流量获取成本和收入分成，谷歌已证明在移动广告上拥有不逊于 Facebook 的市场执行力。

谷歌是人工智能的龙头标的

我们长期看好人工智能，发力语音识别和无人驾驶：我们认为语音识别技术已经足够进入普及。DeepMind 成为谷歌 AI 的标签门面，看好进一步实现前瞻 AI 技术与现有业务的有效整合。

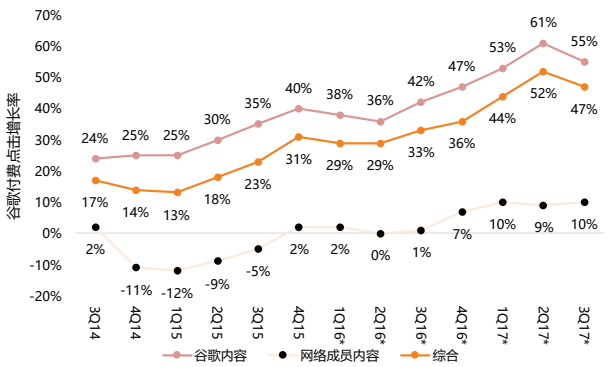
C 端谷歌软硬兼施，Pixel 手机+Home 音箱+Assistant AI 助理打造 AI 生态圈，探索人机交互便捷方式和广告业务协同效应。9 月以 11 亿美元收购 HTC 打造 Pixel 手机的团队。无人驾驶业务 Waymo 初试共享经济，领投 Lyft 把握用户入口将成为未来布局关键。

图 112：谷歌 Cost-per-click 增长率（*号为算法调整后）



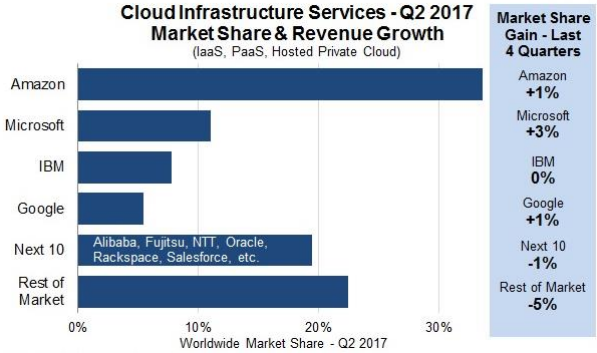
资料来源：谷歌财报，天风证券研究所整理

图 113：谷歌 Paid clicks（*号为算法调整后）



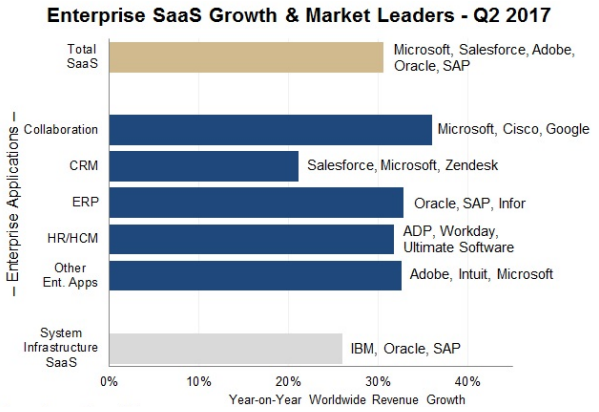
资料来源：谷歌财报，天风证券研究所整理

图 114：全球云计算市场竞争格局



资料来源：Synergy Research，天风证券研究所

图 115：全球云计算企业 SaaS 市场格局



资料来源：Synergy Research，天风证券研究所

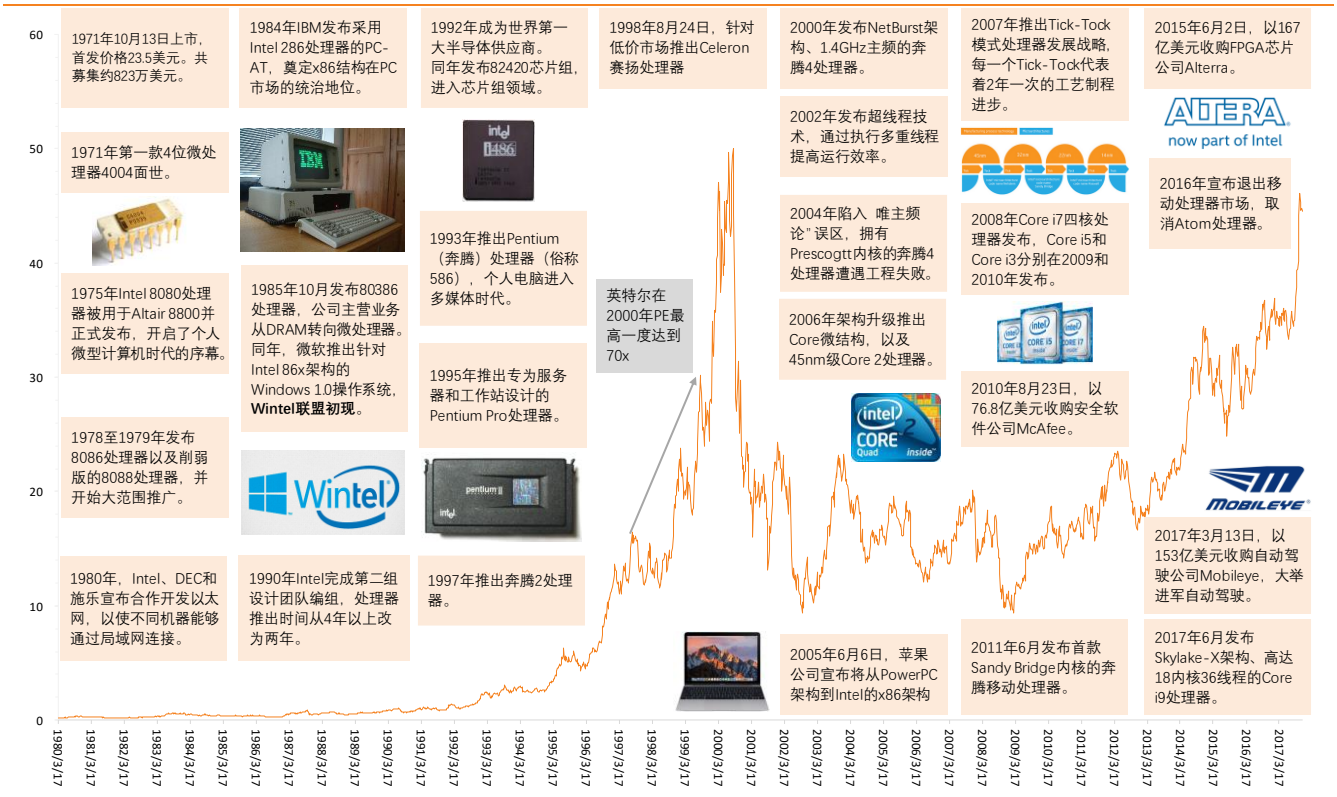
4. 英特尔：老巨头，MAN(Mobileye+Altera+Nervana) up！

英特尔作为 PC 时代无可争议的霸主，进入人工智能时代也在思考如何顺应行业变革。事实上，根据 CB Insights 的统计，Intel Capital 是过去 5 年全球 AI 投资最激进的投资机构之一。英特尔最近几年比较知名的收购包括 FPGA 芯片巨头 Altera、深度学习创业公司 Nervana、无人驾驶行业领导者 Mobileye、机器视觉芯片厂商 Movidius 等，英特尔希望基于自身 Xeon、Xeon Phi 处理器的硬件平台优势，通过大举收购提供端到端的全栈实力，从硬件、库和语言、框架、工具到应用方案，来向全球人工智能市场提供端到端的人工智能解决方案。

但是身为 1968 年就创立的“老巨头”，虽然与微软组成的 Wintel 联盟在 PC 市场呼风唤雨 20 年，“巨头”的标签本身也是一种桎梏，我们看到了英特尔在移动互联网时代的挣扎。今年 3 月，英特尔公司宣布将公司旗下所有人工智能相关业务整合到一起，成立人工智能产品事业部(AIPG)，在公司内部打造更连贯统一的 AI 开发生态环境。我们长期看好英特尔打造的端到端全栈解决方案策略，从最底层的云端数据中心服务器为“开端”，向消费级产品包括智能手机、笔记本电脑、无人机和 IoT 市场的“前端”延伸布局。

不过如何提高执行能力，真正把握人工智能时代的市场节奏，还让我们拭目以待。

图 116：英特尔历史大事件



资料来源：Yahoo Finance，天风证券研究所整理，数据截止至 2017 年 11 月 27 日，采用调整后收盘价

4.1. 收购 Nervana 挑战深度学习上游

2016 年 8 月，英特尔以约 3.5 亿美元收购深度学习企业 Nervana Systems，以获取 Nervana 的软件、云计算服务和硬件技术，直接参与到 AI 芯片的竞争中。英特尔希望能够将 Nervana 专门针对深度学习开发的 ASIC 特制芯片 Nervana Engine 整合到 Xeon 处理器中，实现性能加速。

Nervana 由前高通公司研究员 Naveen Rao（现任英特尔人工智能部门主管）于 2014 年创建，是 Android 之父 Andy Rubin 旗下硬件孵化器 Playground Global 的公司，Nervana 曾与美国情报委员会的风险投资部门 In-Q-tel 签署合作协议，美国国家能源研究科学计算中

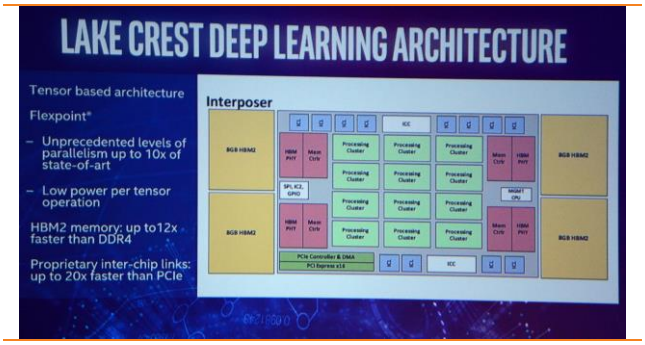
心也正在使用由 Nervana 开发的深度学习软件 Neon 的相关服务。

Nervana Engine 的设计仅保留针对机器学习优化的必要结构，取消了类似 GPU 中的缓存管理结构，含有 12 个双向高带宽链路(12 bi-directional high-bandwidth links)，使得芯片到底层构架之间都能无缝互连，优化可扩展能力，针对线性代数进行加速。

Nervana Engine 设计 8 个 ASIC 以如下所示的环式结构互连，并组成一个完整的加速处理器，每个计算节点都具备独立的储存界面，拥有更好的模型扩展性和数据存储 I/O 能力，实现近 8 倍的线性性能加速。

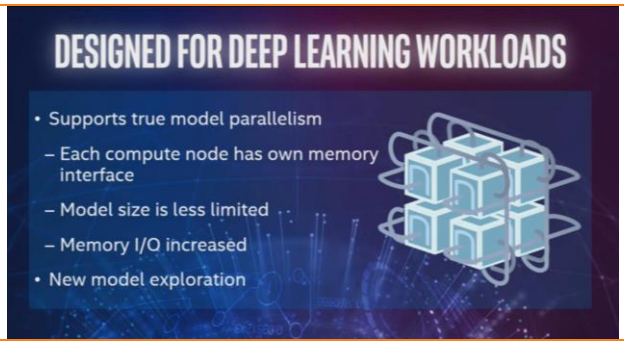
Nervana Engine 被纳入英特尔旗下后更名为 Lake Crest，使用的是台积电 28 纳米工艺，基于张量处理构架，使用 32 GB HBM2 高带宽显存，以及 8 Tera-bits 每秒的处理速度。英特尔计划在今年上半年对 Nervana Engine 进行测试，并在今年下半年向核心客户提供样片。

图 117：英特尔 Lake Crest 深度学习芯片构架



资料来源：公司官网，天风证券研究所

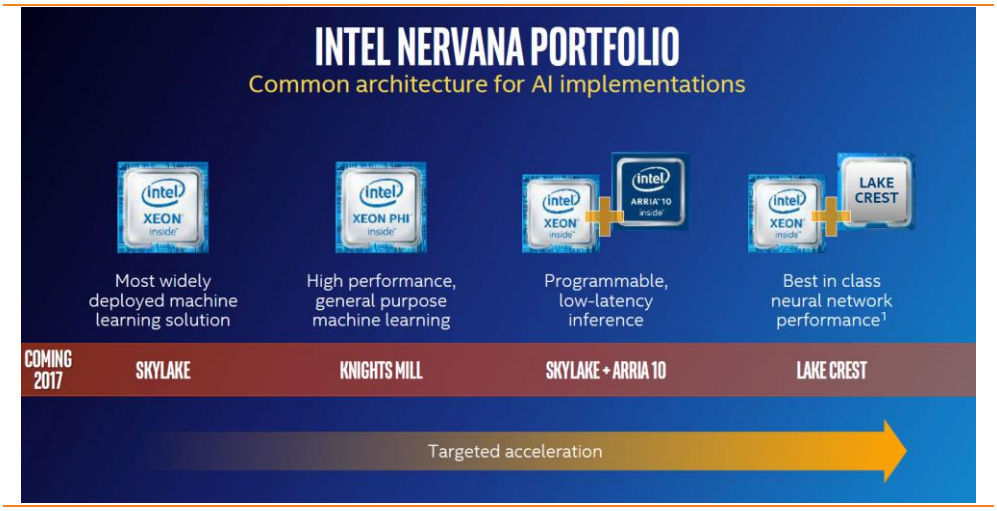
图 118：针对深度学习 workloads 设计的 Nervana Engine



资料来源：公司官网，天风证券研究所

Lake Crest 之后，英特尔正在开发 Knights Crest，将 Xeon 处理器和 Nervana 的工艺技术进行整合，更好地实现在深度学习上游模型训练端的优化。英特尔希望在 2020 年实现较现在计算能力 100 倍的提升。

图 119：英特尔的 Nervana AI 构架布局路线



资料来源：公司官网，天风证券研究所

4.2. 收购 Mobileye 打造自动驾驶新巨头

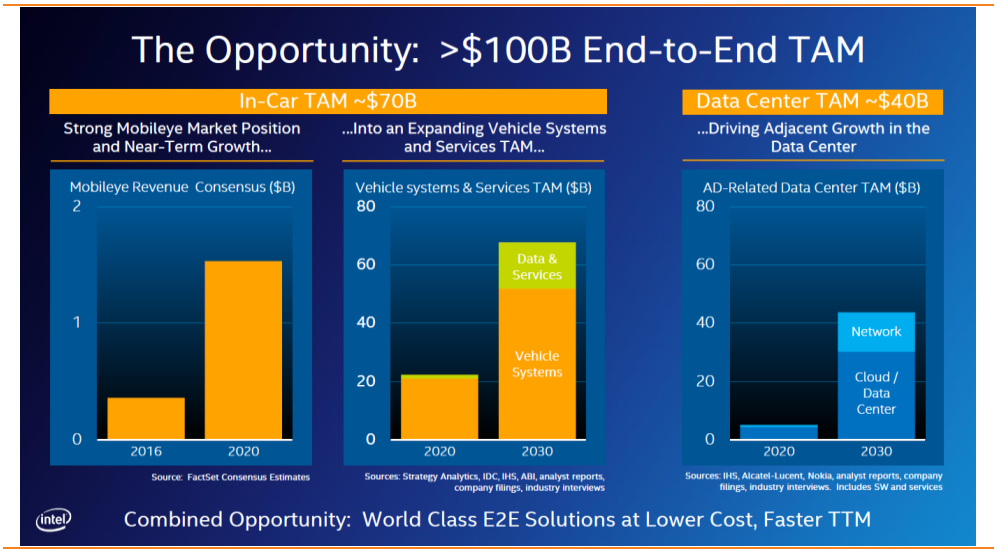
3 月 13 日 Mobileye 和英特尔联合宣布，英特尔将以每股 63.54 美元价格收购 Mobileye，对应股权价值 153 亿美元，企业价值 147 亿美元。收购完成后，英特尔将把自动驾驶部门设在以色列，由 Mobileye 的创始人 Amnon Shashua 继续领导和整合自动驾驶部门，加速 Mobileye 产品开发和上市时间，并极大地增强英特尔在汽车业务上的实力。我们预计 Mobileye 的机器视觉算法与英特尔的芯片、数据中心、AI、传感器融合，以及地图服务等方面产生强大的协同合作效应，联手打造“软硬兼施”的全新无人驾驶供应商。

我们认为，汽车电子化和智能化的方向将持续提高科技类公司在汽车产业链内的重要程度，我们看到三星收购哈曼，高通收购 NXP，到现在英特尔收购 Mobileye，都说明掌握关键技术和客户资源的技术公司的并购价值。

随着英特尔加持，目前自动驾驶上游系统解决方案浮现出英伟达与英特尔-Mobileye 联盟两大竞争者。英特尔-Mobileye 联盟作为 ADAS 行业的龙头，拥有全行业最广泛的车企合作关系。目前正联手宝马和德尔福，共同研发 EyeQ5 第五代车载核心，支持全套开源 SDK 给第三方开发者进行算法开发，配合 OTA 空中升级模块。**我们认为英特尔-Mobileye 联盟的商业路径最为明晰——从 ADAS 出发，逐步完善功能模块，提高自动化程度，进化到 EyeQ5 将会成为一个开源性、定制化、可升级的标准解决方案，打造成为无人驾驶界的 Android 平台。**

Intel 在收购 Mobileye 的时候也给出了市场空间指引，他们认为汽车电子化和智能化的发展方向将是“车轮上的数据中心”。到 2030 年，L1/L2 和 L3/L4 智能驾驶的渗透率将分别达到 40%和 30%以上。而且车载系统的计算能力将从 L2 系统的 0.5 TFLOPS 提升到 L4 系统的 50 TFLOPS。而整个市场空间，包括广告系统、数据和服务也将从 2020 年的 200 亿美元提升到 2030 年的 700 亿美元。

图 120：英特尔给出的市场空间指引



资料来源：公司官网，天风证券研究所

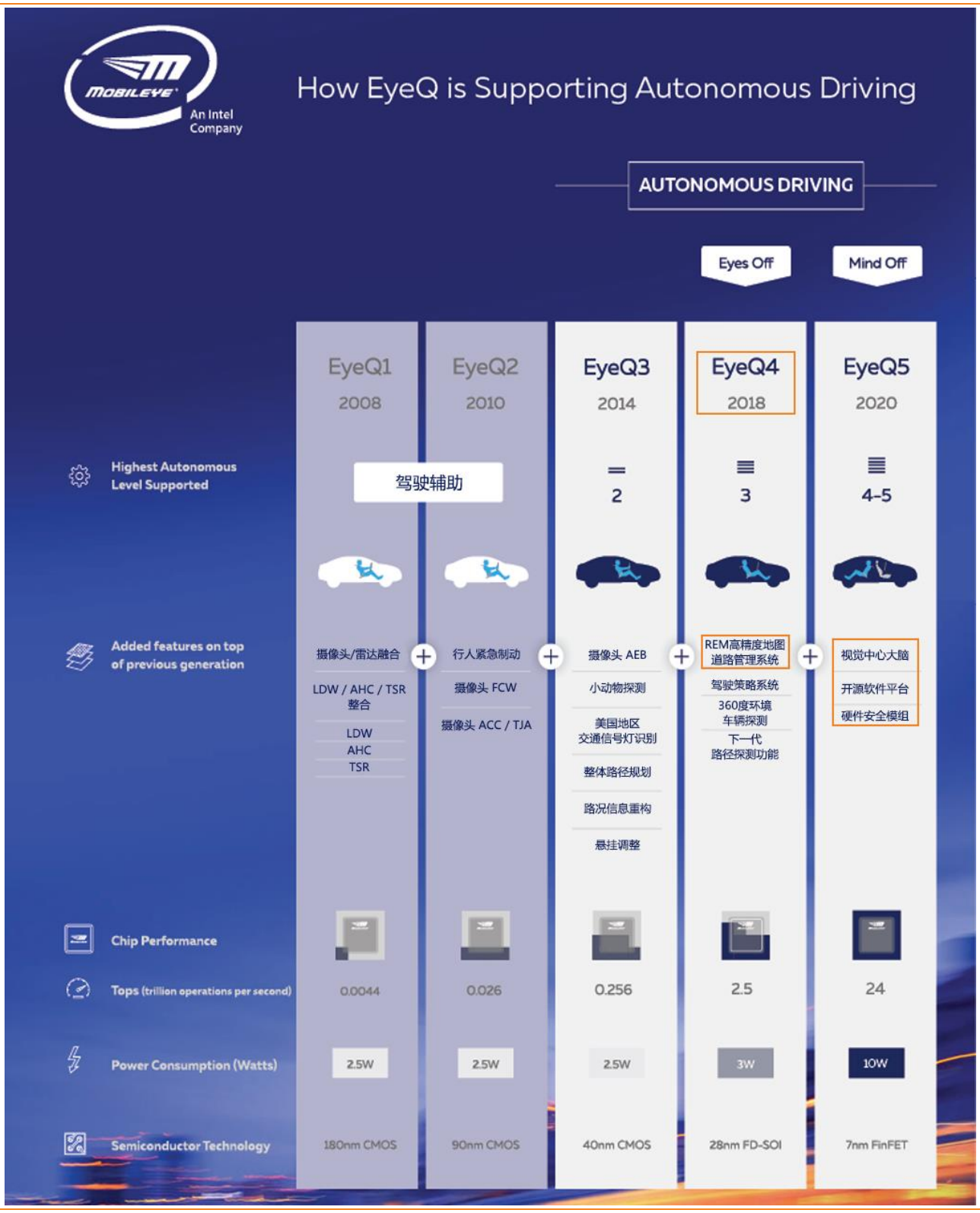
4.2.1. EyeQ 芯片发展之路

Mobileye 从 20 年前开始进行车辆摄像头传感器识别领域的研发，并在 EyeQ 1 和 EyeQ 2 阶段推出了行业首家以摄像头图像为主源数据的 LDW、AHC、TSR、ACC、TJA、AEB 的 ADAS 功能整合，依靠强大的图像识别处理运算能力，逐步成长为 ADAS 行业的龙头。截止到 2016 年 12 月，Mobileye 的合作车企有 27 家，合作车型达到了 313 款，覆盖了全球 90%的主要车企。

定制化片上系统 EyeQ 系列是 Mobileye 的核心技术产品，每一代 EyeQ 芯片都由 CPU 核心以及定制化向量加速器组成，每一代新产品都相较前代产品有 6-8 倍的算力提升，功耗上都保证在 3 瓦以内（EyeQ5 提升至 5 瓦）。

EyeQ 芯片中的加速期核心都具备异构可编程性，用来支持包括机器视觉、信号处理、机器学习任何以及深度神经网络的部署。

图 121：EyeQ 系列芯片参数介绍



资料来源：Mobileye 官网，天风证券研究所整理

EyeQ1：2007 年推出，支持两种 ADAS 套件：1. LDW+TSR+IHC，2. LDW+融合雷达识别的车辆 AEB；

EyeQ2：2010 年推出，在 EyeQ1 基础上加入了 FCW 以及车辆行人 AEB（部分刹车）；

EyeQ3：2014Q4 推出，截止到 2016 年获得超过 12 家车企订单。AEB 升级为全自动刹车，加入交通信号灯识别，路况信息重构(road profile reconstruction)，整体路径规划，自动变道，后车追尾识别以及 REM 等功能。Mobileye 表示，为车辆配备多块 EyeQ3 芯片，叠加

多摄像头识别，可以让车辆具备 L2 级别智能驾驶水平，**基本达到但并不具备脱离方向盘控制的能力。**

EyeQ4：2016 上半年已经推出工程样品，今年下半年搭载于宝马 iNext 概念车中进行真实路测，目前已获得包括宝马、通用、尼桑、大众在内的 5 家车企的合作订单。不过预计上市时间已经从此前的 2017 年下半年推迟至 2018 年。

EyeQ4 搭载 14 核心（4 个 CPU 核心+10 个定制化图像处理加速核心），算力可达每秒 2.5 万亿条命令；支持单车最多 10 个摄像头 36fps 的数据输入，包括一个前向三目摄像头以及数个侧向、后向摄像头。EyeQ4 已经具备多传感器数据处理能力，能够处理车载雷达和激光雷达传感数据。

EyeQ4 引入了多线程处理集群(Multithreaded Processing Cluster, MPC)，拥有比 GPU 更好的通用性以及 CPU 更低的功耗。以及可编程宏阵列(Programmable Macro Array, PMA)，在运算密度上达到固定程序加速硬件的同时还具备可编程能力。

EyeQ5：公司目前正在联手宝马、英特尔共同开发第五代 EyeQ 芯片，计划打造为完全自动驾驶车辆的**“视觉中心大脑”**，预计 2020 年推出一站式自动驾驶解决方案。EyeQ5 能支持超过 16 个摄像头的图像数据以及多传感器融合技术，由英特尔提供的 CPU 核心算力预计达到每秒 15 万亿条命令。

EyeQ5 的加速器核心中搭载了第二代 MPC 和 PMA，同时集成了硬件安全模块(Hardware Security Module)用于支持软件 OTA 空中升级以及安全车内通讯等功能。

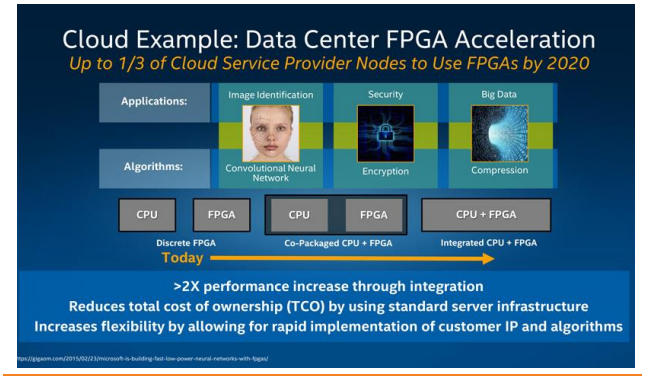
从 EyeQ5 开始，Mobileye 将会正式支持全自动驾驶标准的操作系统以及全套开源 SDK 用于开发者进行算法开发。**我们认为 EyeQ5 会最终形成开放式合作平台，谋求更多的优质合作伙伴加入自动驾驶技术研发领域，成为无人驾驶界的 Android 平台。**

4.3. 收购 Altera 全面加速数据中心

2015 年 6 月，英特尔宣布以每股 54 美元，总价 167 亿美元价格收购 FPGA 芯片厂 Altera，成为当时英特尔历史上规模最大的收购案，被收购后的 Altera 将以“可编程方案事业部”(Programmable Solutions Group)的名义在英特尔运作。

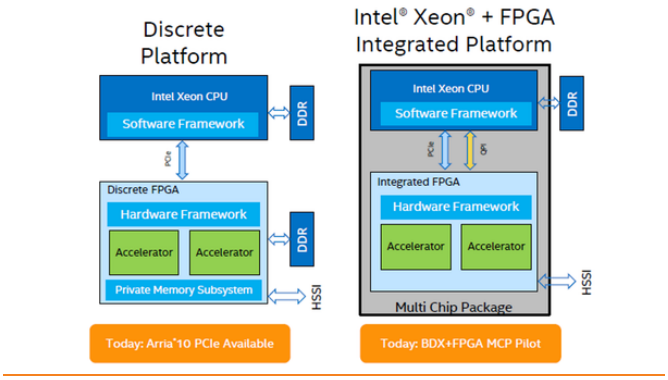
可以通过整合 Xeon 处理器和全定制化的 FPGA 加速器来极大提升应用性能，利用 FPGA 的可重编程能力，在工作负载和计算需求发生波动的时候帮助改变算法。

图 122：英特尔计划在数据中心里提供 FPGA 加速



资料来源：公司官网，天风证券研究所

图 123：英特尔整合 Xeon 处理器和全定制化的 FPGA 加速器



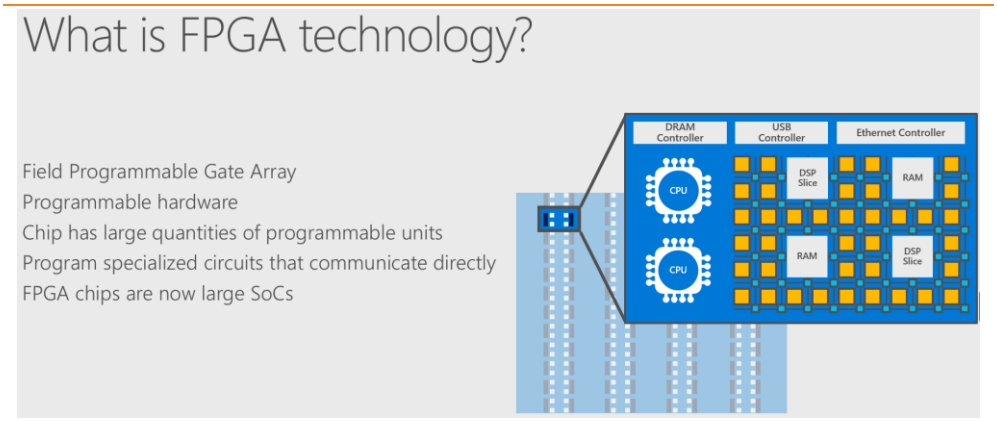
资料来源：公司官网，天风证券研究所

5. 可编程的 FPGA，推理端伸展拳脚

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 采用了逻辑单元阵列 LCA (Logic Cell Array)概念，内部包括可配置逻辑模块 CLB (Configurable Logic Block)、输入输出模块 IOB (Input Output Block)和内部连线(Interconnect)三个部分。它作为专用集成电路(Application Specific Integrated Circuit, ASIC)领域中的一种半定制电路而出现的，既解决了定制电路的不足，又克服了原有可编程器件门电路数有限的缺点。**FPGA 依靠电路级别的通用性，在推理端尤其相较于 CPU 体现良好的低延迟、低功耗等优势，适用于传感器数据预处理工作以及小型开发试错升级迭代阶段。**

Grand View Research 分析，2015 年全球 FPGA 总市场规模达 63.6 亿美元，并预计到 2024 年 FPGA 市场规模将达到 142 亿美元。其中，Xilinx 的市场份额为 49%，主要产品包括，Virtex 系列、EasyPath 系列、Spartan® FPGA 系列等。Xilinx 是 FPGA 的发明者，以技术创新为目标，引领市场的发展趋势。他们的 FPGA 芯片比较大，逻辑门比较多，主要应用到工业和通讯领域，但近年亦致力于在云计算数据中心的服务器以及无人驾驶的应用。

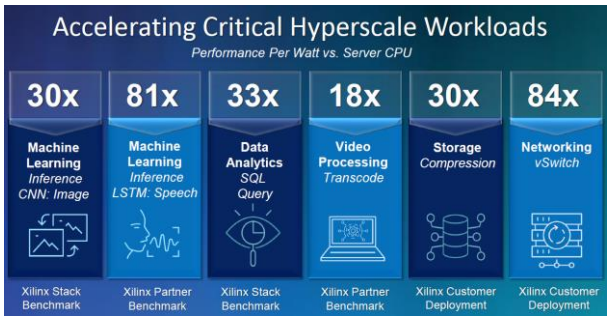
图 124：FPGA 具有低延迟、低功耗、硬件可编程等优良特性



资料来源：微软 Build，天风证券研究所

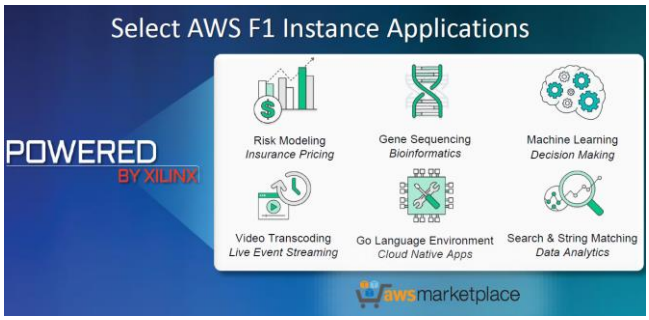
Xilinx 早前公布跟 IBM 合作，通过 CPU+FPGA 的组合，在云计算服务器里提供人工智能辅助应用。鉴于 GPU 的速度和性能，他们主要负责提供人工智能的核心和复杂算法，但是算法编程一旦固定了之后修改比较麻烦，加上机器学习算法里有很多参数(parameters)是需要一边训练一边调整，所以在 GPU 以外使用 CPU+FPGA，可以依靠 FPGA 可编程的性能去做参数调整。

图 125：赛灵思提供的 FPGA 与 CPU 性能对比优势



资料来源：赛灵思官网，天风证券研究所

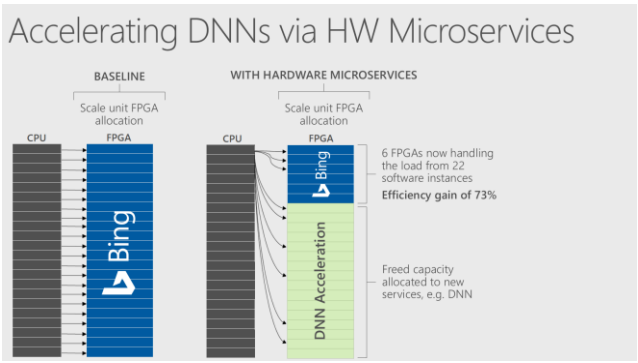
图 126：赛灵思 FPGA 被应用在亚马逊 AWS 中



资料来源：赛灵思官网，天风证券研究所

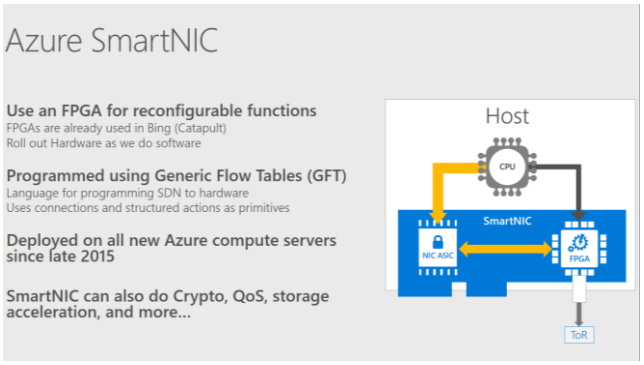
另一巨头 Altera（已被英特尔收购）的市场份额约为 40%。他们的定位跟 Xilinx 类似，同样针对大型芯片和工业、通讯和云计算应用，也比较重视片上系统(SoC)。被收购之后希望与英特尔共同打造完整的嵌入式生态系统。主要产品包括，Cyclone 系列、Stratix 系列等。目前 Altera 的 FPGA 产品被用于微软 Azure 云服务中包括必应搜索、机器翻译等应用中。

图 127：微软使用 FPGA 进行 Bing 搜索加速



资料来源：微软 Build，天风证券研究所

图 128：微软 Azure 从 2015 年开始就布局 FPGA 的使用



资料来源：微软 Build，天风证券研究所

莱迪斯半导体(Lattice Semiconductor)的市场份额约为 6%。他们定位跟 Xilinx 和 Altera 不一样，主要市场为消费电子产品和移动传输，以降低耗电量、缩小体积及缩减成本为主。主要产品有 iCE40 Ultra / UltraLite、MachXO3 Series、ECP Series 等，应用于手机和无人机等。

Microsemi (Actel)的市场份额约为 4%。主要产品为 Fusion、IGLOO、ProASIC3L 等。公司致力于为通信、国防与安全、航天与工业等市场。

6. 冉冉升起的特制芯片新星

6.1. 寒武纪—人工智能 NPU / MLU 芯片

寒武纪作为背靠中科院计算所和中科曙光的 AI 芯片独家首公司，我们认为在芯片开发实力上处于国内领先地位。目前 1A 芯片通过 IP 授权形式进入华为手机，并与中科曙光进行产业链互补，我们看好公司的理论技术储备和研发实力，在人工智能大产业上游 AI 芯片市场蓝海新盛的阶段崭露头角。公司自下而上的策略，从提供低功耗嵌入式终端的本地智能处理芯片解决方案入手，计划逐步向服务器云端的训练处理芯片去布局，我们认为如果能有效利用计算所+中科曙光“背书”的产业资源支持，有效开发硬件指令集和软件平台，有机会构建出健康成长的用户生态圈。

2016 年 4 月 27 日，中科曙光“数据中国加速计划”宣布与北京中科寒武纪科技有限公司，这也是中科院计算所继 2015 年与英伟达签署深度学习联合实验室备忘录后，又一次与深度学习企业进行合作，表现出深度布局人工智能的明显意图。

目前寒武纪主要有三条产品线：1）是 IP 授权，智能 IP 指令集可授权集成到手机、安防、可穿戴设备等终端芯片中，2016 年全年拿到 1 亿元订单；2）在智能云服务器芯片领域：作为 PCIe 加速卡插在云服务器上，希望能布局进入人工智能训练和推理市场；3）开发面向家用智能服务机器人、智能驾驶、智能安防等领域的应用芯片。

寒武纪作为全球第一个成功流片的 AI 芯片公司，从 2017 年起获得了中科院为期 18 个月共计 1000 万元的专项资金支持，用于项目研发及产业化，探索下一代人工智能芯片的架构、算法以及在一些新型场景（如 AR/VR）中的应用开发方法。8 月，寒武纪宣布完成 1 亿美元 A 轮融资，领投方国投创业（国投集团子公司），阿里巴巴、联想、国科投资、中科图灵加入，原 pre-A 轮投资方，元禾原点创投、涌铧投资继续跟投。寒武纪目前估值已接近 10 亿美元，成为了全球第一家智能芯片领域独角兽公司。

图 129：寒武纪产品研发历程

资料来源：公司官网，天风证券研究所

另外公司在最近的发布会上公布机器学习处理器 MLU 的研发规划，并与中科曙光合作开发 Phaneron 智能推理服务器，将搭载 MLU 面向深度学习的在线推理业务环境。目前 Phaneron 服务器已进入最后测试阶段。我们认为寒武纪逐步向服务器云端的训练处理芯片去布局，针对中国市场有机会打开中小型客户需求，我们将持续关注公司新品研发动态。

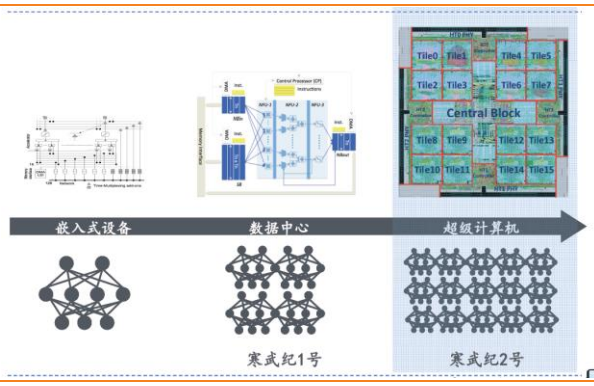
图 130：机器学习处理器 MLU 系列，布局服务器级加速

2018	
MLU100	**MLU200**
• TSMC 16nm工艺 • 支持推理和训练，偏重推理 • 数据中心、小型服务器 • PCIE板卡	• 支持推理和训练，偏重训练 • 企业级人工智能研发中心 • PEIC板卡

资料来源：公司官网，天风证券研究所

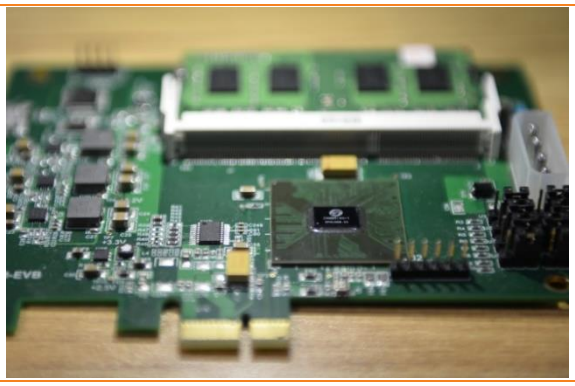
主要技术：从 2014 年至今，寒武纪一共推出了三款寒武纪深度学习芯片，分别是：面向神经网络的原型处理器结构的寒武纪 1 号、面向大规模神经网络的寒武纪 2 号、面向多种机器学习算法的寒武纪 3 号。

图 131：寒武纪处理器系列



资料来源：中科院网站，天风证券研究所

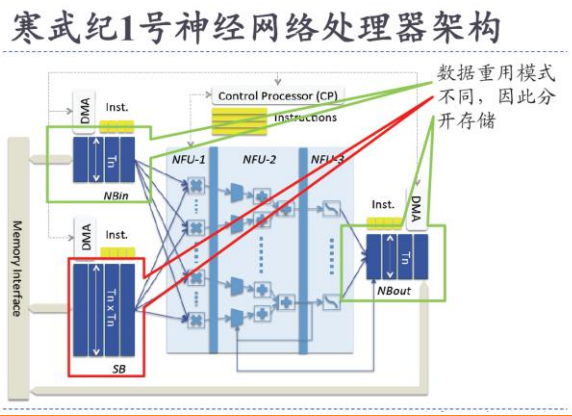
图 132：寒武纪芯片的板卡



资料来源：中科院网站，天风证券研究所

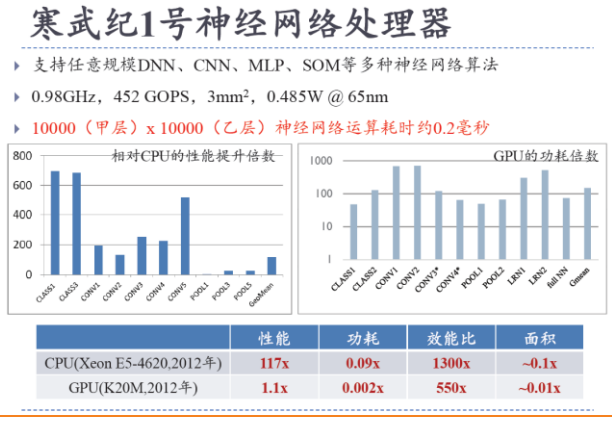
寒武纪 1 号是寒武纪系列的原型处理器结构，支持任意规模 DNN、CNN、MLP、SOM 等多种神经网络算法，包含 16 个神经元硬件运算单元。

图 133：寒武纪 1 号架构



资料来源：中科院网站，天风证券研究所

图 134：寒武纪 1 号处理器

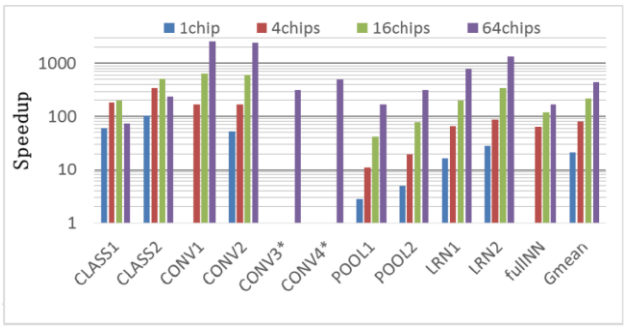


	性能	功耗	效能比	面积
CPU(Xeon E5-4620,2012年)	117x	0.09x	1300x	~0.1x
GPU(K20M,2012年)	1.1x	0.002x	550x	~0.01x

资料来源：中科院网站，天风证券研究所

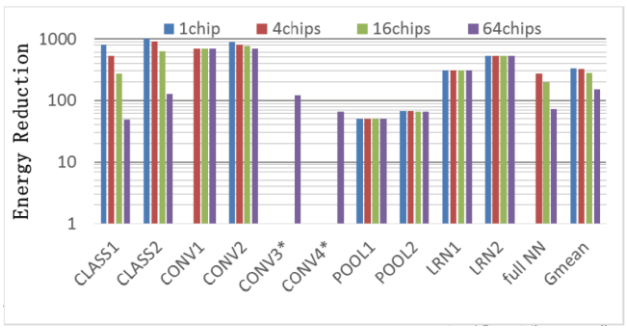
寒武纪 2 号是基于 1 号芯片的结构原理上加以大规模的处理器扩张，2 号一共有 16 个处理器核，运算所需数据全部存储在片上，单芯片片上存储(eDRAM)达 36MB，单芯片运算速度达 5.58 TeraOps/s。单芯片相对英伟达 K20 GPU 性能提升 21 倍，64 结点系统相对 GPU 性能提升 450 倍；能耗降低 330 倍，64 结点系统相对 GPU 能耗降低 150 倍。

图 135：寒武纪 2 号性能比较



资料来源：中科院网站，天风证券研究所

图 136：寒武纪 2 号能耗比较



资料来源：中科院网站，天风证券研究所

寒武纪 3 号则是极大了扩充了人工智能算法，包括 k-最近邻、k-均值、朴素贝叶斯、线性回归、支持向量机、决策树、神经网络等近十种代表性机器学习算法，目的是制造满足不同用途的深度学习处理器，论文中寒武纪 3 号与英伟达的 K20M GPU 作出了性能对比，在 65nm 的工艺下，寒武纪 3 号性能比 K20M 高 1.2 倍，而性能只是其 1/128。

图 137：寒武纪 DianNao 系列主要产品与性能

寒武纪NPU主要产品与性能

产品名称	核数量(个)	核心频率(Core Block)(GHz)	峰值性能(亿次运算/s)	不同nm工艺下功耗(Watt)	面积(mm2)	性能与功耗	
						传统CPU	主流GPU、GPGPU
寒武纪1号(DianNao)	1	0.98	1350	0.485W (65nm)	170.9	100倍、功耗1/10	性能相当、功耗1/100
寒武纪2号(DaDianNao)	16	0.606	-	16W (28nm)	67.7	-	21倍、1/330
寒武纪3号(PuDianNao)	-	1	10560	0.596W (65nm)	3.51	-	性能相当、功耗1/100

资料来源：公司网站，天风证券研究所

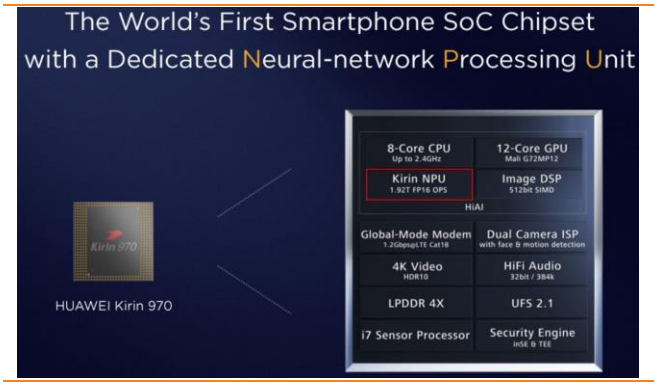
个人成就：陈云霁（中科院计算技术研究所研究员、正教授）、陈天石（博士）兄弟一步一步从学术到产业，将深度学习处理器做到了极致。从 2002 年起，陈云霁一直从事国产处理器的研发工作，先后负责或参与了多款龙芯处理器的设计。2014 年，凭借“寒武纪”神经网络专用处理器，他们在 ASPLOS 的文章两获 CCF 推荐 A 类国际学术会议的最佳论文奖。2016 年 3 月，二人的课题组与寒武纪公司提出的深度学习处理器指令集 Cambricon 首次被 ISCA 2016(International Symposium on Computer Architecture)接受，并排名第一。ISCA 被世界公认为计算机体系结构领域最重要的国际会议，人工智能深度学习的多项重要技术多源自 ISCA。到目前为止，二人已经获得了两篇 ASPLOS、两篇 ISCA、一篇 MICRO，

一篇 HPCA，这些是计算机体系结构方面国际四大顶级会议。

6.1.1. IP 授权进入华为手机

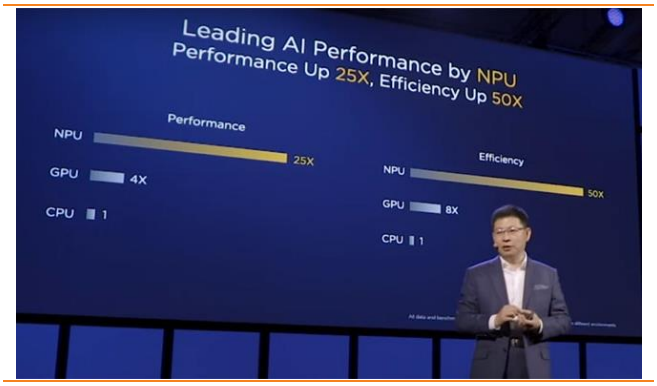
今年 9 月，华为发布全球首款人工智能移动计算平台麒麟 970，并搭载业界首颗带有独立 NPU(Neural Network Processing Unit)专用硬件处理单元。根据市场反馈，这块 NPU 是寒武纪科技去年发布的寒武纪 1A 处理器(Cambricon-1A Processor)，寒武纪通过 IP 授权形式进入华为手机芯片。华为表示，相较于四个 Cortex-A73 核心，新的异构计算架构拥有约 50 倍能效和 25 倍性能优势。

图 138：华为海思麒麟 970 架构搭载寒武纪 IP 的 NPU



资料来源：华为官网，天风证券研究所

图 139：寒武纪 NPU 的性能优势



资料来源：华为官网，天风证券研究所

华为和寒武纪的合作，力图解决端侧 AI 的应用需求，在低功耗、低延迟、便携性等要求下，实现端侧设备的离线智能化。手机中的主要应用场景包括：语音识别、人脸图像识别、传感器数据采集融合等。

7. 量子计算是啥？具体用来干嘛？

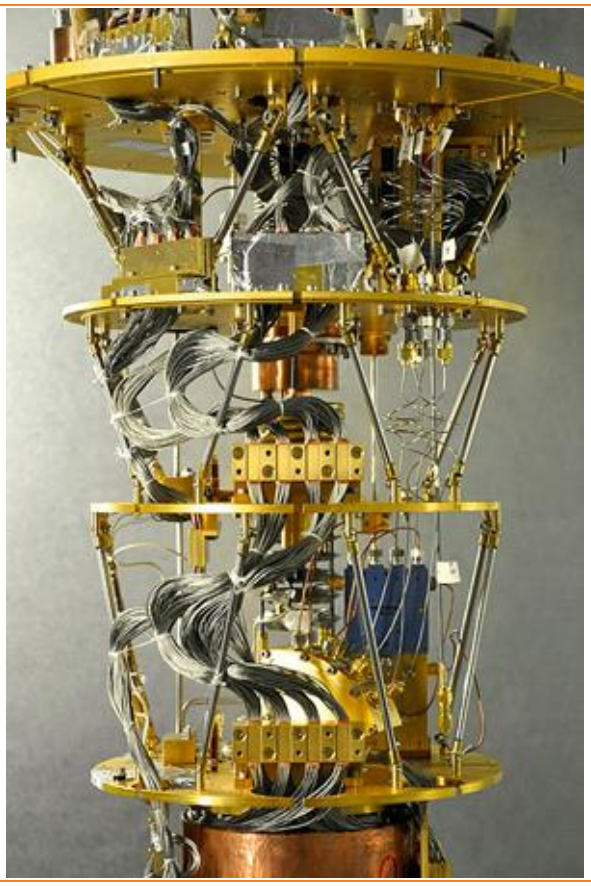
传统电脑是以二进数方式(0&1)存储数据，然后以逻辑来操作。每一个记忆单元叫做比特(bit)，而每个 bit 只能在一个时间内维持一个状态。就是说，要么就是 0，要么就是 1，不能同时是 0 和 1。

量子计算是量子物理学的一门，最初是由 Paul Benioff, Yuri Manin, Richard Feynman 和 David Deutsch 在 80 年代初开始研究。相对于传统电脑，量子电脑的记忆单元叫做 qubit (quantum bit)，而它的形态可以是 0、1 或两个都是(superposition of 2 states)。所以，量子电脑可以操作更高维度的计算。这个特性正好用来解决加密、解密和一些机器学习的优化问题(optimization problem)。

量子计算主要应用量子力学的特性，superposition (量子叠加)、quantum tunneling (量子隧穿) 和 entanglement (量子纠缠)，来进行混算。量子叠加就是信息可以从一个 qubit 移动到另外一个 qubit 而无需在任何地方之间。而量子纠缠就是在一个 qubit 上发生的情况可以影响到其他 qubit，哪怕他们是在不同的地方。编程量子电脑就是用量子纠缠来配置相邻量子之间的关系，而量子隧穿就是去解决最小能量值的量子。最小能量值就是最佳的答案。

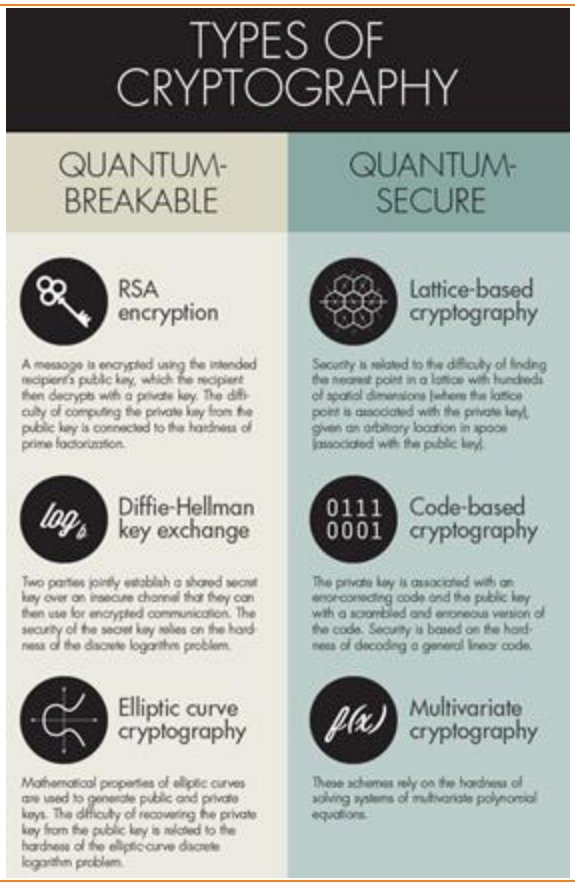
如果一个 qubit 在叠加，而它跟另外一个 qubit 在同样的叠加上，计算时就会产生 4 个结果：0/0，0/1，1/0 和 1/1，或者每 n 个 qubits 就可以有 n^2 个状态。以上例子里的量子特性，让量子电脑可以同时平衡的执行不同计算(quantum parallelism)。鉴于量子计算要在叠加的状态中进行，如果在计算的过程中有任何的观察行为，叠加的状态就会终止，然后量子会返回单独的状态，这就是消相干性(decoherence)。消相干性就是当量子从叠加的状态回复单独（就是 1 或 0）可观察的状态。所以观察行为要等到计算完成后才能执行，而结果就会是单独的（1 或 0）状态。

图 140：D-Wave 2 量子计算机支撑结构，机器被冷却到接近绝对零度



资料来源：MIT Technology Review，天风证券研究所

图 141：可以与不可以被量子计算攻破的加密技术



资料来源：Wired，天风证券研究所

在经典的“薛定格(Schroedinger)的猫”试验中，猫是放在一个密封的盒子里，然后里面有一种放射性同位素，有 50%的机会会发生衰变，然后发射出一个粒子来触发毒气设置，猫就会死。如果中间没有人在观察猫是生还是死，那么猫就会在既是生又是死的状态。但是，如果有人在观察，那么猫的状态就会是生或者是死。所以，这个例子描述了观察会瓦解量子物理的操作。

理论上，大型的量子电脑对于某一些计算问题是可以做得比传统电脑要快，比如说人工智能、机器学习和破解现代常用的加密技术等。美国的国家安全局(National Security Agency, NSA)也估计现代常用的加密技术将会在量子电脑普及之后变得过时。现在，美国的科技巨头包括谷歌、IBM、微软、惠普和加拿大的 D-Wave 都有参与研究和尝试将这个科技商业化。

现代的三款普遍的加密方法有 RSA、Diffie-Hellman 和 Elliptic Curve。RSA 的基楚是两个超大（比如说 300 位数）的质数(prime number)互相乘以的积，要用质数因数分解法(prime factorization)去破解。Diffie-Hellman 的破解方式是解决离散对数(discrete logarithm)，而 Elliptic Curve 是 Diffie-Hellman 的变种，破解方式是解决椭圆曲线离散对数。这些加密技术的具体计算方式不是重点，而重点是他们的破解方法都是利用传统电脑需要长时间甚至没办法去计算的问题。虽然量子电脑理论上可以用 Shor's Algorithm 去解决这种问题，但到目前为止，量子电脑能够因数分解的最大数值为 56153。

那么，有没有一些不能被量子电脑破解的加密技术呢？是有的。这些技术被称为后量子加密法(post-quantum cryptography)。他们包括：1、Lattice-based (以格子为基础，在数百空间维度的格子里查找最近的点子，而密钥跟这点子相关), 2、code-based (以代码为基础，密钥跟一个纠错码相关，而公开钥匙是在密钥上加扰)和 3、multivariate (基于求解多元多项式方程组的难度)的方法。

7.1.1. 量子电脑的历史

简短的说下量子电脑的历史。从 2001 年开始已经有 7-qubit 的量子电脑可以用 Shor's 来因数分解 15。到 2011 年以 10 qubits 和 Shor's 因数分解 21，接着因数分解了 143 和 56153（但不是用 Shor's）。在 2005、2009 和 2010 年，美国密歇根大学、耶鲁大学和英国的布里斯托大学等分别研究量子硬件。

2011 年，加拿大公司 D-Wave Systems 推出了首部商业用的量子退火炉(quantum annealer)，叫做 D-Wave One，拥有 128 qubits。美国国防军工公司 Lockheed Martin 采购了一部，放在南加州大学的量子计算机中心。D-Wave Systems 的第一部量子电脑是在 2007 年建成的，拥有 16 qubits。电脑要在超低温，约 20mK 的接近绝对零度(Absolute Zero,即摄氏-273 度)下操作。公司称主要的电耗来自冷却方面，而整个系统消耗约 15 千瓦(kilowatt)的电，但芯片只需要少于 1 微瓦(microwatt)。每一个 qubit 都是一个超导体电路，在超低温下电流才能同时双向流。D-Wave 的产品其实不是一部全面的量子电脑，就是说它不是所有问题都合适。他们的电脑最合适的是解决优化问题。所以，对于大数据的处理，包括人工智能和机器学习，是非常合适的。其他应用之处还包括加密技术、蒙地卡罗模拟技术、预测分子的化学作用来设计药物等。

2012 年，IBM 也发表一些突破性的研究。同年 10 月，诺贝尔奖颁给了量子学家 David J. Wineland 和 Serge Haroche。12 月，第一家量子软件公司 1Qbit 在加拿大成立。同年，亚马逊的 CEO Jeff Bezos 和美国中央情报局(CIA)也投资了 3000 万美元进去 D-Wave。

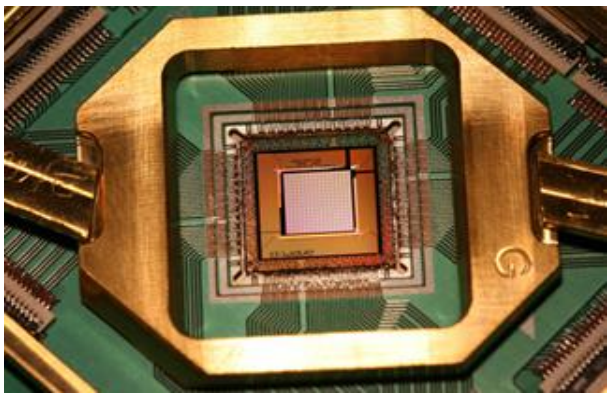
2013 年，谷歌和美国太空总署(NASA)合作的量子人工智能研究所采购了一部 512-qubit 的 D-Wave 量子电脑。谷歌希望量子电脑可以解决人工智能的问题，而 NASA 希望这电脑可以协助安排火箭升空的时间和模拟日后的太空任务和穿梭机飞行。

2014 年，前 NSA 员工斯诺登(Edward Snowden)透露了 NSA 花了约 8000 万美元去研究量子电脑破解加密方式。

到 2015 年，量子电脑的开发仍然处于幼嫩期，电脑里面的 qubits 还是比较少。9 月份，D-Wave 公开发售最新的 D-Wave 2X 量子电脑。D-Wave System 公司宣布与谷歌签订新的合约，谷歌将在未来 7 年内继续使用 D-Wave System 提供的量子计算机，而且 D-Wave System 也会在 NASA 的研究中心内安装其最新一代 D-Wave 2X 量子计算机，相较于 D-Wave 2 的 500 个量子核心数，2X 的将含有超过 1000 个量子处理器核心。

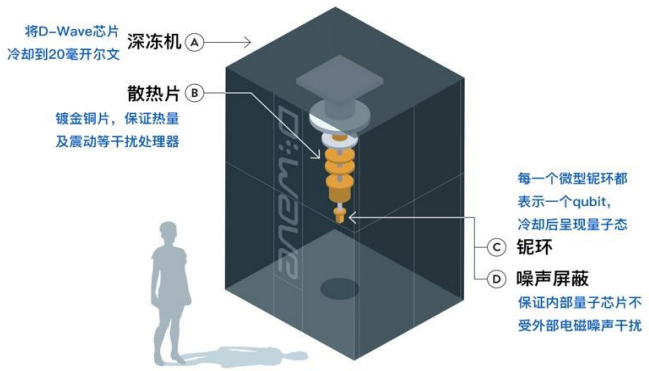
同一时间，英特尔宣布在 10 年内投资 5000 万美元和荷兰的 Delft 大学的 QuTech 合作研发量子计算科技。英特尔认为量子计算的具体应用还有好几年的时间才成型，而虽然现在有一些突破，但是正式普及也最少需要十年时间。这也反映在他们的投资时间和额度里。

图 142：D-Wave 的量子处理器



资料来源：D-Wave Systems 网站，天风证券研究所

图 143：D-Wave 量子计算机示意图



资料来源：天风证券研究所整理

同年 12 月，NASA 公开展示了价值 1500 万美元的 D-Wave 量子电脑，这是上述提到跟谷歌合作的项目。在选择 D-Wave 的量子计算机之前，谷歌、NASA 以及 USRA 运行了 D-Wave 的机器进行基准测试。D-Wave 标榜自己为第一家生产商业化量子计算机的公司，但因为

他们使用的量子运行原理与传统方法不一致，业界一直有对其计算机是否真的符合量子计算机标准的质疑，不过经过检验，在某些特定问题的运算速度上 D-Wave 的机器运算能力是普通计算机的 3600 倍。谷歌也在 2015 年 12 月发表的一片论文中表示，D-Wave 的量子计算机在某项特定而且精心设计的问题上，量子电脑的速度要比传统电脑快一个亿倍。这部量子电脑现在已经由原来的 512 qubits 升级到超过 1000 qubits。而且，qubit 的增加对于耗电量也不会有太多影响。

7.1.2. 谷歌的量子计算机之路

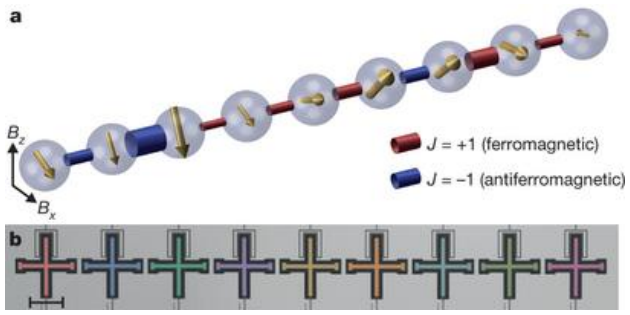
除了上面提到的这个利用 D-Wave 公司设计的量子计算机作为研究核心的人工智能实验室。谷歌还与世界上量子计算领域最前沿的学者，加州大学圣巴巴拉分校的 John Martinis，合作设计量子计算机。2014 年 6 月，谷歌聘请了 John Martinis 及他的团队，并预计能在 2017 年设计出含有 100 个量子核心的量子退火炉。John Martinis 希望借助 D-Wave 计算机在规模化方面的优势以及自己团队在稳定性方面的研究，有力地推动整个量子计算领域的研究发展。

2016 年 6 月，谷歌与西班牙巴斯克大学的研究人员共同宣布了在量子计算机领域的重大研究突破，并表示有希望通过更为便捷的方法建造出一台更能够发挥量子计算能力的计算机样机。谷歌这次从之前耗费了大量研究人力和资源的“数字量子计算机”(digital quantum computers)，转移到“模拟量子计算机”(analog quantum computers)的研发上。研究机构一直以来都把重心放在数字量子计算机上，就是针对特定问题，构建特殊排列的量子位设计的数字电路。这种方法类似传统微处理器中的订制数字电路，但缺陷在于需要大量纠错资源(error-correcting)来弥补脆弱的量子效应，而且无法在量子效应上提高数量级。

目前包括谷歌、IBM 等在内，都转移到模拟量子计算机的加速研发上。模拟量子计算机与传统电子计算机的相似程度更小，而且背后的运算原理目前还不能完全解释清楚，不过系统纠错所需要的资源相较数字量子计算机少很多，从而能更好的发挥量子计算的能力。

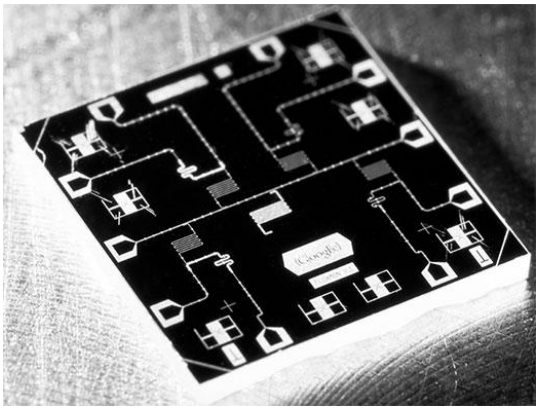
现在谷歌量子计算机项目的负责人 John Martinis 和他的团队搭建的计算机样机，打造了一款超导量子芯片来模拟 9 个相互磁力作用的原子。谷歌的原型机结合了两种量子计算的方法，第一种就是前文提到的数字量子计算机方法，第二种模拟量子计算方法，被称作“绝热量子计算”(adiabatic quantum computing, AQC)。计算机将给定问题编码为一组 qubits，逐步调整这组 qubits 的相互作用以达到最终共同量子态，从而解决给定问题。这个方法具有普适性，所有问题都可以使用同一组 qubits。

图 144：谷歌量子计算机 9 个量子位排列示意图



资料来源：谷歌论文资料，天风证券研究所

图 145：谷歌制造的量子计算实验芯片



资料来源：MIT Technology Review，天风证券研究所

不过 AQC 这个方法也有局限性，就是会受到随机噪声效应的影响，这个影响会带来无法消除的系统误差。谷歌了解 AQC 也不能保证每一个问题都得到有效解决。在没有纠错技术的帮助下，随着系统量级的提高，利用 AQC 技术去扩大计算规模将非常困难，因为在更大的系统中，误差的积累将会很快。所以，他们需要找到更好的方式来应用纠错技术。他们利用了此前在数字量子计算机研究中积累的纠错技术。Martinis 团队表示，目前这款量子芯片只有 9 个 qubits，而要让量子计算机的运算能力达到传统计算机无法匹及的程度，

需要至少拥有 40 个 qubits，D-Wave 安装在 NASA 实验室里的最新代量子计算机 2X 拥有超过 1000 个 qubits。

这 9 个 qubits 采用固态量子位，量子位由十字形的铝制薄膜制成，宽度约为 400 微米，被固定在蓝宝石表面上。研究团队将铝制薄膜降温到 0.02 开氏（约-273 摄氏度），将金属转变为零电阻的超导体。在超导状态下，研究团队可以将信息编码到 qubits 中。

相邻 qubit 的相互作用通过“逻辑门”控制，驱使 qubits 达到能够得出问题解决方案的量子态。在样机中，研究人员调整 qubits 的排列序位来模拟具有自旋态的磁性原子阵列（这个问题在凝聚态物理学中已经得到深入研究），然后研究人员就可以观察 qubits 来确定自旋态原子的最低能量集体态(lowest-energy collective state)。同时，谷歌的量子样机还能够解决传统计算机不能解决的“non-stoquastic”问题，包括在化学研究中所需的对多个电子的相互作用的准确电脑模拟。量子计算机最具实用价值的功能之一就是能够在量子层级模拟分子材料。

南加州大学量子计算专家 Daniel Lidar 表示，新的量子计算方法能够进行量子误差修正，并且可以在谷歌的 9 个 qubits 的样机中完成。谷歌团队表示，凭借量子误差修正能力，他们的量子计算方法能够扩展为通用算法，进而扩展至更为大型的量子计算机使用中。

在未来的几年内，谷歌希望能制造出包含 40 个 qubits 的量子计算机，这个时候“量子优势”(quantum supremacy)才会真正建立起来，进而用来分析并解开医学和能源领域需要进行大规模原子级别仿真计算的难题。谷歌 CEO 皮查伊表示公司已经进入了“人工智能先行”的时代，谷歌量子计算应用设计团队的负责人 Hartmut Neven 也表示，在 10 年之内，人们将会放弃传统机器学习方式，转而拥抱量子机器学习。

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告所表述的所有观点均准确地反映了我们对标的证券和发行人的个人看法。我们所得报酬的任何部分不曾与，不与，也将不会与本报告中的具体投资建议或观点有直接或间接联系。

一般声明

特别声明

在法律许可的情况下，天风证券可能会持有本报告中提及公司所发行的证券并进行交易，也可能为这些公司提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。因此，投资者应当考虑到天风证券及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突，投资者请勿将本报告视为投资或其他决定的唯一参考依据。

投资评级声明

类别	说明	评级	体系
股票投资评级	自报告日后的 6 个月内，相对同期标普 500 指数的涨跌幅	买入	预期股价相对收益 20%以上
		增持	预期股价相对收益 10%-20%
		持有	预期股价相对收益-10%-10%
		卖出	预期股价相对收益-10%以下
行业投资评级	自报告日后的 6 个月内，相对同期标普 500 指数的涨跌幅	强于大市	预期行业指数涨幅 5%以上
		中性	预期行业指数涨幅-5%-5%
		弱于大市	预期行业指数涨幅-5%以下

天风证券研究

北京	武汉	上海	深圳
北京市西城区佟麟阁路 36 号	湖北武汉市武昌区中南路 99 号保利广场 A 座 37 楼	上海市浦东新区兰花路 333 号 333 世纪大厦 20 楼	深圳市福田区益田路 4068 号卓越时代广场 36 楼
邮编：100031	邮编：430071	邮编：201204	邮编：518017
邮箱：research@tfzq.com	电话：(8627)-87618889	电话：(8621)-68815388	电话：(86755)-82566970
	传真：(8627)-87618863	传真：(8621)-68812910	传真：(86755)-23913441
	邮箱：research@tfzq.com	邮箱：research@tfzq.com	邮箱：research@tfzq.com