计算机行业

评级：增持　维持评级　　行业研究

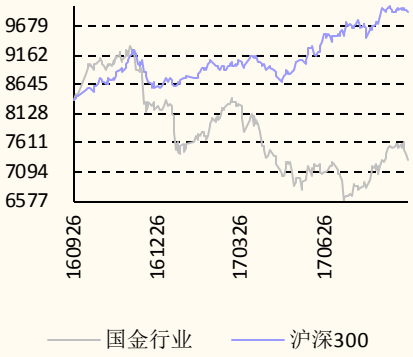
长期竞争力评级：高于行业均值

智能时代，芯片先行

--人工智能系列报告之四

市场数据(人民币)

行业优化平均市盈率	24.68
市场优化平均市盈率	18.43
国金计算机指数	7300.57
沪深 300 指数	3817.79
上证指数	3341.55
深证成指	10930.67
中小板综指	11639.16



相关报告

1.《人工智能时代：AI 赋能，世界重塑》，2017.9.18

2.《无人零售：技术破局引领商业变革》，2017.9.15

3.《企业级服务景气度提升，AI 延伸至消费电子端》，2017.9.11

4.《WannaCry 勒索病毒爆发，利好网络安全行业》，2017.5.15

5.《行业继续保持高景气，板块分化抓住结构性机会》，2017.5.4

行业观点

- **神经网络算法助推人工智能普及，计算能力需求爆发式增长：**神经网络算法是当前最主流的人工智能算法，其通过海量样本数据进行机器学习，从而生成具备智能判断能力的模型。近年来，随着 GPU 等硬件计算平台性能的提升，以及互联网带来的大数据资源，神经网络算法已经被应用到人脸识别、语音识别等多个领域，实现了极高的准确率。但同时，神经网络算法的精确度提升十分依赖于海量的计算资源，计算能力的需求将在人工智能时代迎来爆发式的增长。

- **软件加速遭遇瓶颈，AI 专属芯片势在必行：**神经网络算法应用的不断发展，使得传统的 CPU 已经无法负担几何级增长的计算量。传统 CPU 支持的指令集更加通用，主要针对非计算密集型的程序，其优化在于加速分支判断、逻辑判断等操作，对神经网络算法这种计算密集型的应用并不适合。在芯片产业的发展历史上，当软件加速方案已经远远达不到需求时，针对某一应用的硬件解决方案就会填补这一空白，从信号处理芯片 DSP、图形芯片 GPU 到图像处理芯片 ISP 等，都是这一技术演进路径的案例。基于当下人工智能算法的广泛应用，AI 专属芯片已成行业发展的必然选择。

- **服务器端 AI 芯片：多种技术路线实现高并发计算：**神经网络训练（Training）阶段的加速主要在服务器端完成。在这一过程中，计算节点的处理芯片已不再是传统的 CPU，更适合神经网络计算特点的芯片方案被采用，包括 GPU、FPGA、以及专属的 ASIC 芯片方案。其中 GPU 的浮点计算能力较为出色，FPGA 架构更为灵活，适合迭代开发计算，而专属的 ASIC 芯片方案性能最优，但是初期研发成本较高，目前主要是谷歌、英特尔等巨头玩家参与。三种方案在成本、功耗、速度方面各有优劣，在当前的一段时间内会并存。

- **终端 AI 芯片：应用场景驱动，市场前景广阔：**随着人工智能场景的应用深入渗透到行业的各个领域，在终端，推理（Inference）阶段的计算能力越来越成为瓶颈。一些对即时性要求很高的应用场景，已经无法通过在云端进行推理计算的方式满足，终端 AI 芯片加速成为了必选的方案。于是，以低延时、低功耗为目标的定制化终端 AI 芯片成为了各种应用场景的选择，该领域典型的参与者有专注无人驾驶场景的 Mobileye、机器视觉领域的 Movidius、消费电子端的寒武纪等。我们认为，终端 AI 芯片更接近消费者，在硬件先行的前提下，未来如能形成丰富的终端 AI 应用生态圈，则快速增长的出货量将摊薄前期研发成本，形成行业发展的正反馈。

投资建议

- 我们认为，在人工智能的变革正在深入渗透到各行各业的时代，AI 专属芯片作为计算能力的保障，将迎来巨大的需求。从云端的高性能服务器到终端的视频监控、消费电子等领域，在产业链上均有机会享受这场变革带来的红利。在上市公司受益标的方面，我们建议关注：拥有深度学习平台及产品 XSystem 的中科曙光(603019.SH)、在移动端 AI 芯片产业链布局较早的中科创达(300496.SZ)、与百度联合推出人工智能 ABC 一体机的浪潮信息（000977.SZ）、专注于视频监控前端智能化芯片的富瀚微(300613.SZ)。

风险提示

- 人工智能技术推进不达预期；芯片性能提升无法突破计算瓶颈；消费者对终端 AI 应用接受程度未达预期；

钱路丰　分析师 SAC 执业编号：S1130517060003
qianlufeng@gjzq.com.cn

潘宁河　联系人
panninghe@gjzq.com.cn

蒲梦洁　联系人
pumengjie@gjzq.com.cn

内容目录

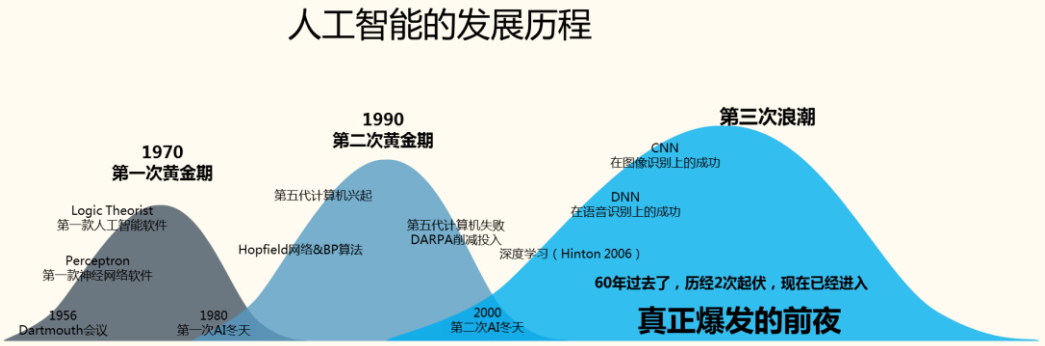
图表目录

一、人工智能深入渗透，催生 AI 芯片产业

神经网络算法逐渐成熟，推动人工智能跨越式发展

- **人工智能的发展经历了较长时间的摸索期。**人工智能（AI：Artificial Intelligence）是一种能够模拟人类智能行为和思维过程的系统，其概念于 1956 年达特茅斯会议上，首次被学术界提出。在人工智能前几十年的发展历程中，聚类算法、决策树算法、支持向量机等算法相继被提出，然而由于实际应用效果欠佳，并未被大规模采用，人工智能的发展甚至一度陷入了停滞。直到近年，深度神经网络算法的大规模应用，使得人工智能技术跨过了漫长的摸索期，打开了下游应用多点开花的新局面。

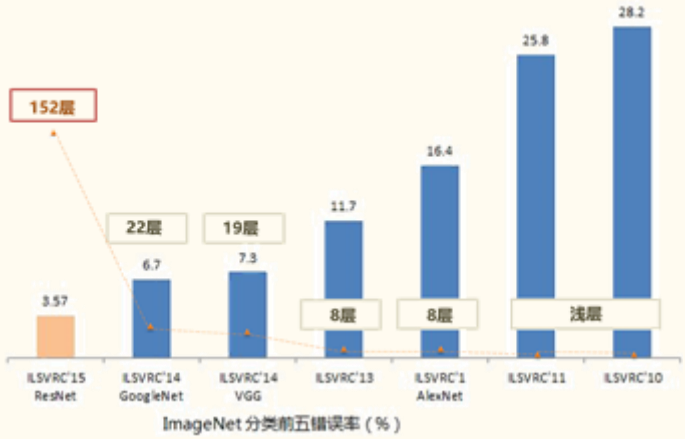
图表 1：人工智能的发展经历了较长时间的摸索期



来源：科大讯飞官网，国金证券研究所

- **硬件性能提升和大数据的发展，使神经网络算法成为当前 AI 主流算法。**神经网络算法，是一种通过模拟人脑神经元的结构，结合机器学习的理念，构建人工智能解决方案的算法。从 1957 年第一个神经网络——感知器被提出以来，神经网络算法便不断被优化，优质算法的不断迭代使得模型的准确率不断提高，另一方面神经网络自身隐含层数的增加也明显提升了模型性能及模型对现实的刻画能力。在算法不断实现突破的同时，数据和计算能力成为主要瓶颈，神经网络隐含层数的增加要求拥有充足的用于训练模型的数据集，但是 80 年代可以采集到的数据极为有限，而且用大量的数据进行训练也要求相应的硬件计算资源及计算能力的提升，当时的硬件水平也无法满足大规模的计算需求，以上因素使得深度神经网络曾经一度“无用武之地”。近年来，随着 GPU 等硬件计算平台被应用到神经网络算法，以及互联网带来的海量大数据资源，神经网络算法已经被应用到人脸识别、语音识别等多个领域，实现了极高的准确率，其已成为当前人工智能领域的主流算法。

图表 2：深度学习算法隐含层数越深，则错误率随之显著降低



来源：ImageNet 测试结果，国金证券研究所

神经网络算法对计算能力提出了极高的要求

- **神经网络算法结果的精确度提升往往依赖于海量的计算。**例如，在 2012 年斯坦福大学黎越国、吴恩达和谷歌科学家联名发表的的论文中，使用了九层神经网络，网络的参数数量高达十亿，是之前论文模型的十倍。而用于训练这个神经网络的图像，是从视频网站 YouTube 上截屏获得的一千万个原始录像，每张图片有四万个像素。该计算模型分布式地在一千台机器(每台机器有 16 个 CPU 内核) 上运行，花费三天三夜才完成训练过程。训练一个复杂的模型的时间有可能更多，例如 Alpha Go 的策略神经网络和价值网络的训练合计需要一个月的时间。并且，在训练模型的过程中，往往需要尝试很多组参数，每调整一次就需要重新训练，所花费的时间将更长，例如在训练百度的机器翻译系统的过程中一共尝试了 10 组参数，整个训练时间达到 100 天。由此可见，如果计算资源不足以支撑大规模计算，那么对于人工智能产品的更新换代和技术创新将十分不利。

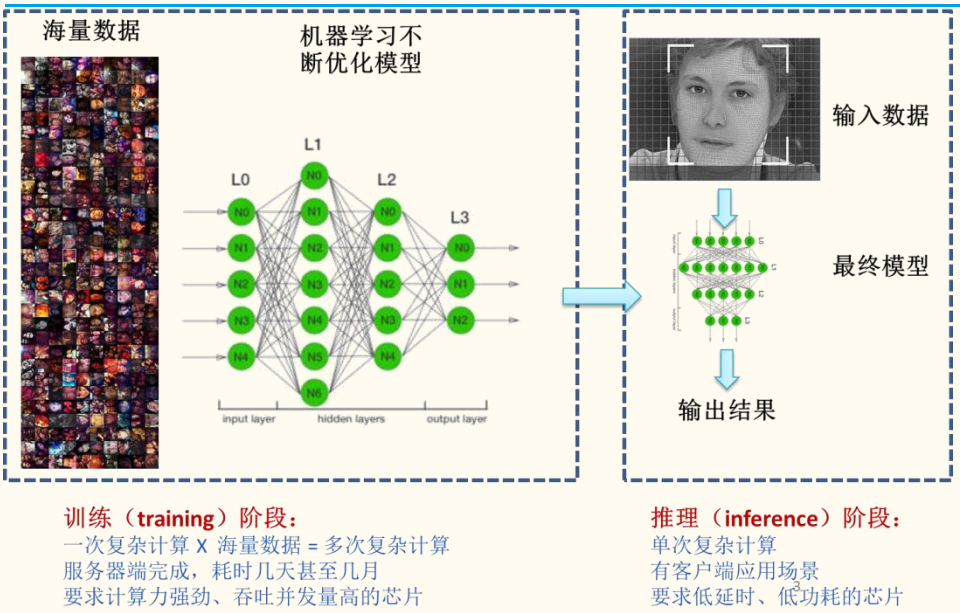
图表 3：神经元数量的增长对计算性能提出了更高的要求

来源：百度官网，国金证券研究所

- **训练（Training）和推理（Inference）阶段均有迫切的加速需求。**一个完整的人工智能解决方案需要经过训练（Training）和推理（Inference）两个过程：训练阶段是通过大量的数据输入，或采取增强学习等非监督学习方法，训练出一个神经网络模型。以人脸识别为例，训练阶段即通过输入几十万甚至几百万张标记信息的人脸图片，从而不断优化模型参数。一旦

训练完成，即可进入推理阶段，同样以人脸识别为例，推理阶段通过输入单张人脸图片和之前训练阶段计算好的模型进行推理计算，得出识别结果。由此可见，神经网络训练（Training）和推理（Inference）阶段都需要很大计算量，但两者场景不同，需求也不一致。训练（Training）阶段由于面对的是海量数据输入循环计算，因而大多是在后台离线计算，其计算系统的设计目标是高并发高吞吐量，加速过程主要是能将计算时间从几个月变成几天左右。而推理（Inference）阶段是直接面对应用场景的即时计算，其往往只需要较少的数据输入，计算系统的设计目标是低延时低功耗，技术突破主要在于将响应时间从几分钟变成几秒甚至零点几秒，增强即时用户体验。

图表 4：神经网络算法训练（Training）和推理（Inference）阶段各有需求

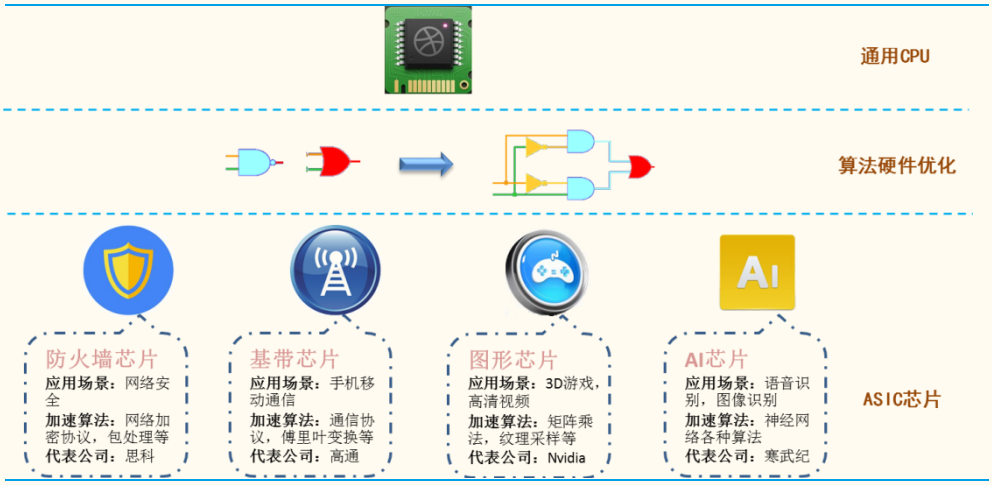


来源：国金证券研究所整理

计算需求几何级增长，催生 AI 硬件解决方案

- **软件加速遭遇瓶颈，硬件优化势在必行。**神经网络算法应用的不断发展，使得传统的 CPU 已经无法负担几何级增长的计算量。传统的 CPU 支持的指令集更加通用，主要针对非计算密集型的程序，其优化在于加速分支判断、逻辑判断等操作，其对神经网络算法这种计算密集型的应用并不适合。在芯片产业的发展历史上，当软件加速方案已经远远达不到需求时，针对某一应用的硬件解决方案就会填补这一空白，从信号处理芯片 DSP、图形芯片 GPU 到图像处理芯片 ISP 等，都是这一技术演进路径的案例。

图表 5：将专用算法硬件优化是 ASIC 芯片的发展方向



来源：国金证券研究所整理

- **巨头纷纷布局人工智能芯片。**随着人工智能的应用领域不断扩张，专属的人工智能芯片已成为了战略高地，从互联网巨头到传统芯片设计公司纷纷布局这一领域。从下图可以发现，GPU 在服务器端的训练应用中使用最为广泛，而设备端的推理应用，不用的场景会采用不用的芯片方案。而谷歌的 TPU 产品目前主要是针对内部应用进行加速，目前尚未对外界发售。

图表 6：互联网和芯片行业巨头纷纷布局 AI 芯片领域

来源：智东西网站，国金证券研究所整理

二、服务器端 AI 芯片：多种技术路线实现高并发计算

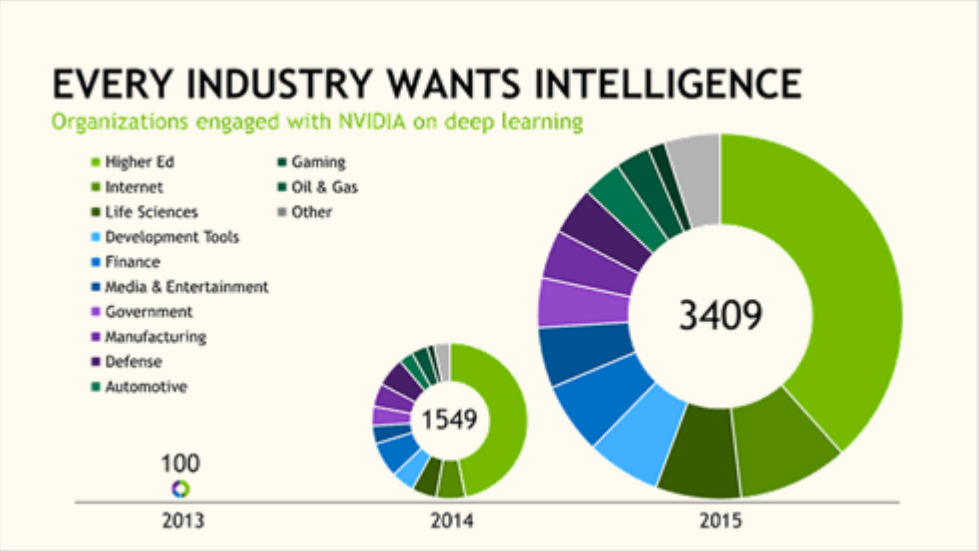
当前，神经网络训练（Training）阶段的加速主要在服务器端完成。通过将算法分布到多个计算节点并行计算的方式，一次复杂的模型训练时间已经从几个月降低到几天至一周左右。在这一优化过程中，计算节点的处理芯片已不再是传统的 CPU，更适合神经网络计算特点的芯片方案被采用，包括 GPU、FPGA、以及专属的 ASIC 芯片方案。

GPU：出色的浮点计算性能加速神经网络计算

- 2011 年 NVIDIA 研究院的 BryanCatanzaro 与斯坦福大学的吴恩达教授展开合作，将 GPU 应用于深度学习，此次合作的事实数据表明 12 颗英伟达

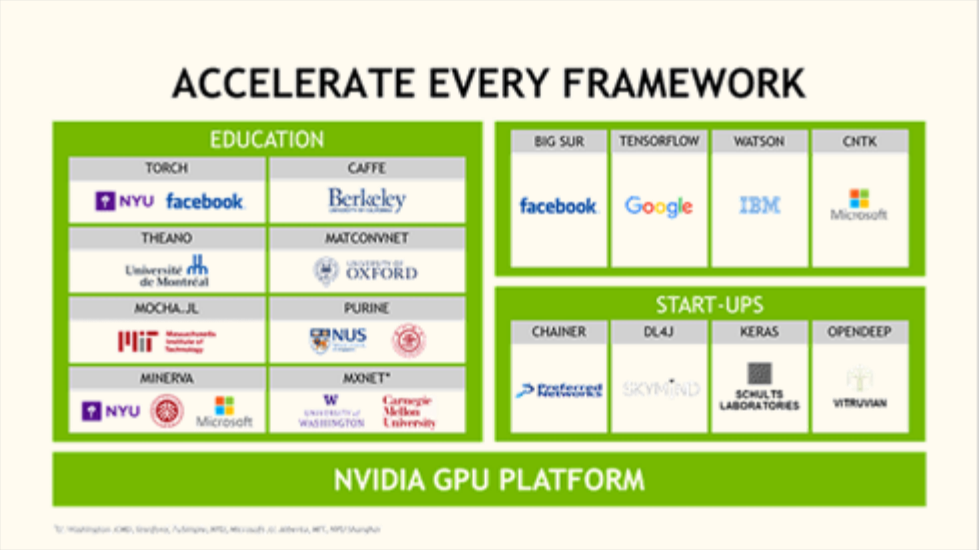
GPU 可以提供相当于 2,000 颗 CPU 的深度学习性能。此后，大家纷纷用 GPU 加速深度学习。从 2013 年开始 GPU 开始被应用于各行业的深度学习加速中，至 2015 年使用 GPU 加速 AI 的企业已经达到 3409 家，短短两年时间增长了 33 倍。

图表 7：GPU 从 2013 年开始应用于各行业的深度学习加速中



来源：英伟达官网，国金证券研究所

图表 8：众多教育机构、科技巨头公司及初创公司都在使用 GPU 加速深度学习



来源：英伟达官网，国金证券研究所

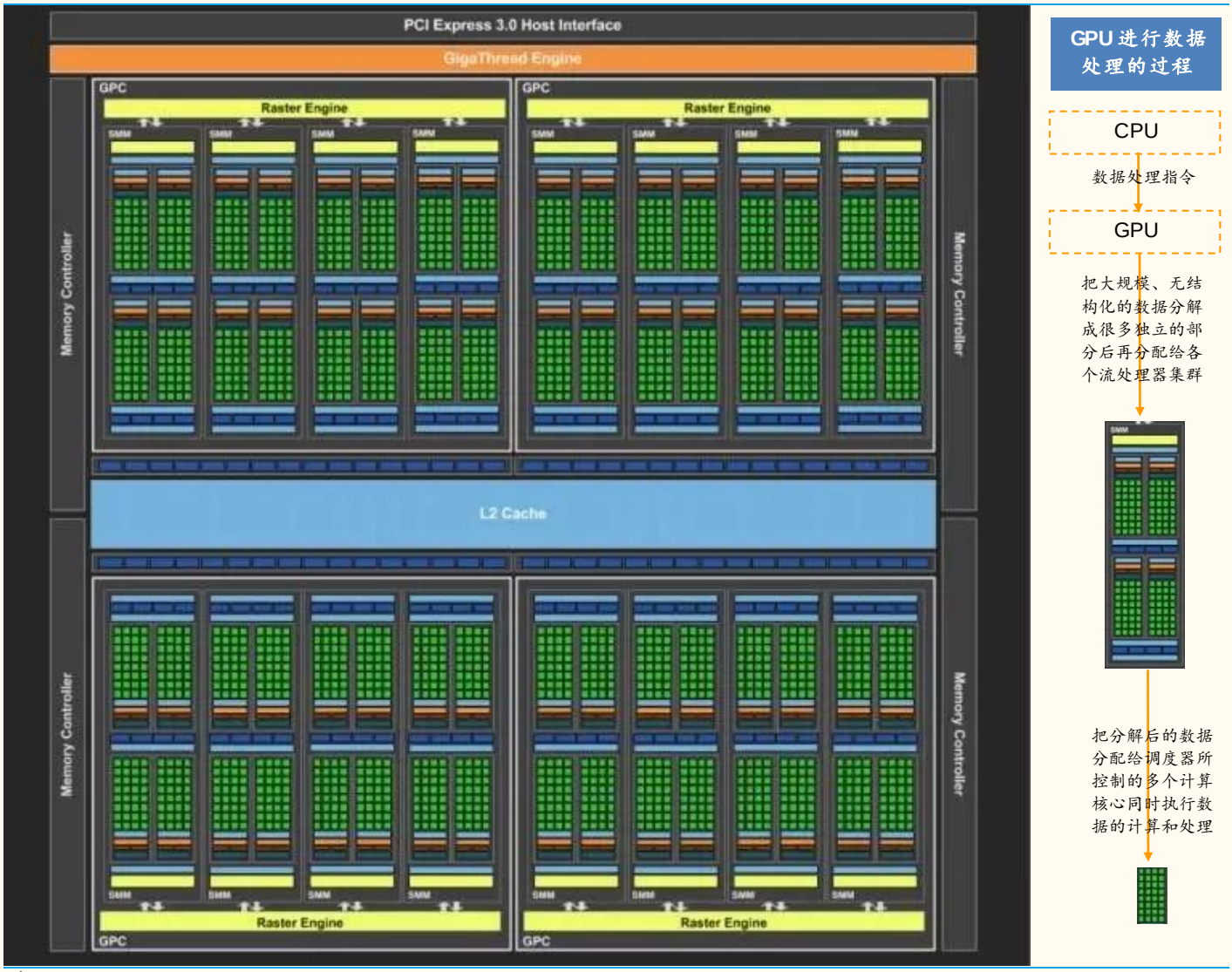
- **由于架构设计不同，使得 GPU 适用于计算密集型程序以及并行计算，而 CPU 擅长于逻辑运算和串行计算，这使得 GPU 比 CPU 更适合用于深度学习。**（注：串行计算可以简单理解为一个人负责完成 10 道计算题，而并行计算则是由 10 个人分别同时完成这 10 道算术题中的一道题，后者所花时间远小于前者，解决问题的速度更快）。CPU 架构设计以低延时为导向，而 GPU 架构设计以大吞吐量为导向。从下图可以看出，GPU 的架构相对于 CPU 而言一个显著的特点是“很多的 ALU+很少的 Cache+很少的 Control”，这样使得 GPU 并行计算能力很强，数据吞吐量大（很多的 ALU），但是牺牲了 CPU 所具备的低延迟性能（因为 Cache 和 Control 很少）。

图表 9：CPU 是基于低延时的设计而 GPU 是基于大吞吐量的设计

- ✓ 运算器（ALU）：GPU>>CPU，ALU 主要用于完成算数运算（加减乘除）及逻辑比较（比较两个数的大小），GPU 主要通过大量的 ALU 实现了大吞吐量，但同时占用了缓存和控制器的空间，所以牺牲了低延迟性能。
- ✓ 控制器（Control）：CPU>GPU，负责从内存中读取指令，把指令翻译后发送到其他部件中，并命令其他部件予以执行，从图中可以看出 CPU 具有复杂且强大的逻辑控制单元，这使得 CPU 可以通过程序分支预测能力来降低延时。GPU 的控制单元主要是把多个访问合并为数量较少访问。
- ✓ 缓存（Cache）：CPU>GPU，CPU 的大缓存可以存储数据，当需要计算这些数据时，直接从缓存里读取即可，从而降低了延迟；但是 GPU 的缓存很少，与 CPU 的缓存主要用于存储需要被访问的数据不同，GPU 的缓存主要是用于合并那些对同一个数据进行访问的多个线程，合并之后由缓存去访问 Dram 从而获取数据（因为数据被保存在 Dram 而不是缓存里），所以这里相对于 CPU 而言存在着延迟问题，因为 ALU 占用了太多空间所以缓存很小，使得 GPU 牺牲了部分延迟。

来源：Nvidia 官网，国金证券研究所

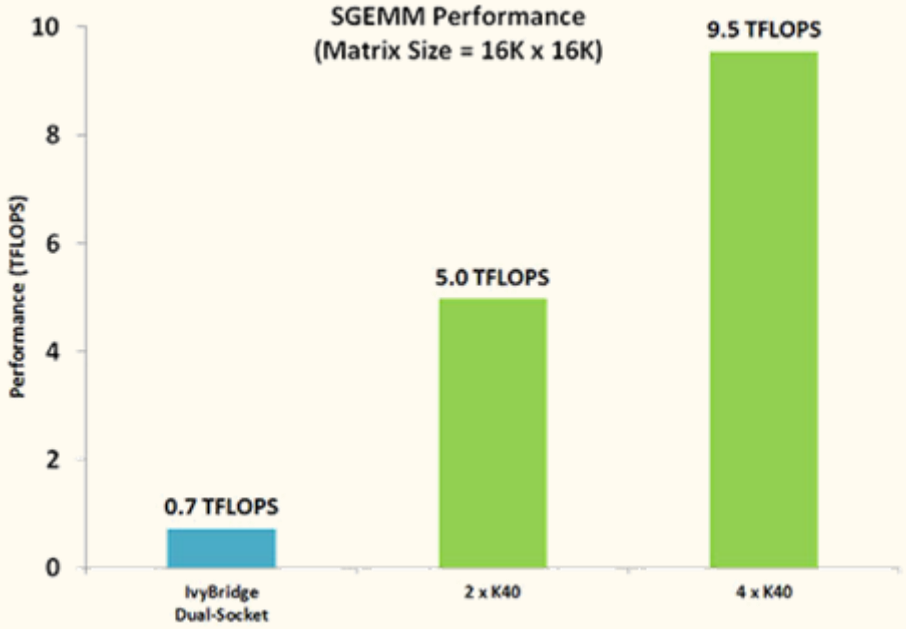
图表 10：英伟达基于 Maxwell 构架的 GPU 结构及其数据处理过程



✓ 这颗代号 GM200 的显示核心主要由 **4 个图形处理集群**（GPC：Graphics Processing Clusters），**16 个流处理集群**（SMM：Steaming Multiprocess）和 **4 个 64bit 显存控制器**组成。每个流处理集群中包含了 4 个调度器（Warp），**每个调度器又控制着** 32 个逻辑计算核心（Core），因此这颗 GPU 芯片包含 32×4×16=**2048 个 Core**，这些 Core 是实现逻辑计算的基本单元

来源：CSDN，国金证券研究所

图表 11：利用 GPU 加速后，浮点运算性能得到极大提升

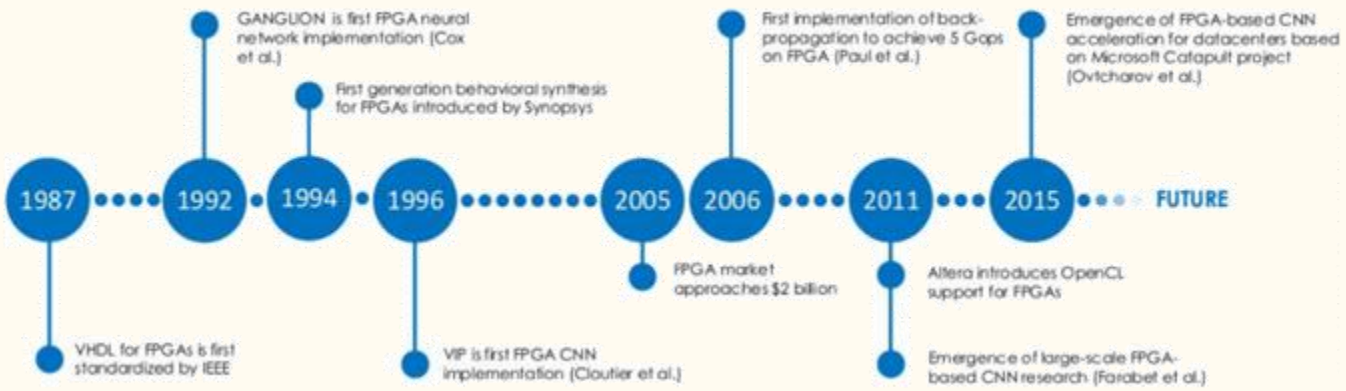


来源：Nvidia 官网，国金证券研究所

FPGA：架构最为灵活的可编程加速器

- FPGA 在 1992 年被首次应用于神经网络，1996 年被首次用于卷积神经网络（CNN）的加速，但是因为那时的 FPGA 没有密集的乘法累加器（MAC）以及 FPGA 尺寸的限制，导致计算效率较低且算法精度较低，应用效果不佳。但是近年来 FPGA 相关技术已经获得显著突破，其中最显著的就是晶体管尺寸不断减小，FPGA 上硬化计算单元数量的增多使得 FPGA 核心逻辑单元的密度显著增加。另外在 2011 年，Altera 推出 Open CL 语言使得在 FPGA 上实现异构计算和并行计算成为可能，极大地推动了 FPGA 在深度学习中的应用。2015 年，微软利用 FPGA Stratix V D5 在 ImageNet 竞赛上实现了 134 张/秒的吞吐量，是第二名参赛者的 3 倍，而且功率仅为 25W，这一事件对于 FPGA 在深度学习中的应用而言具有里程碑的意义。从此，百度、腾讯、亚马逊等科技巨头公司开始纷纷将 FPGA 部署到数据中心中。

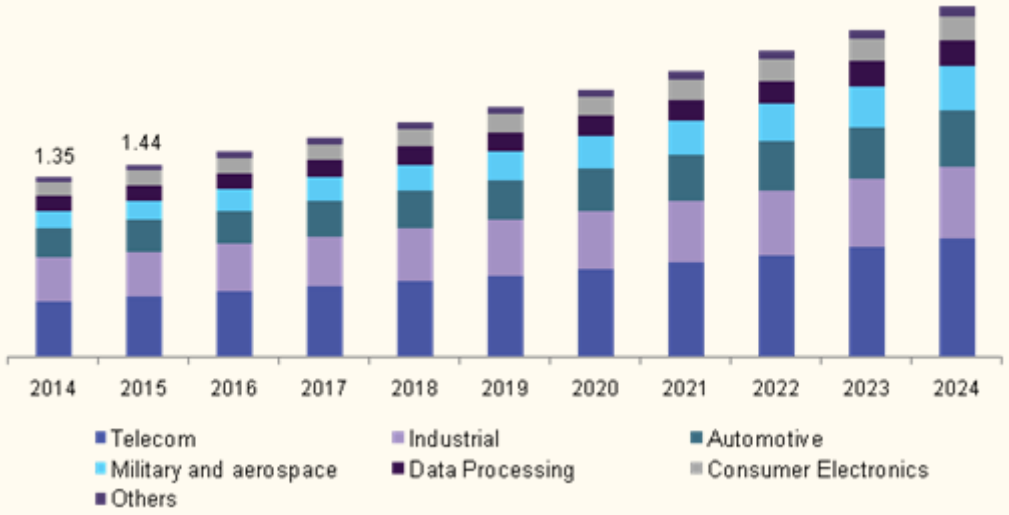
图表 12：FPGA 在深度学习领域应用的重大事件历程



来源：论文《Deep Learning on FPGAs: Past, Present, and Future》，国金证券研究所

- **在 FPGA 被用于深度学习之前，FPGA 主要有 3 大应用方向**：（1）**通信设备的高速接口电路设计**，这一方向主要是用 FPGA 处理高速接口的协议，并完成高速的数据收发和交换；（2）**数字信号处理方向/数学计算方向**（例如金融、医疗数据分析）；（3）**SOPC**，即利用 FPGA 这个平台搭建的一个嵌入式系统的底层硬件环境，然后设计者在上面进行嵌入式软件开发。

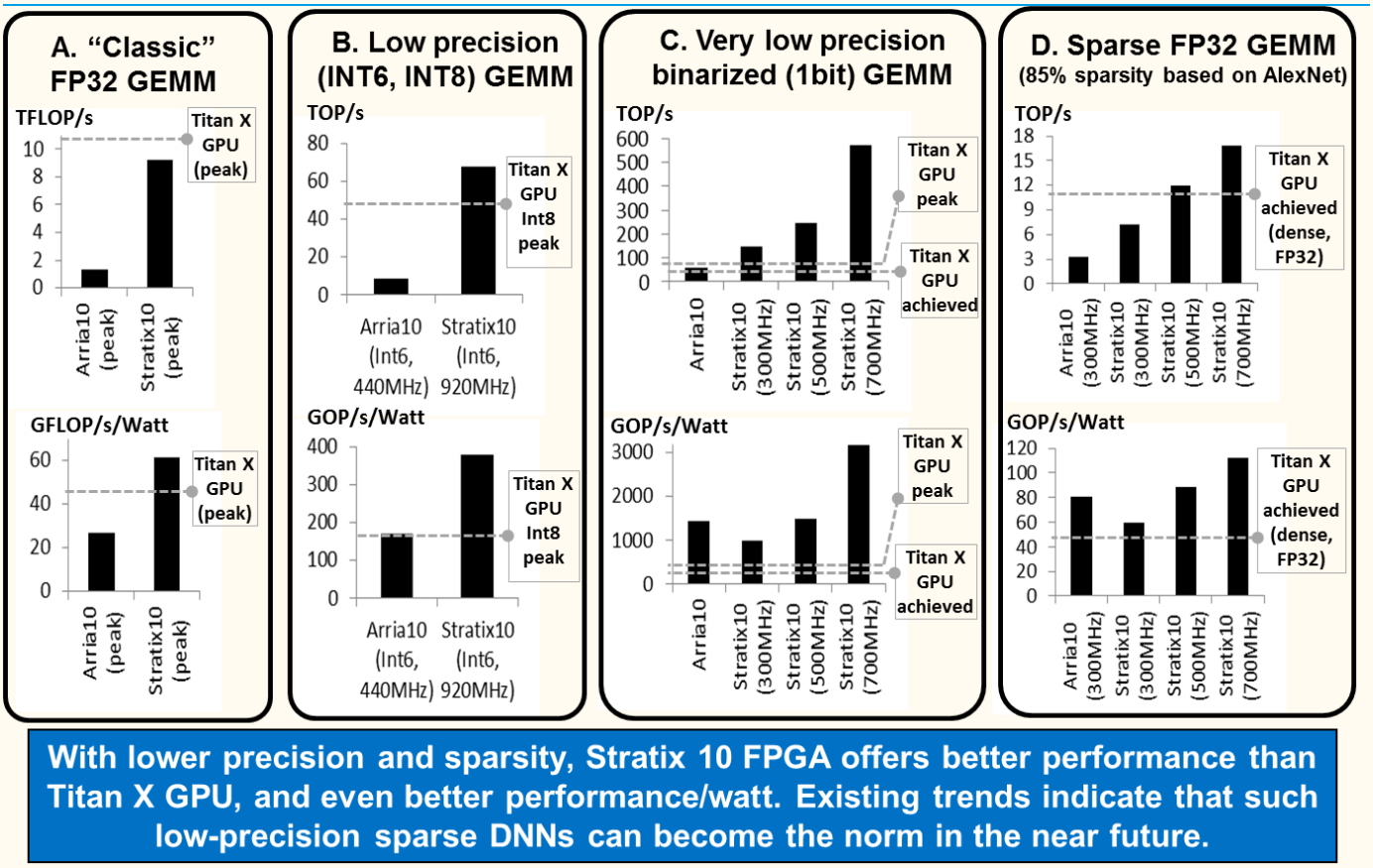
图表 13：美国市场 FPGA 一半以上被用于通信和工业领域（单位：十亿美元）



来源：Grand View Research，国金证券研究所

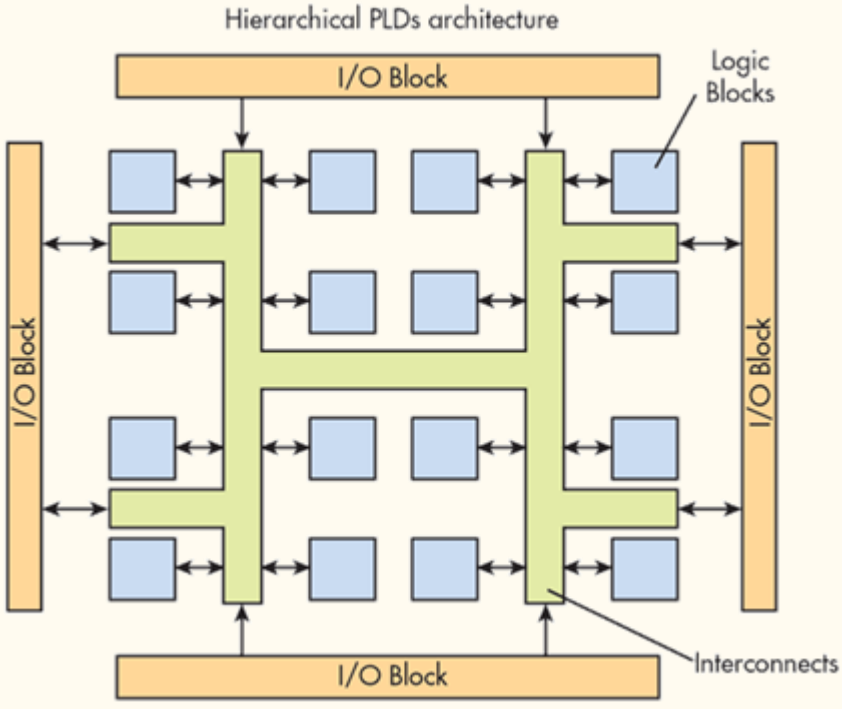
- 神经网络计算中有大部分时间都用在矩阵乘法（GEMM）上，英特尔曾经对 FPGA 和 GPU 在矩阵乘法中的性能表现做过测试，试验结果表明 FPGA 的性能及性能/功耗比都远胜于 GPU。FPGA 的性能及性能/功耗比高于 CPU/GPU 本质上是因为前者不依赖冯·诺依曼结构体系而后者属于冯氏结构。CPU/GPU 是指令译码执行、共享内存的；而 FPGA 是无指令、无需共享内存的体系结构。从架构上看，FPGA 主要由可编程逻辑单元、可编程连接网络、可编程的输入输出（I/O）模块构成，与 CPU/GPU 由运算器、控制器等组成的逻辑结构完全不同。FPGA 每个逻辑单元的功能在重编程时就已经确定，不需要指令；对于通信的需求，FPGA 每个逻辑单元与周围逻辑单元的连接在重编程时就已经确定，并不需要通过共享内存来通信。

图表 14：在矩阵相乘（GEMM）测试中 FPGA 性能均好于 GPU



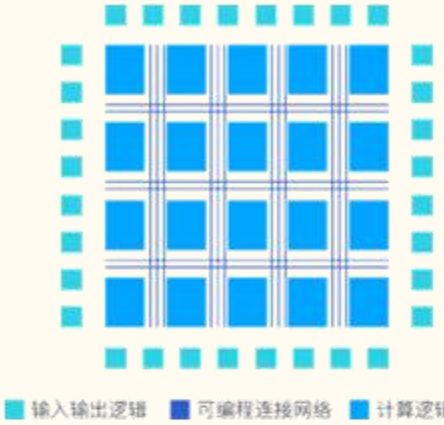
来源：英特尔论文《Can FPGAs beat GPUs in Accelerating Next-Generation Deep Neural Networks》，国金证券研究所

图表 15：FPGA 由可编程逻辑快、连接网络、输入输出模块构成



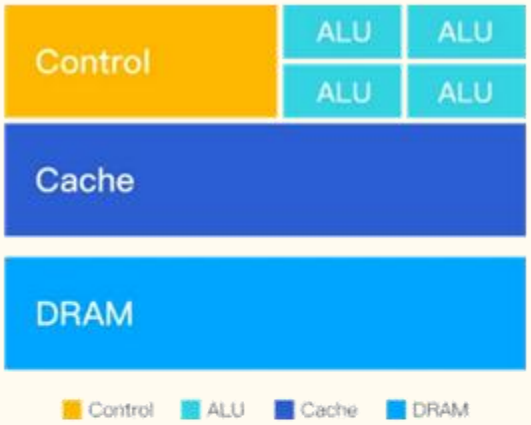
来源：Electronic Design，国金证券研究所

图表 16：FPGA 与 CPU 在处理器逻辑结构、特点、使用场景方面的对比



FPGA

核心数量：	数千个计算逻辑部件
处理器逻辑结构：	非冯诺依曼结构，实现并行操作
特点：	1. 强大的输入输出逻辑，低延时、高吞吐 2. 数量庞大的计算逻辑部件用于建造计算模块 3. 可编程连接网络主要用于连接不同计算逻辑部件
使用场景：	适用有非常大量的重复性、需要并行计算的工作，建议用于基因组学研究、金融分析、图像压缩和实时视频处理等高性能计算服务



CPU

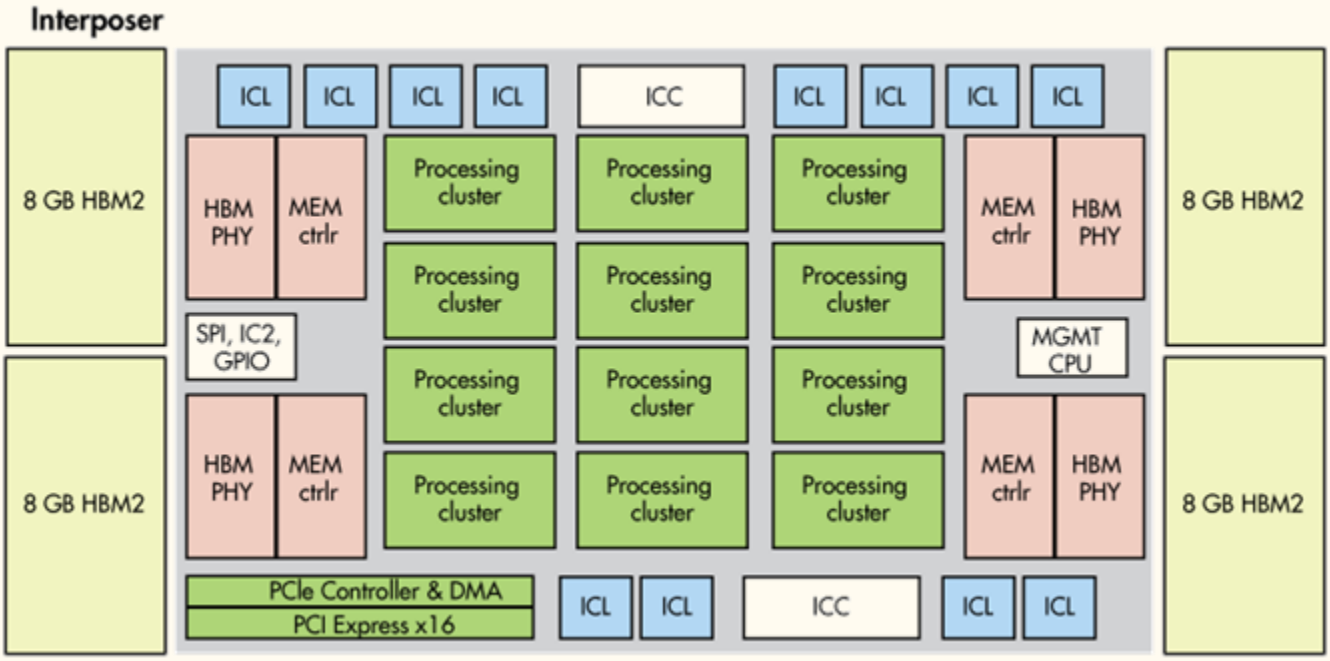
核心数量：	几十个核心
处理器逻辑结构：	冯诺依曼结构，串行的执行一系列指令
特点：	1. 复杂的逻辑控制单元 2. 强大的算数运算单元 3. 较大缓存用户降低延时
使用场景：	逻辑控制、串行运算的程序

来源：腾讯云，国金证券研究所

专属 ASIC 芯片：先行者的游戏

- **目前英特尔和谷歌都在深度学习专用芯片方面有所布局。**2016 年 8 月英特尔斥资 4.5 亿美元收购人工智能 ASIC 芯片商 Nervana，从收购至今英特尔共投入约 3.5 亿美元用于研发“服务于 DNN 的 Nervana 软硬件一体化平台”。2017 年英特尔向外公布整合了 Nervana 之后的第一款 DNN 专用芯片 Lake Crest。2016 年 5 月谷歌在 I/O 大会上发布第一代 TPU，可以用于深度学习推理阶段；2017 年 5 月谷歌发布第二代 TPU，可以同时用于训练和推理阶段。目前 TPU 已经部署在了 Google Compute Engine 平台上，可用于图像和语音识别，机器翻译和机器人等领域。

图表 17：英特尔 Lake Crest 利用处理集群优化 AI 应用



来源：Electronic Design，国金证券研究所

图表 18：用 64 个第二代 TPU 构建的“TUP POD”，可以提供 11500 万亿次/秒浮点运算能力

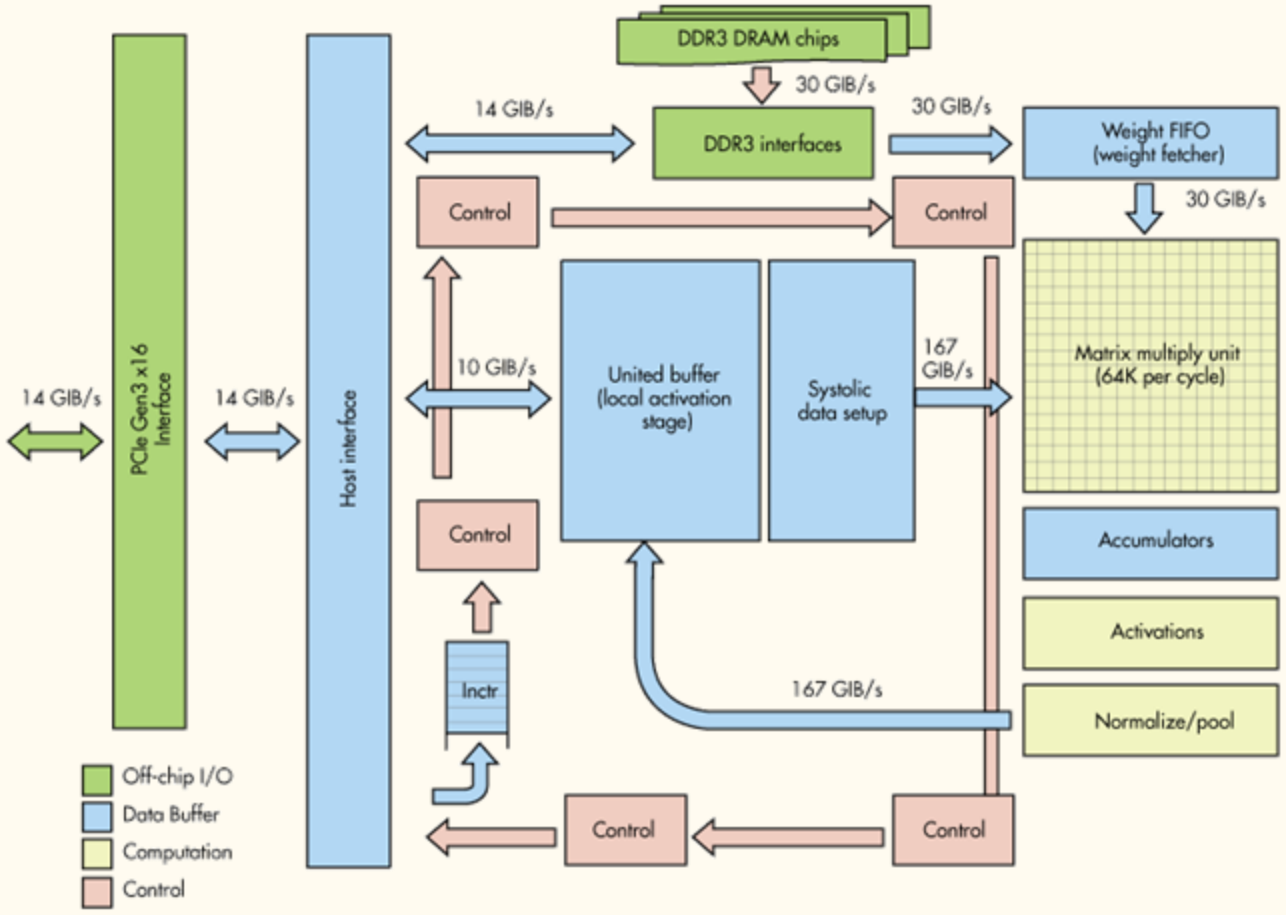
来源：Google Blog《Build and train machine learning models on our new Google Cloud TPUs》，国金证券研究所

■ **TPU 相对于 FPGA/GPU/CPU 而言具有如下优点：**

1) **计算效率更高：**TPU 的特别之处在于其为深度学习所特别设计的矩阵相乘单元。深度学习大部分算法运行实际上就是在进行矩阵相乘运算。谷歌在 TPU 芯片中专门设计了 8bit 的矩阵乘法单元 MUX（Matrix Multiply Unit）。TPU 之所以在运行深度学习算法时比 CPU/GPU/FPGA 快，一方面是因为设计了矩阵相乘单元（MUX），另一方面是因为 MXU 有着与传统 CPU、GPU 截然不同的架构，称为脉动阵列（systolic array）。CPU 和 GPU 在每次运算中都需要从多个寄存器（register）中进行存取；而 TPU 的脉动阵列将多个运算逻辑单元（ALU）串联在一起，复用从一个寄存器中读取的结果，从而避免了数据和指令传输速度跟不上计算速度的限制（即前面所提到的冯·诺依曼瓶颈）。

2) **功耗低：**FPGA 的通用性必然导致冗余。FPGA 的运算电路基于查找表，比如说，FPGA 内部有 1000 万个自定义逻辑部件，一个 4 输入的查找表单元需要 96 个晶体管来支持，而在 ASIC 上来实现只需要 10 个左右。这些冗余也必然体现在芯片的面积和功耗上。

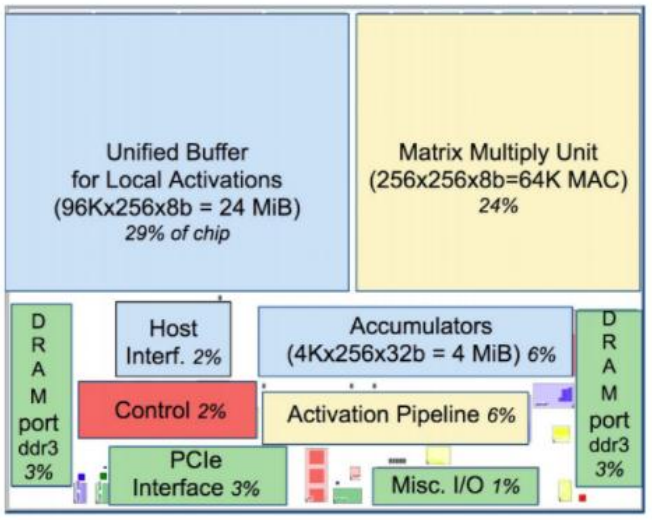
图表 19：TPU 用 8bit 的矩阵乘法单元（MUX）来对 DNN 进行数字处理



✓ MXU 的脉动阵列包含 256×256 =65,536 个 ALU，也就是说 TPU 每个周期可以处理 65,536 次 8 位整数的乘法和加法。TPU 以 700 兆赫兹的功率运行，也就是说，它每秒可以运行 65,536 × 700,000,000 = 46 × 1012 次乘法和加法运算，或每秒 92 万亿（92 × 1012）次矩阵单元中的运算。

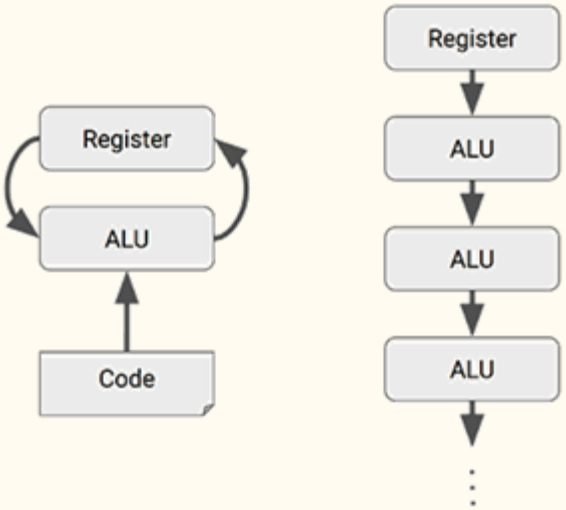
来源：Electronic Design，国金证券研究所

图表 20：MUX 运算单元占整个 TPU 芯片一半面积



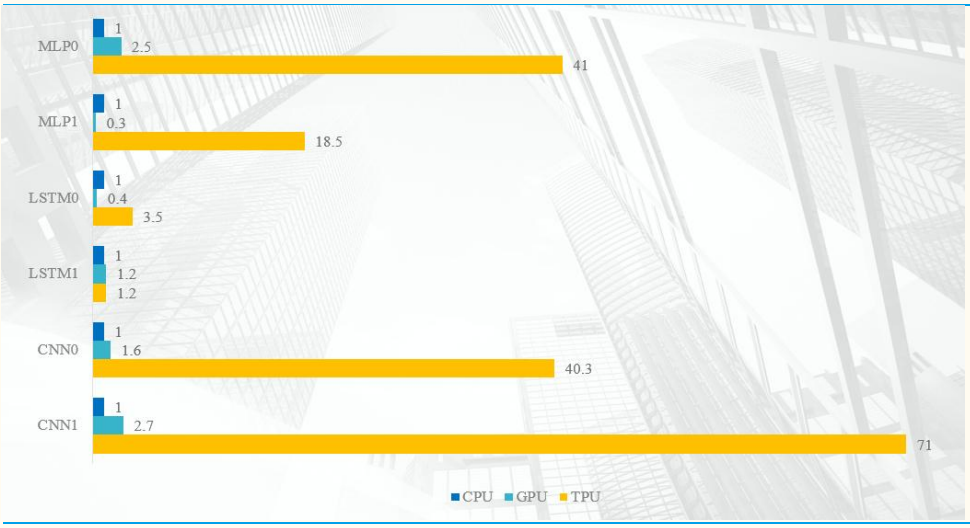
注：黄色-运算单元；蓝色-数据单元；绿色-I/O;红色-控制逻辑单元

图表 21：和 CPG/GPU 相比（左）TPU（右）采用了截然不同的“脉动阵列”



来源：Google Blog《Build and train machine learning models on our new Google Cloud TPUs》，国金证券研究所

图表 22：CPU、GPU、TPU 在 LSTM、CNN 等六种神经网络上的性能表现



来源：量子位，国金证券研究所

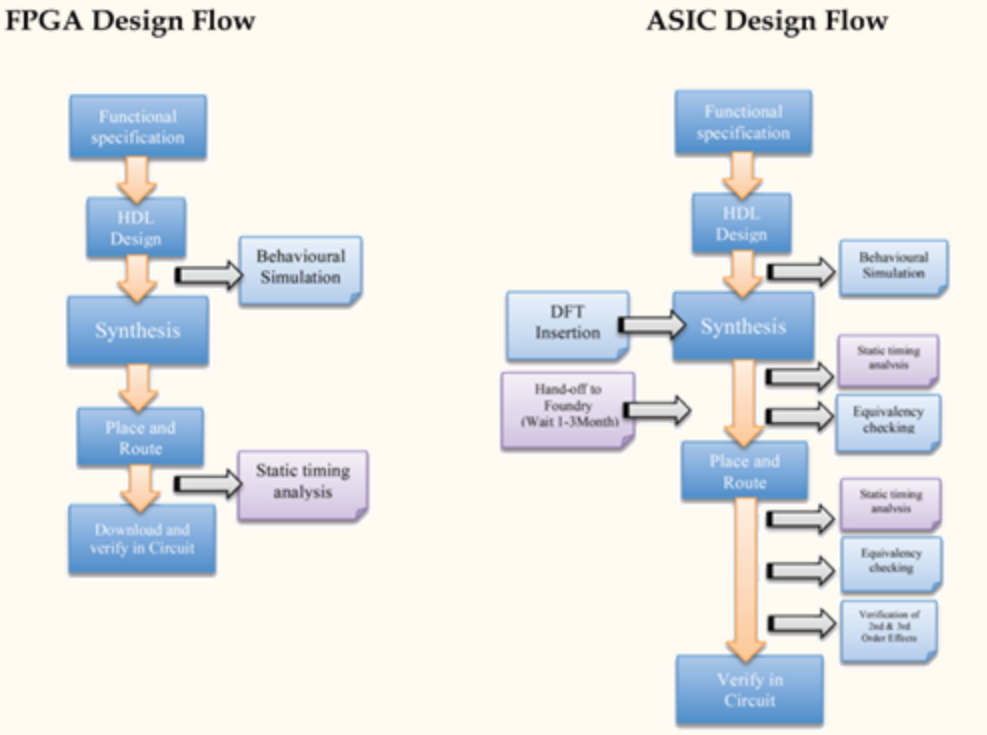
■ **但是以 TPU 为代表的人工智能 ASIC 芯片也具有以下不足：**

1) **太专用而导致灵活性低：**MXU 只适用于 8bit 矩阵乘法运算，不能进行 16bit 矩阵乘法运算，也不能进行减法、除法、倒数等基本计算。

2) **开发周期相对较长：**TPU 从设计到验证、构建和部署到数据中心里共耗时 15 个月；而 FPGA 从设计到部署一般只需要 6 个月。

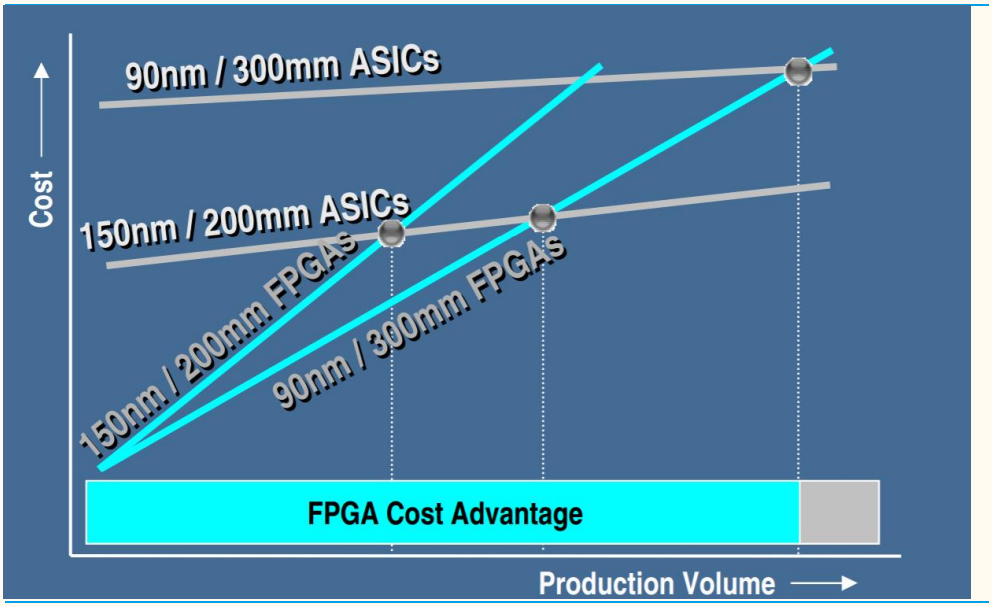
3) **ASIC 一次性开发成本较高：**产量较低时 FPGA 成本小于 ASIC，产量较高时 ASIC 成本小于 FPGA。

图表 23：ASIC 的设计环节比 FPGA 要复杂得多，导致开发周期较长



来源：INTECH，国金证券研究所

图表 24：产量较低时 FPGA 成本小于 ASIC，产量较高时 ASIC 成本小于 FPGA



来源：Xilinx，国金证券研究所

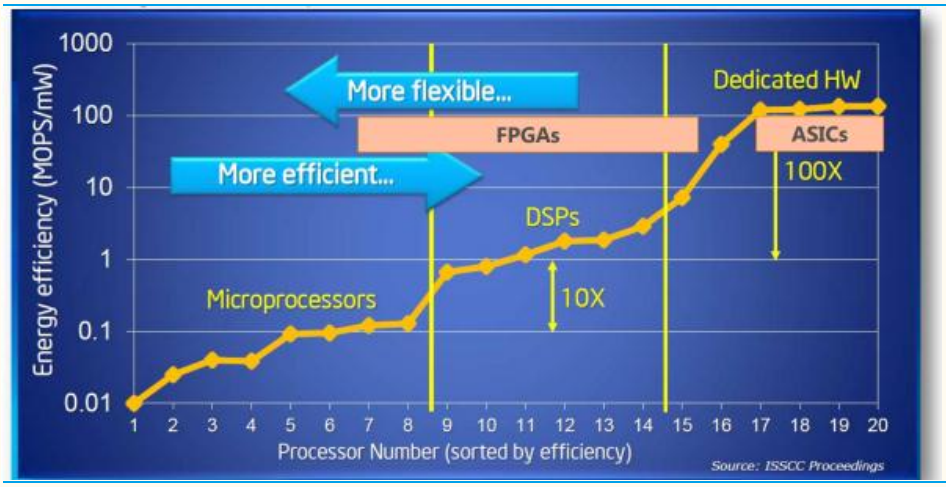
三种方案对比

经过上述分析可以发现以下几个特点：芯片专用性越强，计算效率越高，能耗越低；但与此同时灵活性降低，开发难度增加。具体各芯片特点可作如下归纳：

- **计算效率**方面，CPU<GPU<FPGA<ASIC

- 灵活性方面，CPU>GPU>FPGA>ASIC
- 开发难度方面，CPU<GPU<FPGA<ASIC

图表 25：专用芯片（ASIC）的计算效率虽然最高，但是灵活性最低



来源：: Bob Broderson(Berkeley Wireless group)，国金证券研究所

图表 26：CPU、GPU、FPGA、ASIC 在处理计算密集型任务时的性能比较

体系结构	吞吐量	延迟	功耗	灵活性
CPU	1T	N/A	100W	低
GPU	10T	1ms	300W	高
FPGA(Stratix V)	1T	1us	30W	高
FPGA (Stratix 10)	10T	1us	30W	高
ASIC	10T	1us	30W	低

来源：微软亚洲研究院，国金证券研究所

图表 27：CPU、GPU、FPGA、ASIC 的实现比较

硬件	CPU	GPU	FPGA	ASIC
开发难度	小	较小	大	很大
增加功能	容易	容易	难	不能增加
硬件升级	无需修改代码	无需修改代码	需要修改代码	
成本	低	低	高	很高
开发周期	短	短	长	很长
应用领域	通用		大型企业的线上数据处理中心，和军工单位	消费电子，如移动终端

来源：《CUDA 基本介绍》，国金证券研究所

三、终端 AI 芯片：应用场景驱动，市场前景广阔

随着人工智能场景的应用深入渗透到行业的各个领域，在终端，推理（Inference）阶段的计算能力越来越成为瓶颈。一些对即时性要求很高的应用场景，已经无法通过在云端进行推理计算的方式满足，终端 AI 芯片加速成为了必选的方案，一些典型的场景案例有无人驾驶、机器视觉、消费电子等。于是，针对各种应用场景，均有定制化的 AI 芯片推出。

图表 28：终端 AI 芯片场景案例——翻译机

来源：国金证券研究所绘制

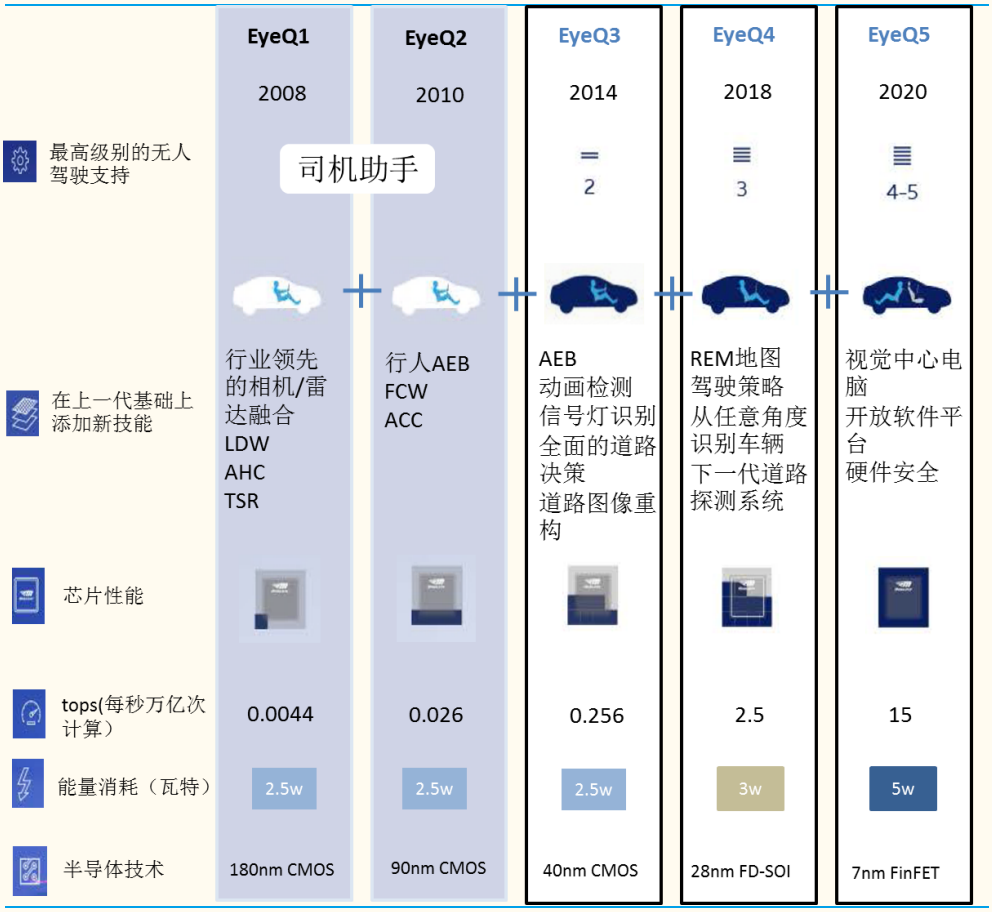
无人驾驶场景：以 Mobileye 公司 EyeQ 系列芯片为代表

- **无人驾驶有望成为汽车产业的下一代革命性技术。**无人驾驶技术是指让汽车自己拥有环境感知、路径规划并自主实现车辆控制的技术，是用电子技术控制汽车进行的仿人驾驶或是自动驾驶。无人驾驶技术是多个技术的集成，其系统包含了多个传感器，包括长距雷达、激光雷达、短距雷达、车载摄像头、超声波、GPS、陀螺仪等。每个传感器在运行时都不断产生数据，而且系统对每个传感器产生的数据都有很强的实时处理要求。根据市场预测，到 2025 年，全球无人驾驶汽车市场规模将达到 420 亿美元。2030 年，将有 1.2 亿辆不同程度的无人驾驶汽车上路；2035 年，无人驾驶汽车将占全球汽车销量的四分之一。
- **人工智能技术已经被广泛地应用到无人驾驶技术上。**随着神经网络算法的不断优化，在无人驾驶领域的使用也越来越普及。无人驾驶汽车通过大量的真实上路数据进行训练，不断提升其神经网络模型的智能水平，从而可以对未来的输入场景做出判断。由于汽车驾驶场景的特殊性，推理（Inference）阶段必须在几毫秒之内完成，否则判断的延时将导致出现安全隐患，因而车载端定制化的 AI 芯片成为了必选项。
- **Mobileye 是无人驾驶芯片研发领域的先驱。**Mobileye 是一家位于以色列的公司，于 1999 年成立，2007 年推出首款产品，2014 年 8 月 1 号在纽交所上市。公司主要从事汽车工业的计算机视觉算法和驾驶辅助系统芯片技术的研究。作为无人驾驶解决方案领域的先行者，Mobileye 的产品覆盖了全球 50 个国家。根据公司官方资料显示，截至 2015 年底，Mobileye 在全球有 1000 万的装载量。2017 年 3 月，芯片巨头英特尔公告将以 153 亿美元收购 Mobieye 公司。
- **EyeQ 系列芯片专注于无人驾驶场景的 AI 优化。**EyeQ 系列芯片是 Mobieye 转为无人驾驶场景打造的针对 AI 算法优化的硬件平台，其中，该系列最新产品是 EyeQ5。2016 年 5 月，Mobileye 联合意法半导体，发布了下一代视觉系统芯片——EyeQ5。该产品装备 8 枚多线程 CPU 内核，同时还会搭载 18 枚 Mobileye 的下一代视觉处理器。据 Mobileye 官方提供的资料来看，EyeQ5 加速器核心经过了优化，如计算机视觉、信号处理、机器学习任务以及深度神经网络。EyeQ5 具有异构性，完全可编程的加速器，芯片内置的四种类型加速器均经过其系列算法优化。加速器结构的多样性，使每种应用执行不同任务时，能够选择最适合的加速器核心，节省计算时间和功耗。这种优化的配置使 EyeQ5 在低功耗封装中，不仅能够提供超级

计算能力，还实现了经济的被动散热。EyeQ5 的运算性能达到了 12 Tera/每秒，相比 EyeQ4 的运算性能提升了 5 倍左右，但其能耗却不到 5W，其技术特点非常适合车载环境。

- **不止是芯片，更是平台。**Mobieye 提供的不仅仅是一款芯片，同时也为汽车制造商和一级供应商提供自动驾驶所需的全套硬件加速算法和应用软件。包括汽车标准操作系统和一款完整的软件开发包（SDK），以便让客户将自己的算法嵌入 EyeQ5，获得差异化的个性解决方案。通过软件开发包，开发者可进行神经网络的原型设计和开发。EyeQ5 在设计上，采用硬件虚拟化和 CPU 与加速器的全高速缓存一致性等结构元素，更适合成为开源软件平台。此外，EyeQ5 还内嵌了基于集成式硬件安全模块 HSM 打造的安全防护系统。这使得系统集成商能够支持 OTA 软件升级，车内通信安全也能得到进一步保障，信息安全的基础源自其加密的存储设备。Mobieye 围绕 EyeQ5 芯片打造了一个完整的芯片生态系统，实现了平台效应。

图表 29：EyeQ 系列芯片针对无人驾驶算法做硬件优化

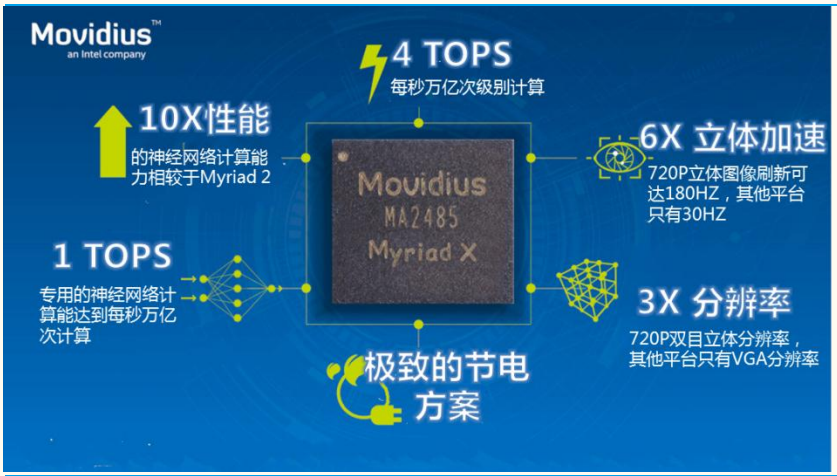


来源：Mobieye 官网，国金证券研究所

计算机视觉场景：以 Movidius 的 Myriad 系列芯片为代表

- **视频结构化分析是计算机视觉发展的大势所趋。**传统的视频数据是一些像素信息的集合，缺乏视频内容的结构化信息，无法实现更深层次的检索、比对等更有意义的操作。随着神经网络算法的应用，视频中的人脸识别成为了可能，视频内容呈现更丰富的信息，同时也对前端视觉设备的计算能力提出了更高的要求，基于计算机视觉的 AI 芯片应运而生。

图表 30：Movidius 的 Myriad X 芯片针对视觉计算有多项性能提升



来源：Movidius 官网，国金证券研究所

- **Movidius 主打产品 Myriad X 芯片专为计算机视觉场景优化。**Movidius 是一家计算机视觉领域的创业公司，其于 2016 年 9 月被英特尔全资收购。公司目前主打产品 Myriad X 视觉处理器（VPU），是一款低功耗 SoC，主要用于基于视觉的设备的深度学习和 AI 算法加速，比如无人机、智能相机、VR/AR 头盔等。在此之前，其前一款芯片产品 Myriad 2 于 2014 年首次推出。Myriad X 能在同一功率范围内的深度神经网络 DNN 推理中，提供 10 倍于 Myriad 2 的性能。基于集成在芯片上的 DNN 加速器。Myriad X 的 DNN 推理吞吐量能达到每秒超过一万亿次运算（TOPS），而理论运算量能达到 4TOPS。Myriad X 的主要特征参数有：
  - 四个支持 C 语言的可编程 128 位 VLIW 矢量处理器（C-programmable 128-bit VLIW），源自 Myriad 2 的可配置 MIPI 通道，以及扩展了的 2.5 MB 芯片内存，更多固定功能的成像/视觉加速器。
  - 向量单元是针对计算机视觉工作负载优化的 SHAVE 处理器,支持最新的 LPDDR4。
  - 4K 硬件编码，支持 30 Hz（H.264/H.265）和 60 Hz（M/JPEG）。
  - 附带 SDK，其中包含神经网络编译器和专用的 FLIC 框架。
- Movidius 的 Myriad 系列芯片已经在多个计算机视觉的下游领域应用。未来，随着视频结构化信息分析功能成为标准配置，Myriad 系列芯片将被更多视觉应用终端采用。

图表 31：Movidius 的 Myriad 系列芯片已经在多个下游领域应用

时间	合作公司	产品
2014 年	谷歌	Project Tango 项目，用 Myriad 1 帮助打造室内三维地图
2016 年	大疆	“精灵 4”无人机的视频分析芯片
2016 年	宇视	基于 Myriad 2 芯片的全系中高端摄像机、卡口抓拍机产品
2016 年	海康威视	全局摄像机使用的是 Movidius Myriad 2 Vision 处理器

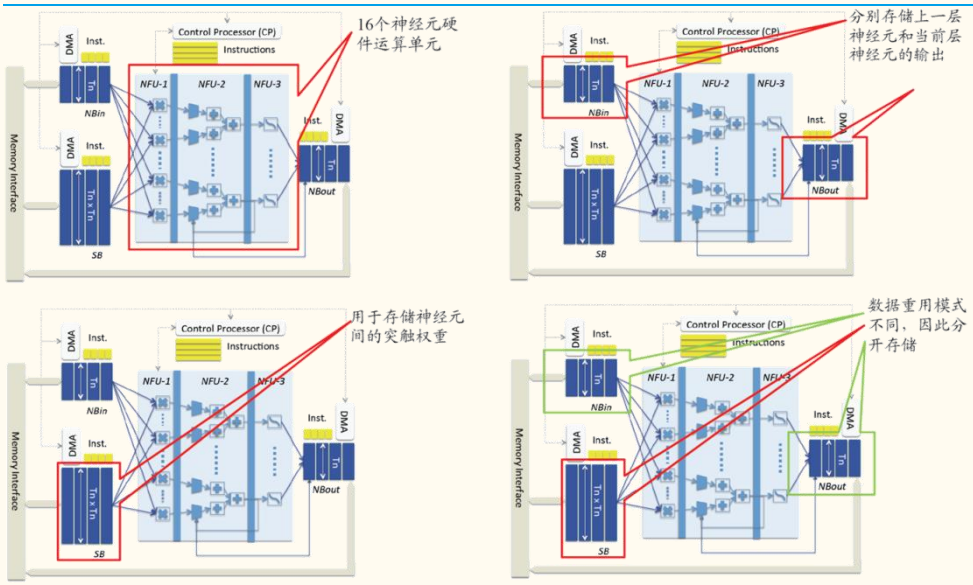
来源：海康威视、大疆科技等官网，国金证券研究所

消费电子场景：以寒武纪 Cambricon-1A 为代表

- **AI 应用正在向消费电子端推进。**随着语音识别、视频识别等技术的成熟，手机、智能音箱等消费电子端的相关 AI 应用也逐渐被大众所接受。简单的服务器计算模式已经无法满足交互的即时性需求，终端的高性能 AI 计算芯片已经成为了消费电子的下一个硬件升级方向。

- **寒武纪公司是较早布局 AI 芯片领域的前沿科技公司。**寒武纪的前身是中国科学院计算技术研究所下的一个课题组，其早在 2008 年就已经开始研究神经网络算法以及芯片设计。寒武纪第一代方案在 2012 年推出，65nm 工艺下功耗为 0.485W，面积 3.02mm ²。平均性能超过主流 CPU 核的 100 倍，但面积和功耗仅为 1/10，非常适合消费电子的应用场景。

图表 32：寒武纪 1 号神经网络处理器架构



来源：寒武纪神经网络计算机介绍，国金证券研究所

- **寒武纪 1A 处理器针对消费电子级别的应用做了成功的优化。**寒武纪公司 2016 年发布的寒武纪 1A 处理器（Cambricon-1A）是首款商用深度学习专用处理器，其集成到终端 SoC 芯片，每秒可处理 160 亿个虚拟神经元，每秒峰值运算能力达 2 万亿虚拟突触，性能比通用处理器高两个数量级，功耗降低了一个数量级。在硬件设计上，该芯片对硬件运算单元分时复用，从而以较小尺的度支持大规模神经网络，同时通过优化片上存储层次尽量减少访存次数降低能耗。在指令集的选择上，其采用自主开发的神经网络处理器指令集 DianNaoYu，参照 RISC（精简指令集）设计思想，所有指令长度都是 64bit，有效简化指令译码器的负担，减少功耗以及芯片面积。一条指令即可完成一组神经元处理，优化了计算数据在芯片上的传输，模拟实验表明，采用 DianNaoYu 指令集的深度神经网络处理器相对 X86 指令集处理器有两个数量级的性能提升。

图表 33：寒武纪 1 号芯片和同期主流芯片对比

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

来源：寒武纪神经网络计算机介绍，国金证券研究所

- **手机端的 AI：硬件先行，应用跟进。**当前手机端真正能使用到人工智能处理器的应用主要包括实时的语音识别、视频语义分析、AR 应用等，总体上来说还不够丰富。 我们认为，未来人工智能芯片在手机端的搭载尝试，将会为手机应用开发者开启一个新的硬件环境，其强大的计算能力一定会孕育更多更出色的移动端 AI 应用。回顾移动互联网产业的发展路线，4G 网络先行，促进了移动端 APP 的繁荣，从而提供了更多网络付费用户，双方形成了正反馈循环。硬件先行，应用跟进，将会是当前时间点终端 AI 的发展方向。我们认为，未来优质的 AI 应用将大大促进 NPU 的出货量，从而

摊薄芯片的研发成本，两者同样形成良好的正反馈循环，将 AI 芯片推向更广阔的市场。

图表 34：AI 芯片和 AI 场景应用有望形成正反馈循环

来源：国金证券研究所绘制

四、投资建议

我们认为，在人工智能的变革正在深入渗透到各行各业的时代，AI 专属芯片作为计算能力的保障，将迎来巨大的需求。从云端的高性能服务器到终端的视频监控、消费电子等领域，在产业链上均有机会享受这场变革带来的红利。

中科曙光(603019.SH)

- 公司是国内领先的高端计算机、存储、云计算和大数据系统提供商，产品能全面满足用户从超级计算机到普通 PC 服务器的各项应用需求，在互联网、金融、电信、生物、气象、石油、科研、电力等多个行业有着大量成功应用。公司掌握服务器等领域多项核心技术，累计申请专利 1937 项，其中发明专利 1496 项。公司与中科院计算所和 Nvidia 公司开展了深度学习战略合作，并发布了 XSystem 深度学习产品。该系统包含深度学习 XSharp 软件栈和 XMachine 硬件平台，为用户提供的一体化的深度学习软硬件整体解决方案。未来公司有望受益于服务器端 AI 计算能力需求的持续提升。

中科创达(300496.SZ)

- 公司是全球领先的智能终端操作系统及平台技术提供商，其多年来致力于提供智能终端操作系统平台技术及解决方案，助力并加速移动终端、智能硬件、智能汽车等领域的产品化与技术创新。公司参与了华为多款手机的芯片层软件优化、系统裁剪等工作，其在底层芯片领域多年的研发经验和客户基础，有望使公司在 AI 芯片普及后，较快切入 AI 产业链中。

浪潮信息(000977.SZ)

- 公司是国内领先的计算平台与 IT 应用方案供应商，其提供的云计算中心的服务器、存储等核心设备以及解决方案多年保持销量领先。2017 年 9 月，公司与百度共同发布人工智能 ABC 一体机。该一体机可支持 TensorFlow、Caffe、CNTK、PaddlePaddle 等几乎所有主流算法框架,面向模型训练 Training 和线上预测 Inference 两大类 AI 计算场景,可以开箱即用,是一款 allinone 解决方案。与国内人工智能巨头百度的合作，有望助力公司的云计算解决方案迅速打开互联网客户市场，从而占据 AI 服务器的制高点。

富瀚微(300613.SZ)

- 公司是模拟高清摄像机（ISP）SOC 市场龙头企业，并与第一大客户海康威视展开长期深度合作。公司第二大产品线网络摄像机（IPC）SoC 系列芯片部分型号也从 2014 年开始量产，目前在 IPC SoC 市场，海思、TI、安霸等大厂占据了 90%以上市场份额。公司在 2017 年 9 月 8 日发布了两款智能 IPC 芯片（FH 8830 和 FH 8630D），已经可以支持人脸识别、智能 编码等人工智能应用。在视频监控行业高清化、前端智能化的趋势下，公司正在开展人工智能相 关技术的跟踪研究。

五、风险提示

- 人工智能技术推进不达预期；
- 芯片性能提升无法突破计算瓶颈；
- 消费者对终端 AI 应用接受程度未达预期；

长期竞争力评级的说明：

长期竞争力评级着重于企业基本面，评判未来两年后公司综合竞争力与所属行业上市公司均值比较结果。

公司投资评级的说明：

买入：预期未来 6—12 个月内上涨幅度在 15%以上；
增持：预期未来 6—12 个月内上涨幅度在 5%—15%；
中性：预期未来 6—12 个月内变动幅度在 -5%—5%；
减持：预期未来 6—12 个月内下跌幅度在 5%以上。

行业投资评级的说明：

买入：预期未来 3—6 个月内该行业上涨幅度超过大盘在 15%以上；
增持：预期未来 3—6 个月内该行业上涨幅度超过大盘在 5%—15%；
中性：预期未来 3—6 个月内该行业变动幅度相对大盘在 -5%—5%；
减持：预期未来 3—6 个月内该行业下跌幅度超过大盘在 5%以上。

特别声明：

国金证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证，对由于该等问题产生的一切责任，国金证券不作出任何担保。且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，在不作事先通知的情况下，可能会随时调整。

客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。本报告亦非作为或被视为出售或购买证券或其他投资标的邀请。

证券研究报告是用于服务机构投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议客户应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保。

在法律允许的情况下，国金证券的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供多种金融服务。

本报告反映编写分析员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，且收件人亦不会因为收到本报告而成为国金证券的客户。

本报告仅供国金证券股份有限公司的机构客户使用；非国金证券客户擅自使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

此报告仅限于中国大陆使用。

上海

电话：021-60753903

传真：021-61038200

邮箱：researchsh@gjzq.com.cn

邮编：201204

地址：上海浦东新区芳甸路 1088 号紫竹国际大厦 7 楼

北京

电话：010-66216979

传真：010-66216793

邮箱：researchbj@gjzq.com.cn

邮编：100053

地址：中国北京西城区长椿街 3 号 4 层

深圳

电话：0755-83831378

传真：0755-83830558

邮箱：researchsz@gjzq.com.cn

邮编：518000

地址：中国深圳福田区深南大道 4001 号时代金融中心 7BD