计算机应用行业

从软件算法生态看 GPU 的发展与局限

行业评级	买入
前次评级	买入
报告日期	2016-05-31

相对市场表现

分析师: 刘雪峰 S0260514030002
02160750605
gfliuxuefeng@gf.com.cn

相关研究:

计算机应用行业:“度秘”机器人入驻肯德基概念店点评	2016-04-26
计算机应用行业:企业级 SaaS 方兴未艾，三细分方向或有大机遇	2016-03-02

联系人: 张璋 021-60759787
zhangzhang@gf.com.cn

核心观点:

- **GPU 作为一种协处理器，传统用途主要是处理图像类并行计算任务;**

计算机系统面对的计算任务有着复杂而不同的性能要求，当 CPU 无法满足特定处理任务时，则需要一个针对性的协处理器辅助计算。GPU 就是针对图像计算高并行度，高吞吐量，容忍高延迟而定制的并行处理器。

- **人工智能加速硬件技术路线尚未确定，GPU 加速受多重挑战;**

在人工智能技术发展早期，GPU 作为一种现成的并行计算加速芯片被使用在多个项目之中，如汽车的自动驾驶，图像识别算法等。

但 GPU 未必为人工智能加速硬件的终极答案。早在 2014 年就有研究表明使用 FPGA（现场可编程逻辑门阵列）与 GPU 在加速图像识别类任务能效比为 7：2；2016 年 5 月末谷歌披露其 TPU（Tensor Processing Unit）专用人工智能加速芯片性能相比之前解决方案高出一个数量级。TPU 已秘密使用在谷歌诸多商业项目中超过一年，并参加了与李世石的世纪人机围棋大战。谷歌专用人工智能芯片实用化超出市场的认知和预期;

芯片上大规模并行计算优化可分为两个主要问题：计算单元优化和片上网络优化。GPU 限于最初设计目标，在两个方向上均不能完美匹配人工智能主流算法。未来随着人工智能技术大规模商用化，从产业链过去发展的历史类比，专用人工智能加速协处理器将对 GPU 这类过渡方案构成挑战。

- **GPU 是 VR 显示性能的保障，VR 是未来 GPU 市场的支撑之一;**

VR 对图形计算性能要求超过现有大众级显卡水平，未来 VR 设备市场将成为 GPU 市场增长的支撑之一。重度 VR 设备主要替换现有的游戏主机和部分客厅高清电视市场，2015 年两者合计最大可替换规模 6000 万台。考虑到 GPU 在游戏主机领域是替换升级而非新增市场，且游戏主机消费人群与高清电视消费人群有一定重合，我们预期未来高性能 GPU 市场空间在游戏主机与游戏主机+高清电视市场之间的某一个数字。

- **GPU 在云计算/大数据等领域也有较好的应用前景;**

GPU 的并行计算能力适合在除图形计算以外的多种特定计算场景中，云计算提供商如亚马逊将 GPU 嵌入云计算服务 EC2 中提供给用户。大数据基础开源软件架构 Hadoop 的部分组件适合使用 GPU 加速；Nvidia 积极推广云+GPU+游戏模式拓展新市场。NV 产品线中供给数据中心的业务虽然体量依然与游戏用 GPU 相比太小，但是还是可以有积极期待。

- **风险提示**

GPU/人工智能/并行计算等前沿技术领域，易被新技术改变行业趋势;GPU 产业长期被寡头垄断，A 股相关公司极少，差距极大;

目录索引

图表索引

投资要点

1、计算机系统面对的不同并行计算任务有着复杂而差异较大的性能要求：计算精度、计算并行度、并行进程交互复杂度、计算吞吐量、计算实时性、计算延迟要求等。由于CPU为适应通用计算要求而无法满足图形计算苛刻而特殊的要求，GPU这种专门针对图形并行计算的协处理器被加入计算机体系之中。

2、在人工智能技术发展早期，GPU作为一种现成的并行计算加速芯片被使用在多个项目之中：谷歌的图像识别项目、AlphaGo项目、特斯拉/沃尔沃等诸多汽车厂的辅助驾驶系统和无人驾驶实验中，均使用了GPU作为加速芯片。

但GPU未必为人工智能加速硬件的终极答案。FPGA同样是业内关注的神经网络加速芯片方案之一。而2016年5月末谷歌披露其TPU（Tensor Processing Unit）专用人工智能协处理器性能超过之前解决方案一个数量级。TPU已秘密使用在谷歌诸多商业项目中超过一年，并参加了与李世石的世纪人机围棋大战。谷歌专用人工智能芯片实用化超出市场的认知和预期；

大规模并行计算优化可分解为两个主要问题：计算单元优化和片上网络优化。GPU限于最初设计目标，在计算单元上提供了过高的精度而在片上网络优化上投入了过少的资源，并不完美匹配神经网络算法。从芯片产业链过去的案例-GPU/DSP的诞生与发展来看一种成熟的需求往往会催生一种新的协处理器。未来随着人工智能技术大规模商用化，专用人工智能加速芯片将对GPU/FPGA等在现有成熟芯片上修改的过渡方案构成挑战。

3、VR对图形计算性能要求超过现有大众级显卡水平，未来理想的重度VR设备市场将支撑高端GPU市场的增长。重度GPU设备将主要替换现有的游戏主机和客厅高清电视市场，考虑到高端GPU在游戏主机领域是替换升级而非新增市场，且游戏主机消费者与高清电视消费者有一定重合，我们预期未来高性能GPU市场空间在游戏主机与游戏主机+高清电视市场之间的某一个数字。

4、GPU的并行计算能力适合在图形计算以外的多种特定计算场景中，云计算提供商如亚马逊将GPU嵌入云服务EC2中提供给用户。大数据基础开源软件架构Hadoop的部分组件适合使用GPU加速；Nvidia积极推广云+GPU+游戏模式拓展新市场。NV产品线中数据中心专用GPU业务最近一个季度增速较快，但体量依然与游戏用GPU相比太小。

风险提示

GPU/人工智能/并行计算等前沿技术领域，易被新技术改变行业趋势；GPU产业长期被寡头垄断，A股相关公司极少，差距极大；

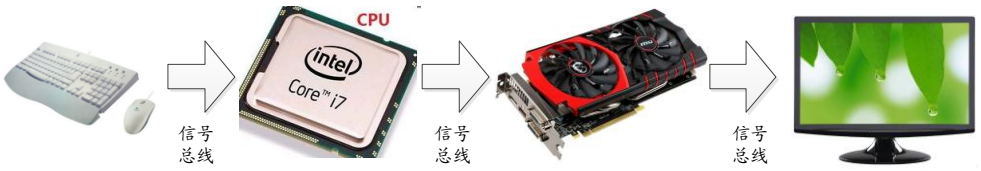
第一章、GPU 简介

GPU（Graphics Processing Unit）的中文名称是图形处理器，原先主要用途是用于协助CPU处理图像计算。其原始设计针对图像计算的特性进行优化，因此也能兼职一些与图像计算特性接近的大规模并行标准浮点数计算任务，如科学计算与数值模拟。但大规模并行计算并非一个笼统的概念，而是一个可以按照计算性能需求在6个维度上进行细分的大类别。因此GPU绝非解决大规模并行计算问题的万金油，无法很好的支持与图形计算特性相差较大的并行计算任务。

1.1、GPU 是什么？

GPU其他名称有显示核心、视觉处理器、显示芯片。顾名思义，GPU最主要的应用场景就是处理图像显示计算。计算机图像显示流程见**图1**，在这个过程中CPU决定了显示内容，而GPU则决定了显示的质量如何。像GPU这类辅助CPU完成特定功能芯片统称“协处理器”，“协”字表明了GPU在计算机体系中处于从属地位。

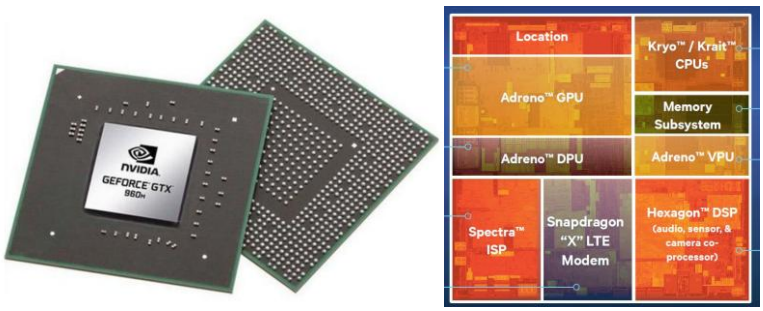
图1：计算机显示的基本过程



数据来源：广发证券发展研究中心

GPU芯片可根据与CPU的关系分为独立GPU和集成GPU。独立GPU通常图形处理能力更高一些，但也有成本更高，功耗和发热较大等问题。近年集成式GPU流行于移动计算平台如笔记本和智能手机。例如高通的智能手机芯片通常将CPU和一个功能较弱的GPU以及其他协处理器通过SoC（System on Chip，片上系统）技术组合在一起。集成GPU图形计算性能相对独立GPU较弱但功耗/成本均针对了移动计算平台的需求做了优化，将长期占据移动计算市场。

图2：独立GPU和SoC集成GPU



Nvidia 的独立GPU芯片

高通骁龙820芯片一部分面积安放了Adreno集成GPU

数据来源：Nvidia官网、高通官网、广发证券发展研究中心

1.2、为什么需要 GPU 等协处理器？

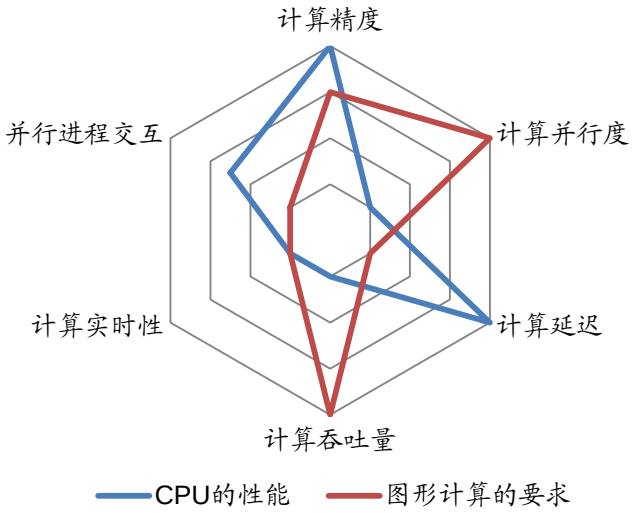
在计算机系统中，之所以出现GPU等协处理器，归根到底在于没有一种芯片设计方案能够满足所有不同类别计算任务所需求的全部性能指标：

- ✓ 计算精度；
- ✓ 计算并行度；
- ✓ 计算延迟；
- ✓ 计算吞吐量；
- ✓ 并行进程之间的交互复杂度；
- ✓ 计算实时性要求；

鱼和熊掌不可兼得；在设计计算机芯片中，以上六个指标不可能在有限的资源约束下同时满足。**图3**的雷达图比较了CPU的设计偏向（蓝线）以及图形计算的要求（红线），越靠近外圈则表示要求高/性能好，如计算延迟低、计算吞吐量大。

我们可以发现CPU设计的一部分偏好，如并行进程交互能力强，低计算延迟是图形计算所不需要的；但图形计算要求的高计算并行度，高计算吞吐量是CPU所不能提供的。将CPU应用在图形处理中会造成一部分性能被浪费，而另一些性能CPU无法满足要求（雷达图上红线和蓝线的显著差异）；这提供了GPU这种针对图形技术优化芯片性能指标的协处理器的生存空间。

图3： CPU的设计性能偏向以及图形计算的要求



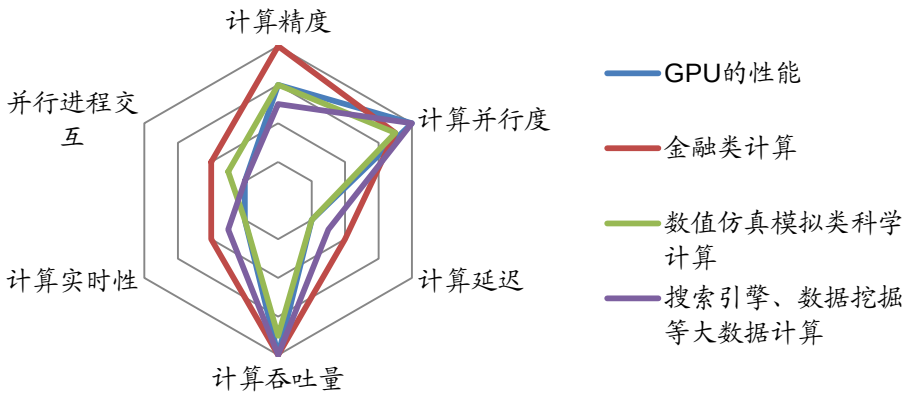
数据来源：广发证券发展研究中心

在广义计算系统体系中，其他类别的协处理器，如DSP（Digital Signal Processor信号处理器），FPGA（Field Programmable Gate Array，现场可编辑门阵列），BP（Base-band Processor,基带处理器）等协处理器之所以独立存在，均因为其所处理的特定计算任务在计算指标雷达图中与CPU以及其他协处理器差异过大。***一个协处理器产业是否有足够的市场空间主要取决于其针对的计算任务在性能雷达图中是否独特（否则会被CPU等“兼职”），以及这种计算任务是否有足够大市场需求。***

1.3、GPU 还能干什么？

GPU生产厂商针对图形处理的性能要求将资源分配强化两个特定指标：**计算并行度和计算吞吐量**。除了图形计算以外，还有一些计算任务的性能雷达图落在GPU的性能范围内或相差不甚太远（见**图4**），比如数值仿真模拟、金融类计算、搜索引擎、数据挖掘等。

图4：一些较匹配GPU特性的计算任务要求雷达图



数据来源：广发证券发展研究中心

正因看中拓展GPU在特殊计算任务的应用前景，主流的GPU厂商纷纷推出软硬件结合的并行编程解决方案。例如Nvidia推出闭源的CUDA并行计算平台，而AMD推出了基于开放性OpenCL标准的Stream技术。这类技术在软件上提供一个定制的编译器，将计算任务尽可能分解成可独立并行执行的小组件（术语为“线程”）；在硬件上对GPU进行小幅度修改，少量提高其在延迟/并行交互等传统弱项的性能。

虽然GPU的并行计算能力与金融数据处理需求存在一定匹配(图4中红线和蓝线相近)，但金融核心账本计算中需要远超过一般计算平台的精度。GPU内部搭载的***2进制***计算单元无法保障账本分毫不差；金融业的核心账本计算业务长期依赖搭载***10进制***计算单元的IBM Power系列高端处理器。如果改造GPU使其搭载***10进制***硬件计算单元，则其又无法适应图形计算的需求。这个案例充分说明：并非***所有并行计算任务就一定适合GPU计算，而需要根据实际情况区分***。

1.4、GPU 不适合干什么？

GPU属于大规模并行计算芯片的一个子类；但其并不能解决所有的大规模并行计算任务。大规模并行计算芯片可粗略划分为两大组成部分：

1）并行计算单元，数目从数个至数千个不等，完成“线程”计算；

2）NoC(Network on Chip，片上通讯网络)，负责在计算单元之间传递数据；

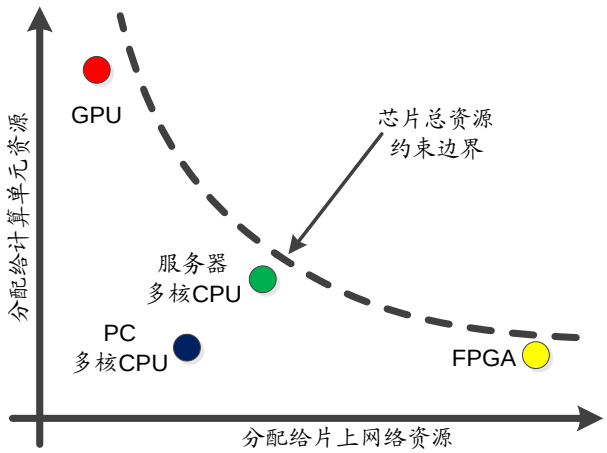
针对不同的计算需求场景，大规模并行计算芯片的设计思路大体有两个方向：

1）处理单元优化：包括增减处理器单元数量或改变处理器单元内部的结构等；

2）NoC网络优化：更改网络拓扑、网络路由算法、优化网络控制机制等；

这两个方向上的优化需要分享芯片上有限的资源；强化一个方向的性能/增加某个方向的资源分配往往就意味着需要牺牲另一个方向的性能。多核CPU、GPU、FPGA是常见的并行计算架构，它们的资源分配倾向示意图见**图5**：

图5：不同并行计算芯片资源分配倾向

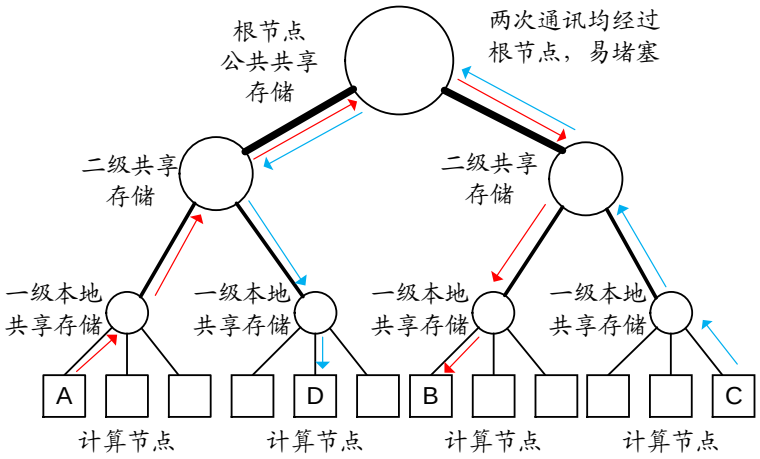


数据来源：广发证券发展研究中心

GPU将主要资源分配给了图形常用计算单元，如浮点数的乘法和加法，而采用了最简单的片上网络拓扑：树状NoC网络，在基本计算单元之间传递数据，见**图6**；这种片上网络的优缺点分别是：

- ✓ 优点1：消耗的资源最小；
- ✓ 缺点1：通过读写片上存储的方式传递数据，速度较慢；
- ✓ 缺点2：树根结点容易因通讯堵塞成为瓶颈，如**图6**中红线和蓝线分别表示A计算节点向B，C向D传递数据，两个传递过程在根节点和二级共享节点交汇，当片上数据传递频繁时，树状拓扑NoC极易发生堵塞问题。

图6：GPU常用的树形片上网络拓扑传递计算单元数据



数据来源：广发证券发展研究中心

GPU之所以采用树状拓扑结构，概因其“主业”-图形计算仅有少量情形需要在计算节点之间做复杂数据通信，因此采用树状拓扑以外的方案是纯粹的浪费。**但树状拓扑结构限制了相当多类别的大规模并行计算任务在GPU上发挥，换句话说，下列这些并行计算任务并不是GPU扩展的强项：**

- ✓ 带有较多分支判断类的并行计算任务，典型任务如人机交互、电脑和环境交互中的逻辑判断计算等；
- ✓ 并行计算中带有较多串行成分，以及反馈算法的并行计算任务，典型例子如控制系统计算任务；
- ✓ 带有网状结构数据流的并行计算。典型案例为FFT（傅里叶分析）计算任务，CUDA中的FFT优化后可以提供相对CPU约10倍的提速，但当FFT长度超过某个门限后GPU的提升性能就发生下滑（资料来源：NV官网）。DSP芯片往往针对FFT的算法特性提供定制优化，没有GPU存在的问题，因此手机SoC中往往由DSP而不是GPU处理FFT这种网状大规模并行计算。

1.5、GPU 总体市场现状

PC/游戏机用独立中高端GPU这个细分市场是GPU整体市场中单价最高，利润率也较高的部分。根据JPD Research发布的2015年/2016年第一季度独立显卡报告：

- ✓ 独显方面，NVIDIA在2015年的表现十分强势。高达81%，AMD只有19%。相比上年的71.5%以及28.4%的差距，2015年差距明显增大了。
- ✓ Nvidia在2015年纯GPU销售额为41.87亿美元，AMD为9.82亿美元，整个独立显卡市场的同比下滑3.6%，系PC下滑拖累。
- ✓ PC市场的独立显卡占有率2016年1季度达到了32.8%，同比上升了1.51%；
- ✓ 桌面系统的插入板式独立GPU环比增长了4.9%；
- ✓ 游戏用PC市场确实在本季是一个亮点，但这是相比PC整体持续下滑而言，游戏用PC对整体PC市场提振有限。
- ✓ Nvidia日前发布了2016年的一季报，其在数据中心和汽车的业务虽然体量较小，和PC游戏等相差数个数量级上，却增速非常高，达到了63%。

对于GPU整体市场体量的估算，主要难点在于SOC类市场的体量，难点有二：

- ✓ IP授权单价往往为商业机密，因此外界难以获得授权金额；
- ✓ 授权后还有生产成本，这一成本与芯片面积和工艺有关，外界更难以拆分；

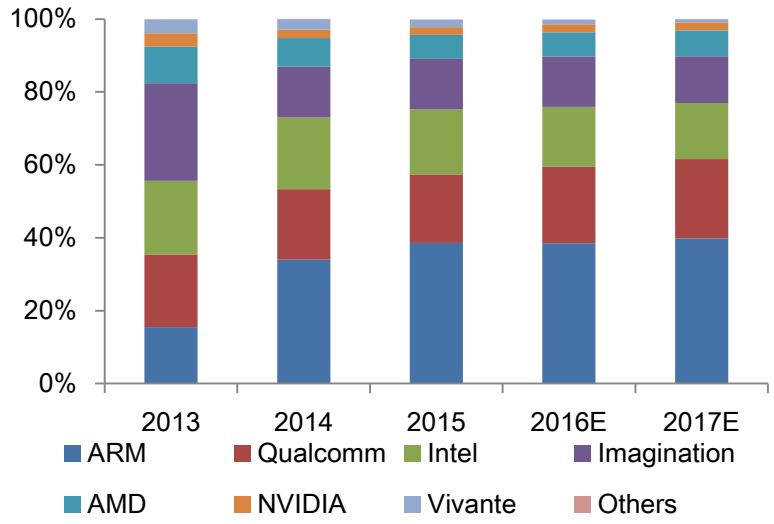
Digitimes Research公布了2013-2015以及预测2016、2017年GPU市场按照出货颗数或SoC中集成的GPU IP数量加总得到的市场占有率变化，比较显著的规律有：

- ✓ **GPU市场未来份额最大的部分将是SoC集成GPU。**ARM/Imagination(大客户为苹果)/Vivante主要提供GPU IP授权，Intel/高通主要在其SoC芯片中集成GPU，而AMD即生产独立GPU芯片也生产集成GPU的SoC。真正专注于独立GPU芯片生产的Nvidia在出货颗数的份额将由2013年的3.7%下滑到

2017年的2.2%。而专注于SoC方案的厂商占比将在2017年突破90%;

- ✓ GPU这个行业和一般IC行业一样呈现了集中度不断提升的特点；ARM和Qualcomm的占比预计将持续提升，而其余厂商的占比都在不断减小。在图中“其余”厂家的份额从2013年的0.2%将下滑到2017年的0.0%。由于SoC IP授权模式下厂商的边际成本接近零，行业竞争格局更接近软件行业，未来呈现7：2：1格局模式甚至单寡头垄断模式也并无不可能;
- ✓ **GPU更主要的未来市场空间将是移动设备市场**，在**图7**中ARM、高通、Imagination、Vivante的主要客户均为智能手机和平板电脑等移动设备，而Intel、AMD、Nvidia的产品主要应用在PC上。移动设备厂商的市场占比持续挤压PC厂商，显示了未来主流GPU的应用还是在移动设备上。

图7：GPU按照出货颗数和SoC IP加总的市场占有率变化



数据来源：Digitimes Research、广发证券发展研究中心

第二章、GPU未来面临挑战应用场景解析：人工智能

最近几年，人工智能技术的实用化取得了显著性突破：

- ✓ 15年6月，Google的无人驾驶汽车在道路实测中突破了100万英里；
- ✓ 像苹果Siri这样的基于语音识别技术的人工智能助手已经普惠大众；
- ✓ 2014年6月，中国香港大学的研究团队发布人脸识别算法准确度超过肉眼；
- ✓ 16年3月，Google人工智能围棋程序AlphaGo击败世界冠军李世石九段；
- ✓ IBM公司尝试使用Watson机器人进行癌症辅助诊断实验；

随着人工智能技术的成熟，利用人工智能替代自然人脑力劳动终将成为一个万亿美元的广阔市场，甚至会成为继互联网之后的下一个生产力革命。目前主流的人工智能软件算法是在神经网络（Neural Networks）技术基础上衍生的几个子类，如CNN（卷积神经网络）、RNN（循环神经网络）、DNN（深度神经网络）等，这些算法的共性特征是都属于大规模并行计算任务。

在人工智能技术发展的早期，多种并行计算芯片被应用于加速人工智能计算，如GPU/FPGA/神经网络专用芯片等。其中GPU作为一种相比其他选项较为成熟的产品，在现有的早期项目中广泛使用。谷歌在图像识别项目、特斯拉与沃尔沃在其辅助驾驶和自动驾驶项目中均使用GPU加速人工智能算法。

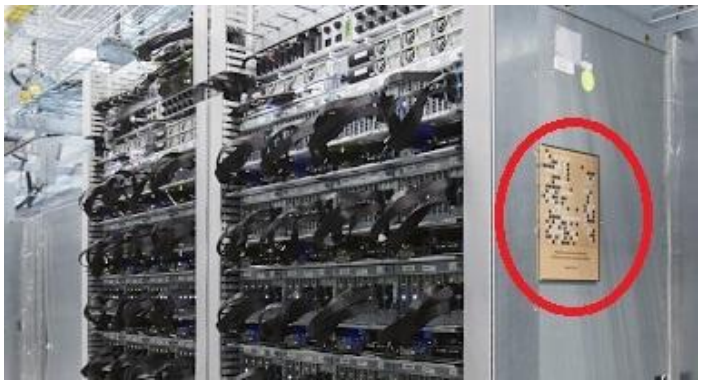
Nvidia日前发布了2016年的一季报，其在数据中心和汽车的业务虽然体量较小，和PC游戏等相差数个数量级上，却增速非常高，达到了63%。Nvidia还针对数据中心云计算推出了Pascal运算平台以及Nvidia自主研发的人工智能算法。***看似GPU已经在人工智能的加速计算中占主导地位；那么，未来人工智能的硬件加速也一定由GPU承担吗？事实并非如此，业内已经存在各种具备竞争力的替代解决方案。***

2.1 谷歌披露实用的全新人工智能专用协处理器：TPU

谷歌在2016年5月末召开的I/O大会披露了TPU（Tensor Processing Unit）专用处理器项目。这种处理器针对谷歌的开源人工智能软件编程框架Tensor Flow进行了优化。资料显示TPU***实际已使用在谷歌诸多商业与科研项目之中超过了一年时间***。今年3月击败李世石的围棋世纪人机大战所使用的服务器集群使用TPU加速围棋中DCNN(Deep Convolutional Neural Network)的计算（**图8**）。谷歌的RankBrain中使用TPU提升搜索结果和街景服务的相关度。

谷歌在人工智能专用处理器的实用化领域再次走在了行业前列。目前业内较为知名的人工智能项目包括中科院的寒武纪芯片项目以及IBM的真北（TruthNorth）项目，而这两个项目距离商业化实用尚有距离。谷歌TPU在***实用一年之后才揭秘面世***，超出业界和市场的普遍认知和预期。

图8：击败李世石的人工智能计算集群中使用了TPU



数据来源：谷歌云平台官方网站、广发证券发展研究中心

谷歌在2016年3月的新闻稿中宣称击败李世石的人工智能计算集群主要使用GPU加速，但在16年5月末突然宣布TPU这种新型芯片，前后略有矛盾让外界困惑，但这证明，***GPU并未在人工智能加速硬件领域相比其他方案占据绝对优势。***

2.2 TPU 主要思路：针对人工智能算法需求裁剪计算精度

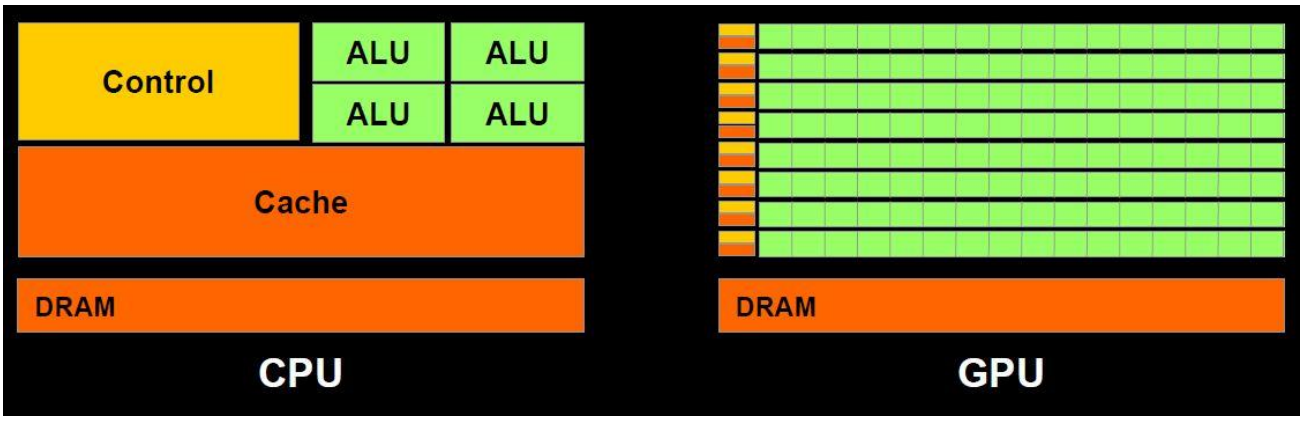
在谷歌官方披露的资料宣称TPU具备强大的计算能力：在机器学习算法上，TPU比传统的加速方案（谷歌之前使用GPU加速方案）在能耗效率上提升一个数量级，相比传统解决方案领先7年（摩尔定律三代节点）。谷歌披露有关TPU如何做到这一点的信息极少，在其官网上仅有一句话有一定信息含量：

“TPU is tailored to machine learning applications, allowing the chip to be more tolerant of ***reduced computational precision***, which means it requires fewer transistors per operation.”

我们判断这里的“reduced computational precision”可能指裁剪数据通道的字宽。例如在GPU中，通常支持IEEE754-2008标准浮点数操作，这一浮点数字宽为32位，其中尾数字宽为23+1（使用隐藏尾数技术）位。如果数据通道中使用8位字宽的低精度尾数，则GPU中各个计算部件所需的晶体管和功耗均会大大减少。

例如，在GPU计算核心中，面积最大，功耗最高的计算部件是ALU（**图9**），ALU中最重要的部件是浮点MA（乘加混合）单元，现有技术下这一单元的延迟与尾数的字宽$\log_2 N$成大致正比，而面积/功耗/晶体管数量大体上与$N^2\log_2 N$成正比。如果字宽由24比特减少到8比特，那么MA的面积可降至约1/14左右，约一个数量级。由**图9**可知ALU占据了GPU芯片面积的很大比例，因此单单优化ALU即可获得足够提高。

图9：CPU与GPU的架构差别概要（示意图，图上比例与实际芯片有差异）



数据来源：Nvida官网、广发证券发展研究中心

除了降低字宽所带来的关键组件优化，GPU原有组件中针对图像处理的组件如光栅、材质贴图单元，均可以根据人工智能的计算需求选择优化或裁剪。对普通GPU进行深度定制处理，削减在神经网络算法不需要的数据位宽和功能即可达到谷歌所宣称的“能耗效率上提升一个数量级”，因此业内有专家认为谷歌采用了此种思路。

但谷歌能实现这一优化的前提，本质上是人工智能软件架构的需求能够容忍低精度的计算。***这表明硬件最终是为算法而服务，算法的需求决定的硬件的设计思路。***

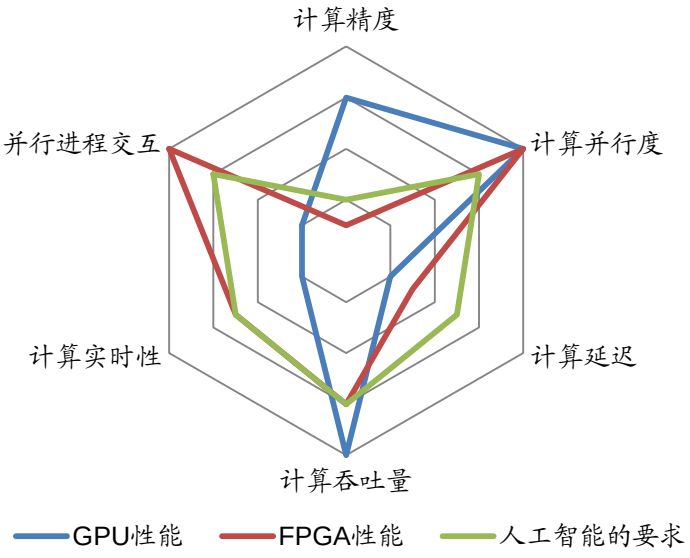
2.3 从谷歌 TPU 设计思路看人工智能硬件发展趋势

目前的GPU加速方案以及FPGA加速方案在人工智能计算领域都存明显缺点：

- ✓ 在计算单元上，GPU的内置计算单元主要针对图像处理设计，计算精度过高存在浪费；FPGA的LUT功能过于弱小，没有针对低精度浮点计算优化；
- ✓ 在NOC架构上，FPGA和GPU原始设计匹配的目标均与神经网络计算存在很大差异性，因此用于人工智能计算加速都存在一定缺憾。

以上表现在计算需求雷达图上即为图10：GPU（蓝线）和FPGA（红线）均不能较好的覆盖住人工智能的需求（绿线）。除了进程交互问题外，实时性和计算延迟同样是人工智能加速的一个重要问题。在人工智能的一些应用场景，如无人驾驶汽车中，汽车的运行速度可能高达40m/s，在计算中额外0.1s的延迟意味着汽车多行驶4米，这就是生与死的差距。GPU的延迟和实时性较差从长期来看会影响其应用在类似无人驾驶这样在实时性和延迟要求较高的场景中。

图10：GPU/FPGA与神经网络计算需求的差异



数据来源：广发证券发展研究中心

虽然谷歌披露有关TPU的实现细节较少，但从已知的细节以及业内共识来看，谷歌实用化人工智能处理器的发展策略是在已有的解决方案架构上针对人工智能处理的需求进行深度定制化开发：谷歌披露了TPU对计算单元做了降精度设计处理，而尚未披露TPU内NOC连接的细节。

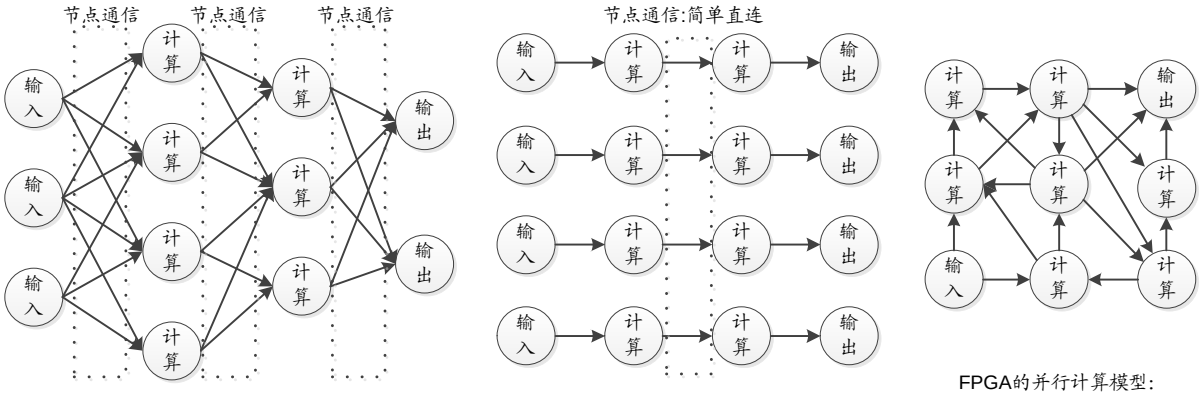
2.4 GPU/FPGA 用于神经网络计算的弱点：片上网络

在人工智能硬件领域，FPGA加速同样是一条有竞争力的技术路径。早在2014年中国搜索引擎巨头百度就尝试与Altera合作探索使用FPGA加速神经网络运算用于搜索结果的优化中，微软也在bing搜索服务中做了相似的探索。Auviz Systems公司在2015年发布了一份研究数据，在神经网络计算中，高端FPGA可处理14个或更多图像/秒/瓦特，而同期一个高端的GPU仅能处理4个图像/秒/瓦特。

但目前学术界已有共识，不管是FPGA还是GPU，由于其最初设计匹配的计算模型与神经网络计算模型存在不同，其并行计算核心之间的通信架构-NOC（Network on Chip，片上网络）应用在神经网络运算中均存在缺点。

图11左展示了神经网络的简要计算模型，可以大致划分为计算节点层和数据通信层，数据通信层把中间结果以较为复杂的方式传送在不同计算层之间。而**图11**中间展示了GPU的SIMT（single instructor multiple threads）并行计算模型，各个通信节点层之间的连接传输为简单直连方式，不需要复杂的通信。**图11**右展示了FPGA的并行计算模型，FPGA允许计算节点之间以任意/可动态编程方式连接并传输数据。

图11：比较神经网络/GPU/FPGA的并行计算模型差别

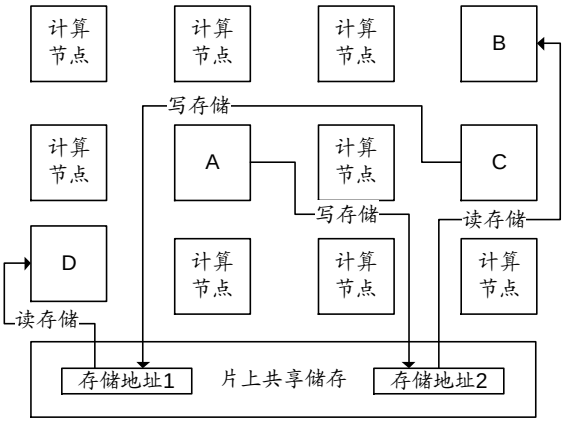


数据来源：广发证券发展研究中心

由于FPGA/GPU针对的并行计算模型不同，其片上网络的实现方式也就不同：

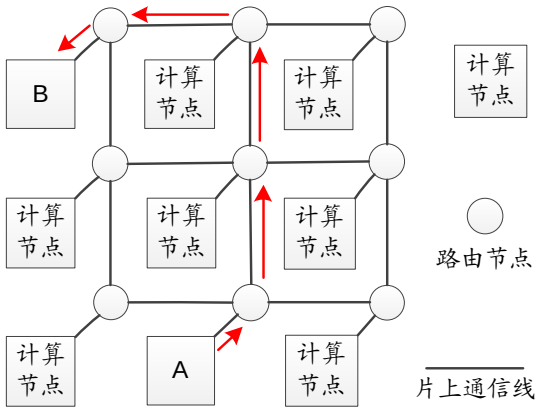
- ✓ **GPU**最初针对图像处理SIMT类任务优化，各个处理核心之间的通信较少且形式简单，因此计算节点主要通过片上共享存储通信，原理如**图12**: A/C计算节点分别向片上共享存储的不同地址写入数据，然后B/D通过读数据的方式完成A->B/C->D的通信。这种片上网络每次通信涉及读写片上共享存储各一次，不仅速度慢，当通信量更多（原本不会发生在图形处理任务中）的时候存储的读写端口还会因堵塞成为系统性能的关键瓶颈。
- ✓ **FPGA**包含大量细粒度，可编程，但功能较弱的LUT（Look up table查找表）计算节点，各个LUT之间通过网格状NOC连接，网格的节点具备Routing（路由）功能。FPGA可以提供计算单元间直接通讯功能（**图13**）：A节点可通过路由网络沿着红色箭头将数据传输至芯片上任意计算节点B，且传输路径动态可编程。因此网格NOC相比共享内存方案能提供大的多的片上通讯容量，相比之下也不易出现瓶颈节点堵塞问题。Auviz Systems能够得出FPGA在神经网络处理中优于高端GPU的方案的结论，很大程度依靠FPGA的片上通信能力而不是羸弱的LUT计算能力。

图12：GPU计算节点片上通信原理



数据来源：广发证券发展研究中心

图13：FPGA计算节点片上通信原理



数据来源：广发证券发展研究中心

神经网络计算的典型的软件架构见**图11**左；这一编程模型的特性为：

- ✓ 节点和节点之间数据传递较为频繁，且互联关系复杂，为典型网状数据流；
- ✓ 特定的神经网络模型下，节点和节点之间的传递关系较为固定。

因此**不管是GPU的共享片上存储NOC，还是FPGA的网格NOC，均非应用在神经网络计算中的最优解**：

- ✓ GPU的片上通信方案中共享存储将成为系统性能的瓶颈：不仅读写存储引入额外延迟和功耗，存储端口一旦在神经网络计算任务较高的片上通信量下堵塞还会造成系统计算整体停顿。
- ✓ 对神经网络计算模型映射优化可使得大部分通信发生在FPGA的临近节点之间且无需经常改变，因此网状神经网络提供的全联通以及动态可编程能力并非必要，一部分用于通信的芯片面积和能量被浪费。

神经网络作为一种并行计算程序，适配的计算节点通讯硬件是提升性能的关键要素之一。目前FPGA和GPU的片上网络架构均不完全匹配神经网络的实际需求，相比之下GPU的共享内存连接的匹配度更差一些。学术界对于定制特殊的NOC去匹配神经网络加速需求已有一定研究，但之前因神经网络算法本身没有商用化，因此定制NOC硬件这一思路也停留在实验室内。随着人工智能实用化和产业化的发展，这些技术将对现有的GPU/FPGA方案形成威胁和替代。

未来人工智能加速硬件一个很自然的发展趋势就是从计算单元和NOC同时入手，研发使用类似TPU这样的专用神经网络协处理器。专用神经网络芯片存在的障碍主要在于：特殊设计芯片需要足够的市场空间去支持开发费用的摊薄（详细论述见我们团队之间发布报告：《中科创达－论下游智能化产业发展及公司角色地位》第一章）。***如果人工智能与神经网络的市场空间是确定性的，从GPU/DSP等专用计算芯片的产业演变历史类比来看，人工智能专用芯片独立发展也将是大概率事件：***

- ✓ 最早电脑图形显示任务由CPU兼任，后来演变出了CPU内置的图形加速器。当图像处理，尤其3D图像类特殊任务需求确立后，GPU自然发展成独立的产品门类并与CPU分离；
- ✓ DSP(Digital Signal Procesor)芯片主要承担特殊数字信号的高速实时特种计算任务，现在广泛用于通信/消费电子/雷达等领域。最早的数字信号处理也由CPU完成，但随后军用数字信号处理的市场空间支撑了DSP专用芯片产业的早期发展，而数字通信时代则强化了DSP独立于CPU的地位。
- ✓ 现有的人工智能技术商业化应用尚未普及，因此设计人工智能专用加速芯片成本过高而无法摊薄，并无商业价值，而FPGA/GPU这样的并行计算架构芯片起到了很好的过渡作用。但当人工智能加速硬件市场真正有了商业化规模后，专用芯片的成本能够被摊薄，从DSP/GPU发展历史来类比，***专用芯片能够比过渡方案更好的适应需求，成为一个独立的产业。***

第三章、GPU 未来较适应场景解析

GPU虽然不能处理所有大规模并行计算问题，但在其适应的特定计算领域，特别是图形优化处理上依然具备绝对性能优势。GPU未来较为适合拓展应用场景应为VR/AR（虚拟现实/增强现实）、云计算+游戏结合、以及云计算服务器中为特定的大数据分析提供加速。在这些领域的增长点有可能是独立GPU突破现有增长迟缓障碍的新增长领域。

3.1 VR 应用：持续增长的优势领域;

在VR（Virtual Reality，虚拟现实）设备性能指标中，图像显示性能是其核心竞争力。在VR中降低从用户头部动作到画面改变的延迟至20毫秒以下是防止用户眩晕的必要条件；而达到这点除了需要软件和OS优化以外，足够的硬件图像计算能力是基础。**表1**举例了VR图形显示的要求以及大众级显卡能够提供的图形显示水平：

表 1：VR 理想的 GPU 性能以及现有解决方案性能指标

性能指标	理想 VR 显示需要指标	大众级显卡现有水平
头部动作到	总延迟 20ms 以下	GPU 计算延迟
图像改变延迟	分配给 GPU10ms 以下	可高达 30ms 以上
视角 FOV	110 度以上	90 度
视频解析度	1512X1680X2	1920X1080
刷新率	90fps	30fps
计算像素要求	4.5 亿像素/秒	6000 万像素/秒

数据来源：Nvidia 官网、广发证券发展研究中心

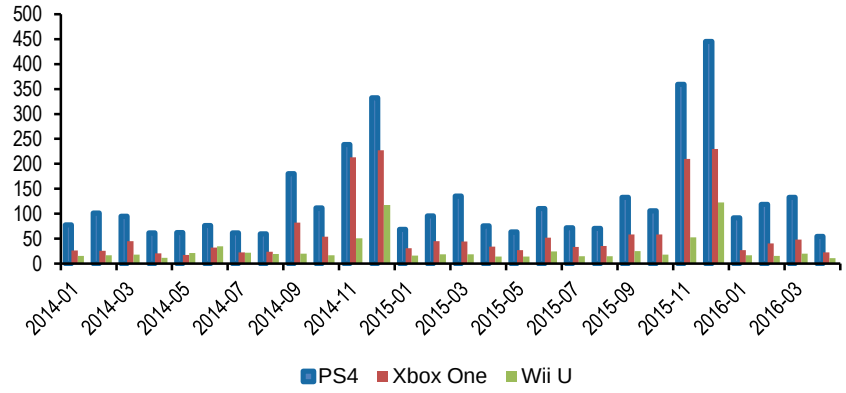
正因目前大众显卡无法提供VR所需的图形处理计算能力，现有的两大主流头显Oculus Rift和HTC VIVE均要求配套的PC配置顶级显卡，如Nvidia GTX970或AMD R9 290级别的显卡，动辄需要￥2000以上的显卡成本。从长期来看，VR/AR设备将拉动中高端GPU市场的持续增长。

目前在VR/AR领域最大的热点应用是游戏和影视。国外第三方机构在其报告中估算了2025年VR/AR的应用场景的比例,其中游戏与影视类占比超过50%。

游戏和视频娱乐是典型的客厅场景应用，我们可以预期未来的客厅游戏机（如PS/Xbox/Wii）和高清电视机都将受到VR设备替代的冲击，**因此供应给VR的专用GPU市场空间大致可对标客厅游戏机和高清电视机的消费台数。**

图14展示了Vgchartz统计的2014年至今三大主流游戏主机的全球月销量。2014、2015年全球三大主流游戏主机的销量分别为2615、2952万台，15年同比增长率12.8%。**图14**展现的月销量显示游戏主机的销售旺季在每年11、12月，恰好为西方感恩节、圣诞节两个传统节日月份。这表明游戏主机的消费主力是经济并不独立的儿童和青少年，其获得游戏主机的主要方式为节日礼品，而游戏主机也非刚性需求。我们预期游戏主机这一市场空间难以进一步快速增长。

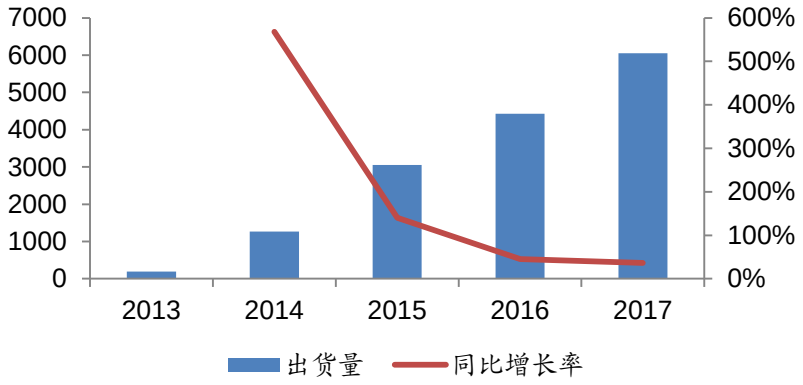
图14：VGchartz统计三大主流游戏主机全球月销量（万台）



数据来源：VGchartz.com、广发证券发展研究中心

而**图15**展示了由NPD DisplaySerach发布全球2013~2017年4Kx2K及以上级别高清电视出货量及同比增长率。NPD DisplaySerach预计到2017年4Kx2K级别高清电视的同比增长率将下滑至40%以下，而全球年出货量为6200万台。

图15：全球4Kx2K及以上级别高清电视出货量及同比增长率



数据来源：NPD DisplaySearch、广发证券发展研究中心

考虑到高清电视和游戏主机的客户有相当一部分是重合的（一般购买游戏主机的玩家有很大概率会配置一台高清电视），而且现有游戏主机也均配置一块定制的中高端显卡。因此可以预测到未来VR专用高性能GPU的市场空间大致在游戏主机数量以上，游戏主机+客厅高端电视总量以下的某一个数字。

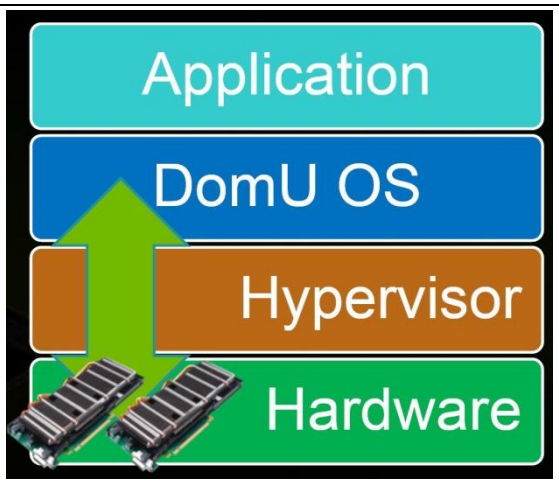
VR以及AR（增强现实）更广阔的应用在于独立一体机上：独立一体机具备移动能力，让VR/AR超脱出了客厅应用这一范畴，与移动互联网结合后成为每个人都需要消费电子产品。但移动一体机对计算芯片的能耗，体积乃至散热都有着严格的要求。目前SoC(System on Chip，片上系统)上集成GPU在移动一体机上的优势是独立GPU显卡暂时无法动摇的。这一领域未来的市场空间较难估算，且市场竞争的结果将更加扑朔迷离。

3.2 云计算/大数据应用

早在2010年，亚马逊风靡全球的计算平台EC2（Amazon Elastic Compute Cloud）中，Nvidia GPU已经被作为一个重要的并行计算组件提供给客户，用作大规模并行浮点数计算。用户每使用一个实例可调用两个Nvidia Tesla m2050 GPU，而当时的价格为每小时2.1$。经过AWS多年的降价之后，目前在EC2上使用一个g2.2xlarge实例的价格为每月475.8$，仅为六年前价格的1/3。

在EC2中调用GPU的原理如**图16**所示：AWS的管理程序Hypervisor被直接跳过，而DomU OS和应用可以直接通过IO与GPU通信，充分发挥GPU在浮点数的并行计算能力。

图16：在AWS EC2中调用GPU的原理



数据来源：Nvidia官网、AWS、广发证券发展研究中心

GPU在云计算中心最有效率的加速工作之一就是大数据分析加速。大数据领域流行的开源软件平台Hadoop与GPU的计算属性较为匹配。Hadoop的核心算法之一为Map-Reduce。在“Map”过程中计算任务被分解成了完全不通信的并行线程，而到“Reduce”阶段线程才开始出现复杂的数据流，因此使用GPU加速一部分计算时完全可行。

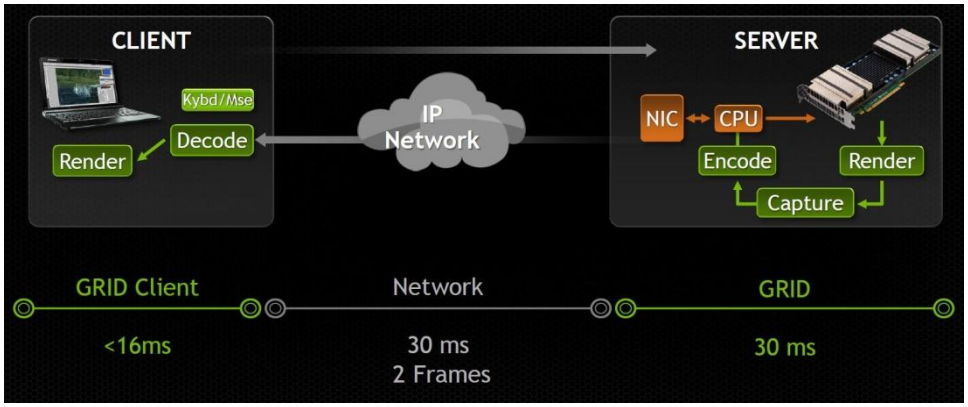
3.3 GPU，云和游戏服务结合

在现如今互联网基础设施已经完善的市场，把GPU和云计算以及游戏结合在一起是游戏产业下一个具有吸引力的发展方向。云计算与游戏结合的好处显而易见：

- ✓ 对于游戏开发者，不需要担心盗版问题；
- ✓ 对于游戏运营商，云服务可以获得更精确的客户资料，开展新式计费；
- ✓ 游戏更新将更快捷，用户无需下载本地更新包；
- ✓ 对于游戏玩家，无需购买昂贵高端游戏主机或PC，初始投资少；
- ✓ 对于游戏玩家，云服务游戏更具备移动性；

展示了云计算+游戏+GPU架构的原理：绝大部分图像计算都在云端完成，云服务器通过互联网将显示图像直接输送至客户端，客户端只需将视频解码显示：

图17：云、GPU、和游戏结合的原理



数据来源：Nvidia官网、AWS、广发证券发展研究中心

目前云计算+GPU+游戏这个模式限于现有网络基础设施限制，依然没有大规模商用，但Nvidia依然对其抱有厚望并积极推动。从这个侧面也可以看出，Nvidia自己也知道GPU未来最主要的应用领域依然是游戏的图像处理上。

最后，GPU还有一块市场是军用GPU市场，这一市场与民用GPU市场有着很大不同。民用GPU追求画面性能的极致，以最好的画面满足消费者，特别是游戏玩家的需求；而军用GPU更多的要求在于高可靠性、高耐用性、抗高空辐射、能在野战环境下安全使用。需求的导向不同导致GPU从工艺到芯片设计理念都截然不同。

对于军品厂商转向民品制造，或从头开始搞GPU芯片的设计与制造，我们持谨慎态度。这些年国外GPU大厂建立起了深厚的技术和生态壁垒，且不谈技术指标上的差距，单是Nvidia私有的CUDA平台成为科学计算极重要标准就让后入玩家望洋兴叹。另一方面，在不掌握核心技术的情况下和垄断寡头合作做其下游是否能赚到钱，我们相信有类似合作经历的国内厂商都会有深刻体会。

风险提示

- GPU与并行计算属于高技术行业，行业趋势易被新技术出现所改变；
- GPU产业长期被寡头垄断，A股相关公司极少，差距极大。

广发计算机行业研究小组

刘雪峰: 首席分析师，东南大学工学士，中国人民大学经济学硕士，1997 年起先后在数家 IT 行业跨国公司从事技术、运营与全球项目管理工作。2010 年 7 月始就职于招商证券研究发展中心负责计算机组行业研究工作，2014 年 1 月加入广发证券发展研究中心。

王文龙: 研究助理，东南大学信息工程学士，香港城市大学金融与精算数学硕士，2015 年进入广发证券发展研究中心。

王奇珏: 研究助理，上海财经大学信息管理学士，上海财经大学资产评估硕士，2015 年进入广发证券发展研究中心。

广发证券—行业投资评级说明

买入: 预期未来 12 个月内，股价表现强于大盘 10%以上。

持有: 预期未来 12 个月内，股价相对大盘的变动幅度介于-10%～+10%。

卖出: 预期未来 12 个月内，股价表现弱于大盘 10%以上。

广发证券—公司投资评级说明

买入: 预期未来 12 个月内，股价表现强于大盘 15%以上。

谨慎增持: 预期未来 12 个月内，股价表现强于大盘 5%-15%。

持有: 预期未来 12 个月内，股价相对大盘的变动幅度介于-5%～+5%。

卖出: 预期未来 12 个月内，股价表现弱于大盘 5%以上。

联系我们

	广州市	深圳市	北京市	上海市
地址	广州市天河区林和西路 9 号耀中广场 A 座 1401	深圳市福田区福华一路 6 号免税商务大厦 17 楼	北京市西城区月坛北街 2 号月坛大厦 18 层	上海市浦东新区富城路 99 号震旦大厦 18 楼
邮政编码	510620	518000	100045	200120
客服邮箱	gfyf@gf.com.cn			
服务热线				

免责声明