从英伟达的DPU谈起

https://mp.weixin.qq.com/s/6BjUGM_fCu8g4cL6hcVh4g

李飞 Fri Oct, 09 00:53

近日，Nvidia召开了秋季GTC并发布了一系列新的芯片和硬件。从GTC发布的产品来看，我们认为Nvidia最近一系列的动作都是在加强数据中心服务器市场的布局。虽然Nvidia占据了人工智能加数据中心的风口，但是我们仍然认为Nvidia在打入数据中心的道路上有不少挑战需要克服。

Nvidia在数据中心的新产品：DPU

在GTC上，Nvidia着重介绍的一个产品是它在数据中心网络领域的新产品：Data Processing Unit（数据处理器，DPU）。DPU把ARM处理器核、VLIW矢量计算引擎和智能网卡的功能集成在了一起，主要应用在分布式存储、网络计算和网络安全领域。

在数据中心的分布式计算中，存储和网络都需要大量的处理器资源。对于存储，在数据中心中大量存储是分布式的，每个服务器可以访问的存储空间远远不仅限于本地的硬盘，这一方面大大增加了分布式存储的灵活性，但是另一方面这类分布式数据存储系统的访问会需要额外的软件处理来完成，而传统上这类分布式存储的软件处理会使用CPU来完成。

除了分布式存储之外，数据中心服务器之间的网络互联也有一套底层软件系统，这套软件需要能完成网络互联协议，除此之外还需要能跑一套数据中心必备的网络安全系统。传统上，这些处理也会跑在CPU上，而随着智能网卡（SmartNIC）的逐渐普及，智能网卡正在网络安全和网络协议处理方面缓慢地取代CPU。

Nvidia的DPU事实上可以看做是SmartNIC的增强版本，一方面加强了SmartNIC对于网络协议和网络安全的处理能力，另一方面整合并加强了分布式存储的处理能力，从而让DPU能在这两个领域更多地替代CPU。

事实上，Nvidia这次着重介绍的DPU是由它于近年收购的Mellanox内部孵化的。该产品最初的定位就是增强型智能网卡，其第一代产品BlueShield 1已经于去年发布了，只是由于是试水产品因此比较低调。Nvidia内部应该是看到了数据中心中网络和存储相关处理的机会，因此给第二代BlueShield项目倾注了更多资源，同时也提出了DPU的概念。Nvidia这样的操作有点像20多年前由它提出的GPU——在Nvidia提出GPU之前，大多数公司对于同类产品的概念一直是显示加速卡，而Nvidia认为显示卡不仅仅是一种附属于CPU的显示加速，而且在计算机系统中能担负起和CPU接近甚至更重要任务，因此提出了GPU的概念。这次Nvidia提出DPU这个新概念，可以看出Nvidia对于网络数据处理领域的重视。而在未来，Nvidia计划继续推进BlueShield DPU产品线，一方面增强其网络互连处理能力，另一方面还计划加强其AI处理能力，从另一个维度增加产品的功能性。

DPU+ARM是Nvidia替代中低端x86的切入点

自从Nvidia开始收购ARM之后，我们看到Nvidia已经在多个场合显示了其利用ARM处理器进一步占领数据中心服务器市场的决心，而DPU则是Nvidia最新的一个布局。
回顾Nvidia在数据中心市场的策略，最初Nvidia的立足点在于其在人工智能领域无与伦比的优势，其GPU配合CUDA生态成功地抓住了数据中心最大的增量市场，即人工智能计算。
而在此之后，Nvidia显然已经不满足于抓住增量市场，更希望能切入数据中心的存量市场，即设法用自己的芯片产品去取代Intel（以及AMD）主导的x86 CPU。目前，唯一能与x86 CPU性能和生态属于同一数量级的唯有ARM，而这也是Nvidia决定收购ARM的重要原因。一旦Nvidia正式完成收购ARM，那么集成了ARM核的DPU将成为Nvidia打入数据中心存量市场取代x86 CPU的第一个切入点。如前所述，传统上会使用低端x86处理器来完成数据中心中网络协议处理、网络安全和存储控制等任务，而使用DPU之后，就不再需要再使用CPU来处理这些任务了。从另一个角度来看，事实上就是Nvidia的DPU在数据中心的网络协议处理、网络安全和存储控制市场取代了低端x86 CPU。
事实上，Nvidia选择使用DPU来打这个市场，而不是推出ARM核直接和低端x86 CPU竞争是一个明智之举。首先，智能网卡和智能存储领域CPU逐渐被取代是大势所趋，Nvidia在这个领域同时也有Mellanox的技术积累，可以说是顺势而为。此外，DPU在网络和存储领域从概念上来说是CPU的下一代产品，所以说虽然DPU中的ARM核性能未必比同一代对标的低端x86 CPU更强，但是整体来说由于在DPU SoC上集成了专用的处理加速模块，因此总体性能一定是超过x86 CPU的。因此，Nvidia虽然第一步想要取代的是低端x86 CPU，但是其推出的对标产品并不低端，这从市场定位上来说也很讨巧。

Nvidia在中高端数据中心处理器市场仍然面临挑战

DPU为Nvidia替代低端x86 CPU打开了道路，但是在中高端数据中心处理器市场，Nvidia如何把握市场仍然是一个未知数，也面临不少挑战。
首先，在中高端市场，目前Nvidia的GPU和x86处理器是标准配置，或者说Nvidia的产品和x86处理器仍然处于互补的关系。如果想要完全吃下高端市场，即使用Nvidia的GPU加Nvidia基于ARM架构的处理器做高端服务器，目前来看还需要在ARM处理器的性能上更进一步。我们知道，ARM基于RISC的架构能占领移动市场主要是依靠能效比，而在服务器市场，尤其是高端服务器市场，能效比的重要程度较弱，主要还是需要看性能。从技术上来说，Intel和AMD在高性能服务器处理器领域一直是领导者，也有数十年的经验积累，而ARM架构做高性能处理器的时间并不久，而直到今天基于ARM架构尚没有公开发售且成功的服务器端高性能处理器。当然，随着未来人工智能重要性提升，或许在一些应用场景GPU比起CPU来说要更重要，Nvidia因此可以推出围绕人工智能的GPU+ARM处理器中高端服务器产品，但是这样的搭配能实现多高的性能尚不清楚，而且很明显如果ARM处理器的性能相比x86越弱，那么这样组合的服务器的使用场景就越狭窄。
除此之外，Nvidia在中高端数据中心处理器可能遇到的另一个挑战是客户自研芯片。亚马逊为AWS已经发布了基于ARM核的自研处理器Graviton。云提供商对于自身的需求最清楚，因此自研芯片非常合乎情理，而且有机会能为自身的云服务提供差异化竞争的能力。如果谷歌等其它云服务商也跟进使用ARM架构自研芯片，那么这些云厂商就会成为Nvidia的客户同时也是竞争

对手。Nvidia如何在客户自研芯片和x86之间找到一个中高端服务器处理器的切入点，我们还需拭目以待。

*免责声明：本文由作者原创。文章内容系作者个人观点，半导体行业观察转载仅为了传达一种不同的观点，不代表半导体行业观察对该观点赞同或支持，如果有任何异议，欢迎联系半导体行业观察。

今天是《半导体行业观察》为您分享的第2457期内容，欢迎关注。

推荐阅读

★欧盟的半导体担忧

★中国半导体制造的海外扩张之路

★COVID-19如何影响存储器行业

半导体行业观察

『**半导体第一垂直媒体**』

实时 专业 原创 深度

识别二维码，回复下方关键词，阅读更多

晶圆 | IP | SiC | 并购 | 射频 | 台积电 | Nvidia | 苹果

回复 **投稿**，看《如何成为“半导体行业观察”的一员 》

回复 **搜索**，还能轻松找到其他你感兴趣的文章！