# 终于，英伟达发布了一个Arm架构的CPU

https://mp.weixin.qq.com/s/D14hcVmSbzpAEopCS0paBg

李寿鹏 Mon Apr, 12 17:52

受惠于过去几年人工智能的爆发，全球GPU领导厂商英伟达不但成为全球市值最高的半导体企业。与此同时，公司又收购了Mellanox和Arm，打造了完整的数据中心芯片产品线，全新的英伟达俨然成为数据中心最炙手可热的明星。

英伟达创始人黄仁勋在昨晚开幕的GTC大会上也表示，在其CPU、DPU和GPU这三条产品线，公司未来在数据中心将会创造更多可能。而在本次大会上，英伟达也带来了CPU和DPU的更新。尤其是收购Arm之后发布的首款CPU，更吸引了广泛的关注。

在笔者看来，这颗芯片的发布，又一次吹响了英伟达向英特尔盘踞的数据中心市场进攻的号角。

**为什么数据中心需要新的xPU?**

熟悉数据中心构造的读者应该知道，传统的数据中心必不可少的一个芯片那就是英特尔X86架构的CPU。然而伴随着AI的流行，这种传统的架构就不再满足了新兴应用的需求，这就给有用并行计算先天优势的英伟达带来了机会，这也是他们过去几年在数据中心如鱼得水的原因。

但按照黄仁勋的说法，随着数据中心训练模型的变大，加上对芯片处理能力需求的提升，这就给数据中心的CPU的数据“搬运”和网络相关处理带来了新的挑战。换而言之，就是当前数据中心使用的X86 CPU很难兼顾数据在CPU和GPU之间流动和网络处理的需求，为此这就给DPU和英伟达自研的Arm架构CPU带来了机会。这也是英伟达收购Mellanox和Arm的原因。

所谓DPU，也就是Data Processing Unit（数据处理器）。在英伟达的产品线布局中，这是一款把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的产品，主要应用在分布式存储、网络计算和网络安全领域。根据相关资料显示，这款产品是他们基于公司之前收购的Mellanox内部孵化而成的。

据介绍，DPU可从CPU上卸载关键的网络、存储和安全任务，使企业能够将其IT基础设施转变为最先进的数据中心。此类数据中心可实现加速、具有完全可编程性，并具有“零信任”安全功能，防止数据泄露和网络攻击。这就减轻了CPU的负载，让其只专注于亟需处理的各种企业应用程序。

来到CPU方面，正如前文所说，数据中心目前几乎都是X86架构的至强处理器的市场。然而黄仁勋指出，正是因为这样的配置，影响了整个数据中心的数据传输。“现在CPU的存储和PCIE带宽，严重影响了GPU能力的释放”，黄仁勋强调，为此他推出了全新的基于Arm架构打造的CPU Grace，希望借助这个新处理器以及自有的NVlink来解决这个数据瓶颈问题。

**英伟达xPU的强势出击**

在上述思路的驱动下，英伟达推出了全新的DPU Bluefield-3 和新款CPU Grace。

首先看DPU方面，据介绍，BlueField-3将具有16个Arm A78内核，能提供十倍于BlueField 2的计算能力，在带宽方便则可以达到400Gbit / s，同时也加上了对PCIe gen 5的支持，获得了比PCIe gen 3快四倍的速度。在这个芯片中，Nvidia还加入了两个加速器，为软件定义存储、网络、安全、流和TLS / IPSEC加密等应用提供支持。此外和BlueField-2一样BLUEFIELD-3能给5G电信和时间同步数据中心的精确定时。数据显示BlueField-3的额定值为350 SPECINT和1.5 TOPS（TeraOps）。

黄仁勋进一步指出，一个Bluefield-3 DPU约等于300个x86 cpu内核，因此它能够大大减轻CPU的负载。BlueField-3同时还可以充当Nvidia的Morpheus云原生网络安全框架的监视或遥测代理。据介绍，公司将于2024年推出性能更强的Bluefield-3 ，进一步加强公司在这个市场的影响力。

以美国海军少将、计算机编程先驱Grace Hopper的名字命名的CPU，则开启了英伟达数据中心的新时代。据黄仁勋介绍，这个处理器能够与常规GPU产品一起工作，让公司能够获得更全面地垂直集成其硬件堆栈的能力。按照NVIDIA的说法，该芯片是专门为大规模神经网络工作负载设计的，预计将于2023年在NVIDIA产品中使用。

按照anandtech介绍，Grace的发布旨在填补NVIDIA AI服务器产品线中CPU的空白。该公司的GPU非常适合某些类的深度学习工作负载，但GPU并不能执行数据中心里的所有操作，这就是CPU存在的意义，为此NVIDIA当前的服务器产品通常依赖于AMD的EPYC处理器和Intel至强这样的处理器，

但正如前文所说，这些处理器对于一般的计算而言是非常快，但不能满足NVIDIA所追求的那种高速I / O和深度学习优化。特别在NVIDIA目前使用PCI Express来进行CPU和GPU连接时，这种连接方式就成为瓶颈。如果引入NVLink，那么系统中的GPU彼此之间就可以直接快速对话，而不需返回主机CPU或系统RAM。

“基于Grace的系统与NVIDIA GPU紧密结合后，性能将比目前基于X86 CPU的，最先进的NVIDIA DG高出十倍”，黄仁勋表示。“绝大多数的数据中心仍将继续使用现有的CPU，而Grace将主要用于计算领域的细分市场”，黄仁勋进一步指出。

从anandtech的介绍我们可以看到，在之前，NVIDIA为了在数据中心使用NVLink，选择了POWER9处理器作为合作伙伴，但Grace的发布，从某种程度上宣布了这种合作的终结。

**写在最后**

其实在这次发布会上，NVIDIA还发布了他们新一代的AI自动驾驶汽车处理器NVDIA DRIVE Atlan，它能够提供约为上代产品Orin四倍的速度。作为一款集成了DPU、下一代架构GPU、新型Arm 内核和最新深度学习和计算机视觉加速器的SoC，NVDIA DRIVE Atlan能够获得高达1000TOPS的速度，以及能提供极高的带宽，为未来的自动驾驶设计赋能。

除了芯片以外，NVIDIA还带来了各种软硬件的更新，例如用于仿真、协作、和自助机器训练的NVIDIA Omniverse，专为工作组打造的便携式AI数据中心NVIDIA DGX Station，面向企业级的NVDIA EGX，用于训练Transformer框架的NVIDIA Megatron，用于计算药物研发加速库Clara Discovery的一些模型，能够为量子电路模拟器提供加速的cuQuantum，数据中心安全平台NVIDIA Morpheus，先进的深度学习对话式AI Jarvis和推理服务器Triton等一系列产品。此外，

英伟达还宣布了和Ampere computing和Amazon等一系列Arm服务器芯片厂商的合作，共同推动Arm生态的发展。

黄仁勋表示，凭借公司现有的芯片、软件、AI和各种产品，英伟达能助力开发者成就毕生的事业。尤其是在芯片方面，黄仁勋强调，公司数据中心路线图包括CPU、GPU和DPU这三类芯片，而Grace和BlueField是其中必不可少的关键组成部分。每个芯片架构历经两年的打磨周期（周期内可能出现转变），一年专注于 x86 平台，另一年专注于 Arm 平台。

“我们每年都会发布激动人心的新品。三类芯片，逐年飞跃，一个架构。”黄仁勋补充说。



**今天是《半导体行业观察》为您分享的第2644内容，欢迎关注。**

推荐阅读

★中国信号链芯片厂商该如何杀出重围

★芯片行业啥都缺，就是不缺投资人

★以色列芯片实力浮出水面

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 | 集成电路 | 设备 | 封测 | 射频 | 存储 | 美国 | 台积电

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！