英伟达CPU问世：ARM架构，对比x86实现十倍性能提升

https://mp.weixin.qq.com/s/ZAIIqt6Du9YK0tsSwpf40g

Synced Mon Apr, 12 17:46

机器之心报道

机器之心编辑部

收购 Arm 还没有定论，但英伟达的 Arm 架构 CPU 已经出现了。英特尔现在可能正感受到不一样的压力。

「只需一张 GeForce 显卡，每个学生都可以拥有一台超级计算机，这正是 Alex Krizhevsky、Ilya 和 Hinton 当年训练 AI 模型 AlexNet 的方式。通过搭载在超级计算机中的 GPU，我们现在能让科学家们在 youxian 的一生之中追逐无尽的科学事业，」英伟达创始人兼首席执行官黄仁勋说道。

4 月 12 日晚，英伟达 GTC 2021 大会在线上开始了。或许是因为长期远程办公不用出门，人们惊讶地看到在自家厨房讲 Keynote 的黄老板居然留了一头摇滚范的长发：

如果你只是对他的黑色皮衣印象深刻，先对比一下 2019、2020 和 2021 的 GTC，老黄气质越来越摇滚。如此气质，黄仁勋今天推出的新产品肯定将会与众不同。

「这是世界第一款为 terabyte 级别计算设计的 CPU，」在 GTC 大会上，黄仁勋祭出了英伟达的首款中央处理器 Grace，其面向超大型 AI 模型的和高性能计算。

英伟达也要做 CPU 了

Grace 使用相对能耗较低的 Arm 核心，但它又可以为训练超大 AI 模型的系统提供 10 倍左右的性能提升。英伟达表示，它是超过一万名工程人员历经几年的研发成果，旨在满足当前世界最先进应用程序的计算需求，其具备的计算性能和吞吐速率是以往任何架构所无法比拟的。

「结合 GPU 和 DPU，Grace 为我们提供了第三种基础计算能力，并具备重新定义数据中心架构，推进 AI 前进的能力，」黄仁勋说道。

Grace 的名字来自于计算机科学家、世界最早一批的程序员，也是最早的女性程序员之一的格蕾丝 · 赫柏（Grace Hopper）。她创造了现代第一个编译器 A-0 系统，以及第一个高级商用计算机程序语言「COBOL」。计算机术语「Debug」（调试）便是她在受到从电脑中驱除蛾子的启发而开始使用的，于是她也被冠以「Debug 之母」的称号。

英伟达的 Grace 芯片利用 Arm 架构的灵活性，是专为加速计算而设计的 CPU 和服务器架构，可用于训练具有超过 1 万亿参数的下一代深度学习预训练模型。在与英伟达的 GPU 结合使用时，整套系统可以提供相比当今基于 x86 CPU 的最新 NVIDIA DGX 快 10 倍的性能。

目前英伟达自家的 DGX，使用的是 AMD 7 纳米制程的 Rome 架构 CPU。

据介绍，Grace 采用了更为先进的 5nm 制程，在内部通信能力上，它使用了英伟达第四代 NVIDIA NVLink，在 CPU 和 GPU 之间提供高达 900 GB/s 的双向带宽，相比之前的产品提升了八倍。Grace 还是第一个通过错误校正代码（ECC）等机制利用 LPDDR5x 内存系统提供服务器级可靠性的 CPU，同时提供 2 倍的内存带宽和高达 10 倍的能源效率。在架构上，它使用下一代 Arm Neoverse 内核，以高能效的设计提供高性能。

基于这款 CPU 和仍未发布的下一代 GPU，瑞士国家超级计算中心、苏黎世联邦理工大学将构建一台名为「阿尔卑斯」的超级计算机，算力 20Exaflops（目前全球第一超算「富岳」的算力约为 0.537Exaflops），将实现两天训练一次 GPT-3 模型的能力，比目前基于英伟达 GPU 打造的 Selene 超级计算机快 7 倍。
美国能源部下属的洛斯阿拉莫斯国家实验室也将在 2023 年推出一台基于 Grace 的超级计算机。

Grace 可以说是英伟达在今年 GTC 上最引人关注的产品了，它高度特化的设计与通过 PCIe 链接的 x86 版 CPU-GPU 系统大为不同，进而可以实现更好的性能。

GPU+CPU+DPU，三管齐下

「简单说来，目前市场上每年交付的 3000 万台数据中心服务器中，有 1/3 用于运行软件定义的数据中心堆栈，其负载的增长速度远远快于摩尔定律。除非我们找到加速的办法，否则用于运行应用的算力将会越来越少，」黄仁勋说道。「新时代的计算机需要新的芯片、新的系统架构、新的网络、新的软件和工具。」
除了造 CPU 的大新闻以外，英伟达还在一个半小时的 Keynote 里陆续发布了大量重要软硬件产品，覆盖了 AI、汽车、机器人、5G、实时图形、云端协作和数据中心等领域的最新进展。英伟达的技术，为我们描绘出了一幅令人神往的未来愿景。

- 首先是用于训练 Transformers 的框架—— NVIDIA Megatron。Transformers 已帮助开发者在自然语言处理领域取得了突破性进展。
- 面向医药领域，英伟达发布了一些用于计算药物研发加速库 Clara Discovery 的新模型，并介绍了一个基于物理学和机器学习的顶尖药物研发与材料科学计算平台 Schrodinger。
- 在量子计算领域中，英伟达发布了量子计算模拟环境 cuQUANTUM，其有助于加快有赖于量子位（或量子比特，能作为单个的 0 或 1 存在，也可以同时作为二者存在）的量子计算研究，为量子电路模拟器提供加速，从而助力研究人员设计出更完善的量子计算机。

- 为了保障现代化数据中心的安全，英伟达发布了 Morpheus 数据中心安全平台，其基于 NVIDIA AI、NVIDIA BlueField、Net-Q 网络遥测软件和 EGX 而构建，能够对完整的数据包进行实时检测。

- 为加快对话式 AI 的发展，英伟达发布了对话人工智能——NVIDIA Jarvis 的新版本，其能够实现语音识别、语言理解、翻译和表达性语音，同时也支持了更多种类的语言。

- 推荐系统是用于搜索、广告、在线购物、音乐、书籍、电影、用户生成内容和新闻等领域的引擎，为加快推荐系统的速度，黄仁勋宣布 NVIDIA Merlin 现可通过 NGC（NVIDIA 的深度学习框架容器目录）获取。

- 为帮助客户将自身专业知识应用于 AI 领域，同时保护数据隐私，英伟达发布了 NVIDIA TAO，其可以运用客户和合作伙伴的数据，对 NVIDIA 预训练模型进行微调和适配。

- 推理服务器 NVIDIA Triton，它可以从进入客户 EGX 服务器或云实例的连续数据流中获取洞察。黄仁勋说：「这包括任何在 cuDNN 上运行的 AI 模型，也就是几乎所有的 AI，包括来自 TensorFlow、Pytorch、ONNX、OpenVINO、TensorRT 或自定义 C++/python 后台等的任何框架。」

- 黄仁勋发布了 BlueField-3 DPU，其将为构建超大规模数据中心、工作站和超级计算机所需的基础设施提供进一步的加速。这款新一代数据处理器将提供最强大的软件定义网络、存储和网络安全加速功能。据介绍，一个 BlueField-2 能够实现相当于 30 块 CPU 核的工作负载，而 BlueField-3 在此基础上又实现了 10 倍的性能飞跃，能够替代 300 个 CPU 核，以 400Gbps 的速率，对网络流量进行保护、卸载和加速。

黄仁勋表示，英伟达全新的数据中心路线图已包括 CPU、GPU 和 DPU 三类芯片，而 Grace 和 BlueField 是其中必不可少的关键组成部分。投身 Arm 架构的 CPU，并不意味着英伟达会放弃原有的 x86、Power 等架构，黄仁勋将英伟达重新定义为「三芯片」公司，覆盖 CPU、GPU 和 DPU。
对于未来的发展节奏，黄仁勋表示：「我们的发展将覆盖三个产品线——CPU、GPU 和 DPU，以每两年一次更新的节奏进行，第一年更新 x86，第二年就更新 Arm。」

最后是自动驾驶。「对于汽车而言，更高的算力意味着更加智能化，开发者们也能让产品更快迭代。TOPS 就是新的马力，」黄仁勋说道。
英伟达将于 2022 年投产的 NVIDIA 自动驾驶汽车计算系统级芯片——NVIDIA DRIVE Orin，旨在成为覆盖自动驾驶和智能车机的汽车中央电脑。搭载 Orin 的量产车现在还没法买到，但英伟达已经在为下一代，超过 L5 驾驶能力的计算系统作出计划了。

Atlan 是这家公司为汽车行业设计的下一代 SoC，其将采用 Grace 下一代 CPU 和下一代安培架构 GPU，同时也集成数据处理单元 (DPU)。如此一来，Atlan 可以达到每秒超过 1000 万亿次（TOPS）运算次数。如果一切顺利的话，2025 年新生产的车型将会搭载 Atlan 芯片。

与此同时，英伟达还展示了 Hyperion 8 自动驾驶汽车平台，业内算力最强的自动驾驶汽车模板——搭载了 3 套 Orin 中心计算机。

不知这些更强的芯片和系统，能否应付未来几年里人们对于算力无穷无尽的需求。在 GTC 2021 上，英伟达对于深度学习模型的指数增长图又更新了。「三年间，大规模预训练模型的参数量增加了 3000 倍。我们估计在 2023 年会出现 100 万亿参数的模型。」黄仁勋说道。

英伟达今天发布的一系列产品，让这家公司在几乎所有行业和领域都能为你提供最强大的机器学习算力。在黄仁勋的 Keynote 发表时，这家公司的股票一度突破了 600 美元大关。

「20 年前，这一切都只是科幻小说的情节；10 年前，它们只是梦想；今天，我们正在实现这些愿景。

英伟达每年在 GTC 大会上发布的新产品，已经成为了行业发展的风向。不知在 Grace 推出之后，未来我们的服务器和电脑是否会快速进入 Arm 时代。