英伟达官宣全球最强AI芯片：性能提升30倍 将重新设计底层软件堆栈

来源：InfoQ

北京时间凌晨 4：00，大洋彼岸的美国加利福尼亚州圣何塞的圣何塞会议中心，被称为英伟达技术盛宴的 GTC 2024 大会正如火如荼地进行着。作为英伟达 2024 的开年大戏，身着标志性皮夹克的万亿富豪黄教主站在舞台中央，平静地甩出继 H100、A100 后的又一系列“核弹”级超级芯片。

今年的 GTC 之所以万众瞩目，是因为过去一年英伟达在 AI 领域的财务业绩方面取得了巨大成功。从 Volta V100 GPU 系列到最新的 Ampere A100 和 Hopper H100 芯片，该公司一直问鼎 AI 芯片之王。

## 1.GPU 家族再添“新丁”，全新 Blackwell 架构芯片炸场

在本届 GTC 大会开始之前，国外媒体就已经开始盛传：黄仁勋将在 GTC 2024 上发布一款 GPU 家族的新品，果然，采用 Blackwell 架构的 B200 和 GB200 系列芯片如期而至。

据英伟达称，Blackwell 架构系列芯片是迄今为止功能最强大的 AI 芯片家族。

据老黄介绍，B200 拥有 2080 亿个晶体管（而 H100/H200 上有 800 亿个晶体管），采用台积电 4NP 工艺制程，可以支持多达 10 万亿个参数的 AI 模型，而 OpenAI 的 GPT-3 由 1750 亿个参数组成。它还通过单个 GPU 提供 20 petaflops 的 AI 性能——单个 H100 最多可提供 4 petaflops 的 AI 计算。

但值得注意的是，Blackwell B200 并不是传统意义上的单一 GPU。它由两个紧密耦合的芯片组成，这两个芯片通过 10 TB/s NV-HBI（Nvidia 高带宽接口）连接进行连接，以确保它们能够作为单个完全一致的芯片正常运行。

该 GPU 平台以数学家 David Harold Blackwell 的名字命名，继承了英伟达两年前推出的 Hopper 架构，基于该架构一系列产品使英伟达的业务及其股价飙升。

该架构在 AI 安全方面又向前迈进了重要一步。Blackwell 通过 100% 系统内自测试 RAS 服务和全性能加密提供安全的 AI，也就是说数据不仅在传输过程中安全，而且在静止状态和计算时也安全。

Blackwell 将被整合到英伟达的 GB200 Grace Blackwell 超级芯片中，该芯片将两个 B200 Blackwell GPU 连接到一个 Grace CPU。英伟达没有透露价格。

新芯片预计将于今年晚些时候上市。英伟达表示，AWS、戴尔科技、谷歌、Meta、微软、OpenAI 和特斯拉计划使用 Blackwell GPU。



“生成式人工智能是我们这个时代的决定性技术，”老黄在演讲时表示。“Blackwell GPU 是推动这场新工业革命的引擎。与世界上最具活力的公司合作，我们将实现人工智能对每个行业的承诺。”

英伟达还发布了 GB200 NVL72 液冷机架系统，其中包含 36 颗 GB200 Grace Blackwell 超级芯片，拥有 1440 petaflops（又名 1.4 exaflops）的推理能力，它内部有近两英里长的电缆，共有 5000 根单独的电缆。

英伟达表示，与用于推理用途的相同数量的 H100 Tensor Core 图形处理单元相比，GB200 NVL72 性能提升高达 30 倍。此外，该系统还可将成本和能耗降低多达 25 倍。



GB200 NVL72

例如，训练一个 1.8 万亿参数模型之前需要 8000 个 Hopper GPU 和 15 兆瓦的功率。如今，只需要 2000 个 Blackwell GPU 就可以做到这一点，而功耗仅为 4 兆瓦。

在具有 1750 亿个参数的 GPT-3 基准测试中，英伟达表示 GB200 的性能是 H100 的 7 倍，训练速度是 H100 的 4 倍。

此外，英伟达称还将推出一款名为 HGX B200 的服务器主板，它基于在单个服务器节点中使用 8 个 B200 GPU 和一个 x86 CPU（可能是两个 CPU）。每个 B200 GPU 可配置高达 1000W，并且 GPU 提供高达 18 petaflops 的 FP4 吞吐量，因此比 GB200 中的 GPU 慢 10%。

目前，企业客户可以通过 HGX B200 和 GB200（将 B200 GPU 与 英伟达的 Grace CPU 结合在一起）访问 B200。

## 2.全面升级软件服务

市场正在升温，硬件和软件方面的竞争都在加剧。在本次 GTC 中，英伟达不仅通过新的硬件创新来应对竞争，还展示了其 AI 软件战略如何帮助确定其在该领域的领导地位，以及未来几年将如何发展。

黄仁勋还着力推销其 AI 软件订阅服务包，这显然是在配合该公司向“以软件卖硬件”的新战略，也是在与过往的“以硬件卖软件”的战略彻底告别。

英伟达可以访问所有领域的大量模型，但他们认为对于企业来说它们仍然太难使用。他们推出了 Nvidia 推理微服务（NIM），将模型和依赖项整合到一个简洁的包中，根据用户的堆栈进行优化，并与易于使用的 API 连接。



经过打包和优化的预训练模型，可在 NVIDIA 的安装基础上运行，包含运行它所需的所有软件。CUDA 库、API 等，基本上是容器化的 AI 软件包，针对 NV GPU 进行了优化，并带有一个简单的 API 来访问它们。

老黄指出：“这就是我们未来编写软件的方式”——通过组装一堆人工智能。



老黄我们介绍了英伟达如何使用 英伟达推理微服务（NIM）创建一个内部聊天机器人，旨在解决构建芯片时遇到的常见问题。“我们需要一个模拟引擎，以数字方式为机器人呈现世界，”他说，这就是 Omniverse。这些“微服务”将允许开发人员使用专有和自定义模型快速创建和部署“副驾驶”或人工智能助手。

他表示，机器人技术与人工智能和 Ominverse/Digital Twin 工作一起成为英伟达的关键支柱，所有这些都共同努力以充分利用公司的系统。

据悉，Omniverse 是一个专为构建和操作 Metaverse 应用程序而设计的平台，本质上是人们可以交互、工作和创建的共享虚拟世界。Omniverse 平台可以创建数字孪生和高级模拟。英伟达对 Omniverse 的愿景包括成为 Metaverse 的基础平台，创作者和企业可以在共享虚拟空间中进行协作。在 Omniverse 中创建的数字孪生可用于 Metaverse 中的各种应用，例如虚拟培训、产品设计和预测性维护。

老黄表示英伟达已经推出了数十种企业级生成式 AI 微服务，企业可以使用这些服务在自己的平台上制作应用程序，同时保留对其知识产权的完全所有权和控制权。

老黄还宣布将 Omniverse Cloud 流传输至 Apple Vision Pro 耳机。



他也表示，英伟达表示正认真考虑从根本上重新设计整个底层软件堆栈，希望借 AI 之力为人类生成更优质的代码。

之所以会有这样的想法，原因非常简单：几十年来，整个世界一直受制于围绕 CPU 发展出的传统计算框架，即由人类编写应用程序以检索数据库中准备好的信息。

黄仁勋在发布会上指出，“我们今天的计算方式，首先需要确定信息是由谁编写、由谁创建的，也就是要求信息先要被记录下来。”

而英伟达的 GPU 为加速计算开辟出一条通往算法化计算的新路，可以依托创造性推理（而非固有逻辑）来确定相关结果。

此外，英伟达希望通过发布另一个新的 API 集合 Project GROOT 来推动人形机器人的开发。

Project GROOT 是一个人形机器人模型，英伟达与 Jetson Thor 一起生产，Jetson Thor 是一款 SoC，也是 Nvidia Isaac 的升级版。英伟达表示，GROOT 机器人将理解自然语言并模仿人类动作来学习灵活性。Jetson Thor 运行基于 Blackwell 的 GPU，可在 8 位数据处理中提供 800 teraflops 的 AI 性能。



老黄透露，由该平台驱动的机器人将被设计为能够理解自然语言并模仿机器人的动作，观察人类行为。这使 GROOT 机器人能够快速学习协调性、灵活性和其他技能，以导航、适应现实世界并与之互动——并且绝对不会导致机器人叛乱。

“为通用人形机器人构建基本模型是我们当今人工智能领域能够解决的最令人兴奋的问题之一，”老黄说。“这些使能技术正在融合在一起，使世界各地领先的机器人专家能够在人工通用机器人领域取得巨大飞跃。”

## 3.对开发者的影响

根据专家预测，五年之后，文本、图像、视频和语音等形式的信息将全部被实时输入大语言模型（LLM）。届时计算机将直通所有信息源，通过多模态交互不断实现自我改进。

黄仁勋表示，“未来，我们将步入持续学习的时代。我们可以决定是否部署持续学习的成果，而且与计算机的交互不会再借助 C++。”

这就是 AI 技术的意义所在——人类可以在推理之后，要求计算机生成代码以实现特定目标。换句话说，未来人们可以用简单的语言、而非 C++ 或者 Python，与计算机实现顺畅交流。

“在我看来，编程本身的价值正在悄然跨过历史性的衰退拐点。”黄仁勋还补充称，AI 已经在弥合人类与技术之间的鸿沟。

“就在当下，约有上千万人凭借自己的计算机编程知识来谋取职位、赚得收益，而余下的 80 亿人则被他们远远甩在身后。未来的情况将有所改变。”

在黄仁勋看来，英语将成为最强大的编程语言，而个性化交互则是缩小技术鸿沟的关键因素。

生成式 AI 将成为一种宏观层面的操作系统，人类可以在其中用简单的语言指示计算机创建应用程序。黄仁勋表示，大语言模型将帮助人类通过计算机把自己的灵感转化为现实。

例如，人类已经可以要求大语言为特定领域的应用程序生成 Python 代码，且全部提示内容均使用简单英语编写而成。

“我们要如何让计算机按自己的想法做事？我们要如何在计算机上实现指令微调？这些问题的答案就是提示词工程，而且更多是种艺术、而非单纯的技术。”

也就是说人类将可以专注于领域专业知识，而生成式 AI 将补齐编程技能这块短板。黄仁勋认为这将彻底颠覆软件的开发格局。

黄仁勋此前曾将大语言模型比作经过预培训且头脑灵光的大学毕业生。英伟达正围绕大模型提供医疗保健与金融等领域的专业知识，借此为企业客户提供高效支持。