8 Big Announcements at Nvidia's GTC 2023: From Generative AI Services to New GPUs

MARCH 22, 2023, 05:25 PM EDT

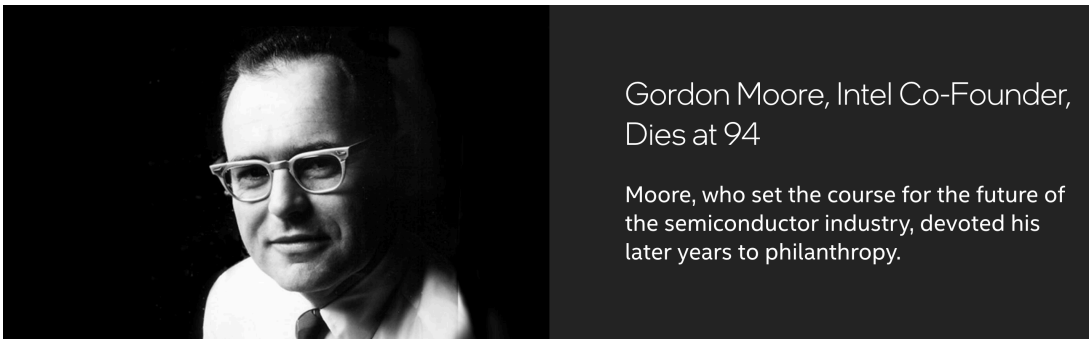
At Nvidia's GTC 2023 event, the chip designer revealed new cloud services meant to help enterprises build generative AI and metaverse applications as well as new GPUs, systems and other components. CRN outlines the biggest announcements that will open new opportunities for partners.

New Cloud Services, GPUs And More

Nvidia unveiled new cloud services aimed at helping enterprises build generative AI and metaverse applications alongside new GPUs, systems and other components at the chip designer's GTC 2023 event.

Nvidia 首席执行官兼联合创始人 Jensen Huang 在周二的主题演讲中，将 Nvidia 定位为加速计算需求解决方案的卓越供应商，为生成式人工智能等应用提供算力支持。并表示，无论从性能还是效率的角度来看，传统 CPU 都不适合这些要求越来越高的工作

负载。他声称摩尔定律正在放缓并即将结束（Gordon Moore died on March 24, 2023）。



“As Moore’s Law ends, increasing CPU performance comes with increased power. And the mandate to decrease carbon emissions is fundamentally at odds with the need to increase data centers,” Huang said. “Cloud computing growth is power-limited. First and foremost, data centers must accelerate every workload. Acceleration will reclaim power. The energy saved can fuel new growth.”

虽然在过去几年中英特尔最新的芯片制造技术的多次延迟引起了人们对摩尔定律的持久力的质疑，但其首席执行官帕特-盖尔辛格曾表示，他的公司可以 "在未来十年内保持甚至甚至比摩尔定律更快"。

Nvidia 在 GTC 2023 上宣布了一些新的产品和服务，以扩大 GPU 支持的应用，提高 GPU 性能，使 GPU 计算更容易获得。这些产品和服务包括 DGX Cloud AI 超级计算服务，用于定制生成性 AI 应用的 AI Foundations 服务，L4 和 H100 NVL 专用 GPU，以及 Omniverse Cloud 平台即服务。该公司还披露了围绕其基于 Arm 的 Grace 和 Grace Hopper 芯片的更多细节，并说它已经开始生产其 BlueField-3 数据处理单元。

以下是 Nvidia 在 GTC 2023 上发布的八大公告的综述，这些公告将为渠道合作伙伴创造新的机会，其中包括新产品和服务以及支持它们的云服务提供商和服务器供应商。

Nvidia Launches DGX Cloud with Support from Oracle, Microsoft And Others

Nvidia 在 GTC 2023 上推出了新的人工智能超级计算服务，并从 Oracle 云基础设施中获得了初步可用性。

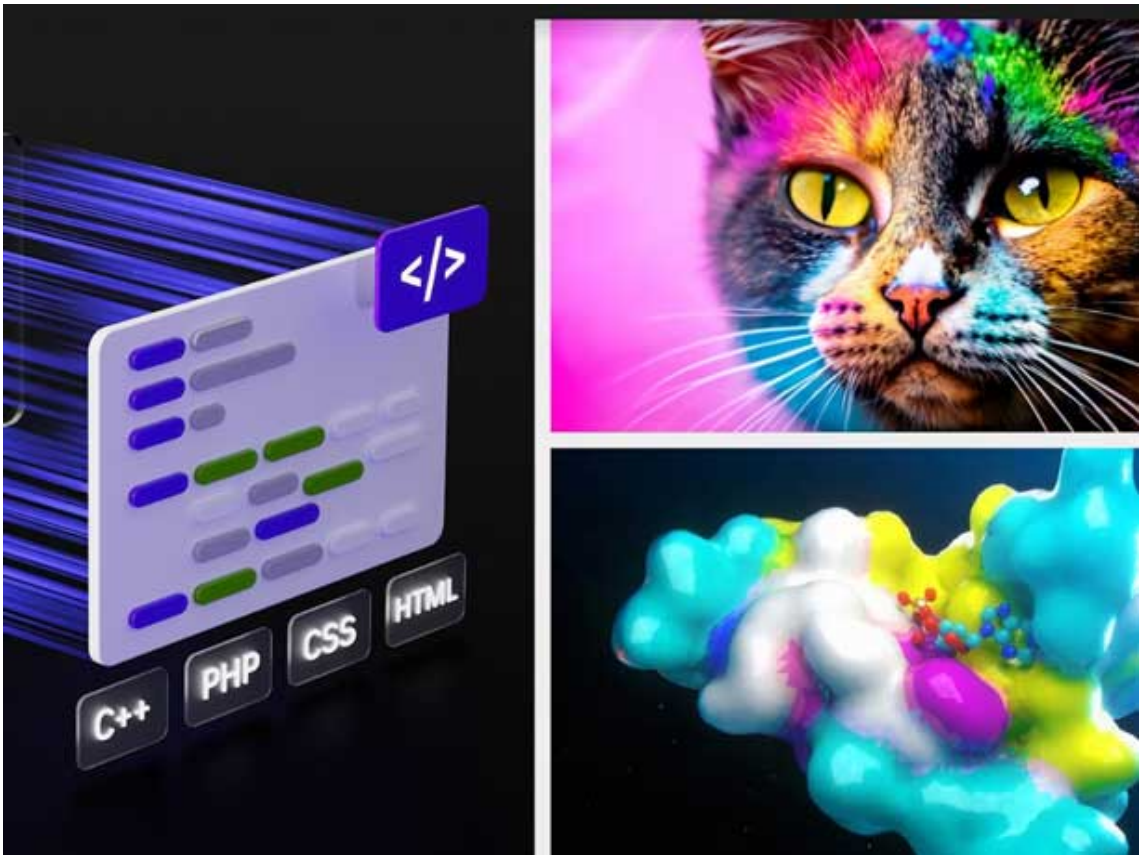
DGX Cloud 是一项服务，旨在让企业快速获得他们所需的工具和基础设施，为生成性人工智能和其他应用训练深度学习模型。它通过将 Nvidia 的 DGX 超级计算机和人工智能软件结合起来，并通过云服务提供商托管的实例在网络浏览器中提供这些服务。

“We are at the iPhone moment of AI. Startups are racing to build disruptive products and business models while incumbents are looking to respond. Generative AI has triggered a sense of urgency in enterprises worldwide. To develop AI strategies, customers need to access Nvidia AI easier and faster,” Nvidia 首席执行官 Jensen Huang 在 GTC 2023 的主题演讲中说。

DGX 云实例现在可以从甲骨文云基础设施获得。该服务预计将在第三季度登陆微软 Azure，Nvidia 说它将 "很快扩展到谷歌云和其他地方"。

在软件层，DGX 云包括 Nvidia 的 Base Command 平台，该平台管理和监控训练工作负载，它可以让用户根据自己的需要调整基础设施的规模。该服务还包括对 Nvidia AI Enterprise 的访问，这是一个软件套件，包括 AI 框架、预训练模型和用于开发和部署 AI 应用程序的工具，以及其他组件。

在推出时，每个 DGX 云实例将包括 8 个 A100 80GB 的 GPU，这些 GPU 是在 2020 年底推出的。这八颗 A100 的组合使节点的 GPU 总内存达到 640GB。基于 A100 的实例的月费将从 36,999 美元开始，长期承诺可获得折扣。
Nvidia 计划在未来某个时候提供配备 Nvidia H100 80GB GPU 的 DGX 云实例。



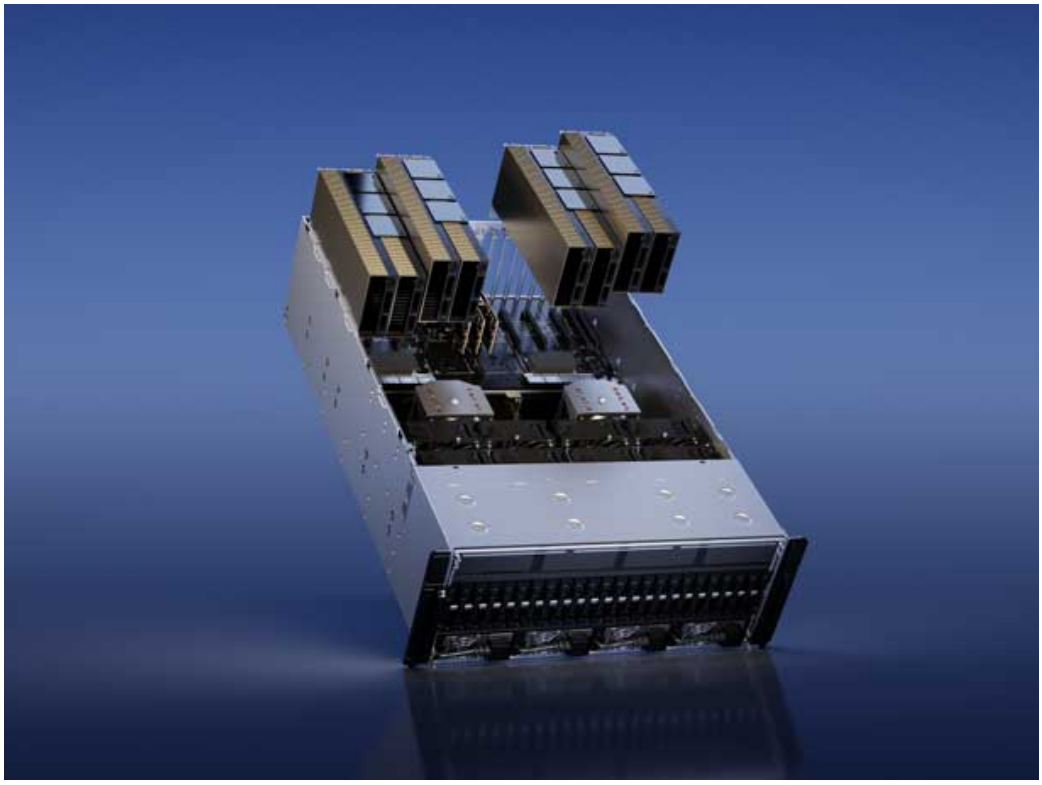
New Services Help Enterprise Build Generative AI Models from Proprietary Data

Nvidia 披露了一个新的人工智能云服务，统称为 Nvidia AI Foundations，旨在帮助企业使用**专有数据集**建立和运行定制的大型语言模型和生成型人工智能模型。

这些服务在 Nvidia 新的 DGX Cloud AI 超级计算服务上运行，带有预训练的模型、数据处理框架、矢量数据库和个性化功能、优化的推理引擎、API 和企业支持。

Nvidia NeMo 是一项大型语言模型定制服务，包括企业提取专有数据集的方法，并将其用于生成式人工智能应用，如聊天机器人、企业搜索和客户服务。NeMo 可在早期访问中使用。

Nvidia Picasso 专注于从创意、设计和模拟应用的文本提示中生成 AI 驱动的图像、视频和 3D 模型。该服务允许企业在专有数据上训练 Nvidia 的 Edify 基础模型。该服务还配备了基于完全授权数据集的预训练 Edify 模型。Picasso 在预览模式下可用。



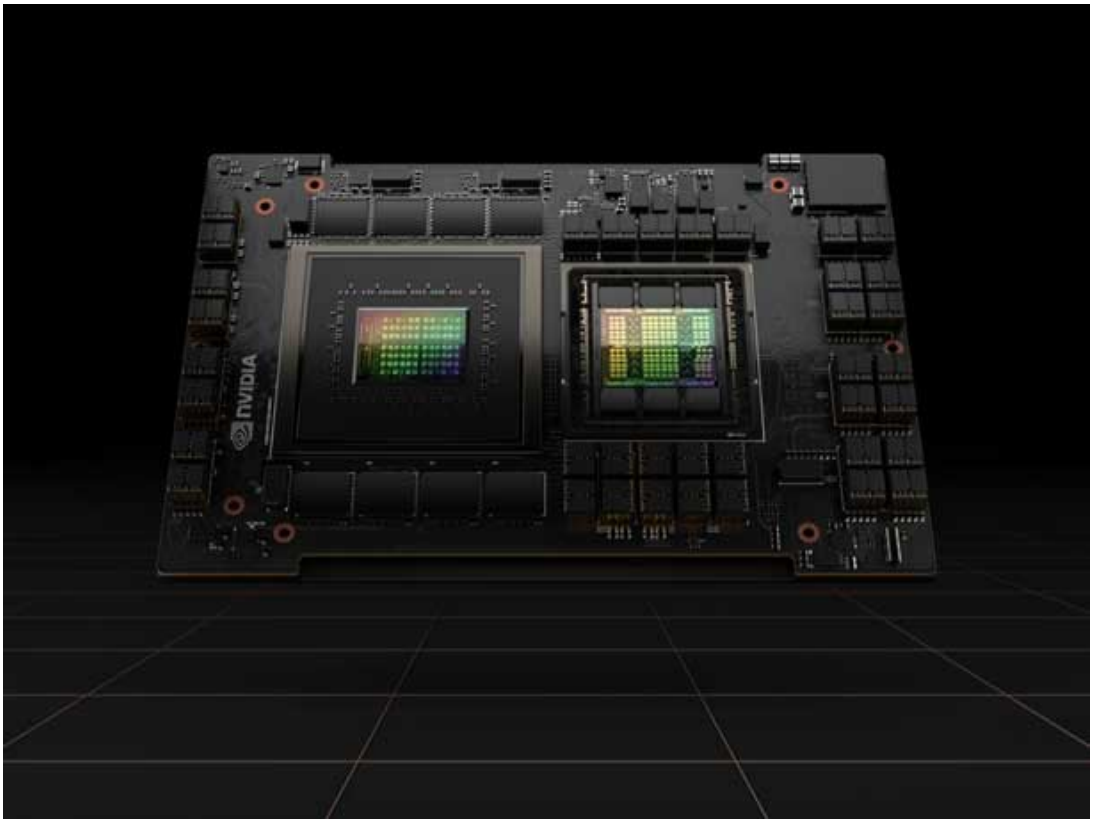
New Specialized Data Center GPUs for AI, Graphics

Nvidia 扩大了其数据中心 GPU 的组合，推出了专门用于 AI 驱动的视频性能和大型语言模型的产品。

为人工智能驱动的视频应用而设计，Nvidia 正在推销 L4 GPU，在这些工作负载中，它比 CPU 更快、更节能。据 Nvidia 称，与配备两个英特尔至强铂金 8380（来自 2021 年第三代至强可扩展处理器）的双插槽服务器相比，配备八个 L4 GPU 的服务器速度快 120 倍，能效高 99%。GPU 是为视频解码和转码、视频流、增强现实、AI 生成的视频和其他视频工作负载而构建的。

谷歌云已经推出了由 L4 驱动的实例与 G2 虚拟机的私人预览。该 GPU 也可用于 30 多家供应商的系统，包括华硕、Atos、思科系统、戴尔科技、技嘉、惠普企业、联想和超微。

该公司还披露了 H100 NVL，它结合了两块 H100 PCIe 卡，并通过 NVlink 桥连接它们。这款 GPU 旨在对大规模的大型语言模型（如流行的 ChatGPT）进行推理，它配备了 94GB 的内存，由于 Hopper 的 Transformer Engine，与 A100 相比，GPT-3 模型的推理性能快了 12 倍。H100 NVL 预计将在今年下半年推出。



Grace, Grace Hopper Superchips

Grace 超级芯片和结合了 CPU 和 GPU 的 Grace Hopper 超级芯片将如何加速新工作负载。虽然这两个芯片模块现在正在向客户提供样品，但在 2022 年多次承诺系统将在 2023 年上半年到达后，它已将其在系统内的可用性推迟到下半年。

预计支持 Grace 超级芯片的服务器供应商包括华硕、Atos、技嘉、惠普企业、QCT、超微、纬创和 ZT。

Nvidia 首席执行官黄仁勋在 GTC 2023 上强调，传统的 CPU 并不适合许多加速计算应用，新设计的 Grace CPU 由 72 个 Arm 兼容的内核组成, to “excel at single-threaded execution and memory processing” and provide “high energy efficiency at cloud data center.” These facets make the CPU “excellent for cloud and scientific computing applications,” Huang added.

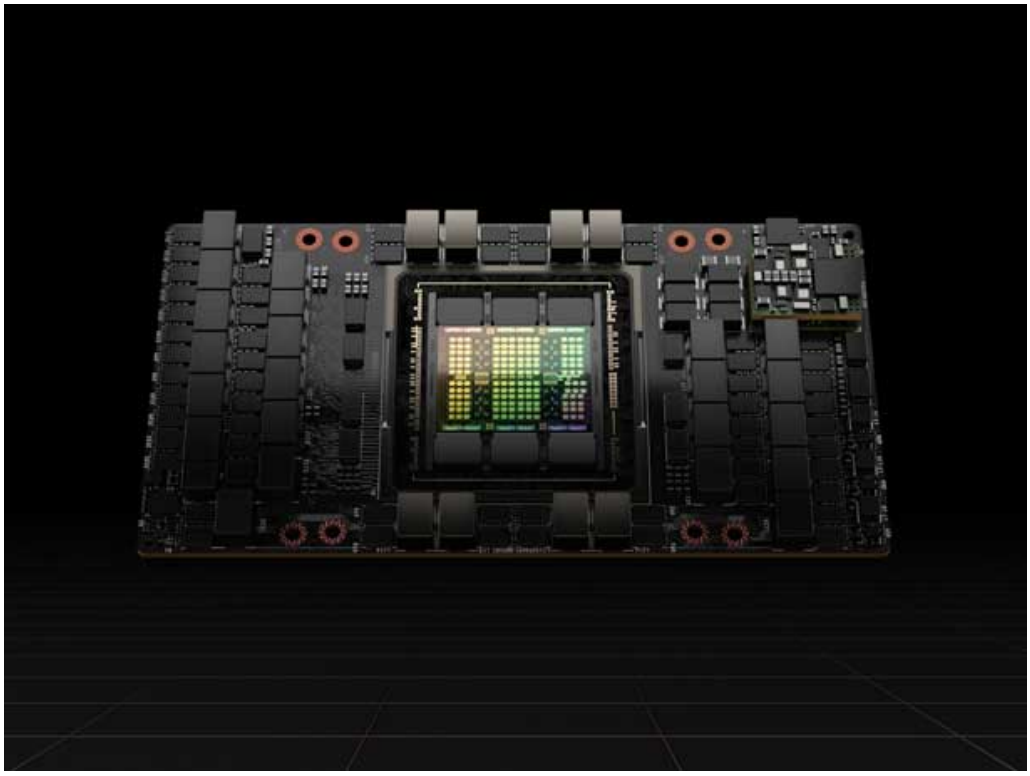
Grace 超级芯片结合了两个 Grace CPU，共提供 144 个内核，这些内核通过 900GB/s 的低功耗芯片间相干接口连接。在内存方面，该超级芯片拥有 "服务器级 "的 LPDDR5X，Nvidia 称这是第一个这样做的数据中心 CPU。

黄仁勋说，Nvidia 在一个流行的谷歌基准上测试了 Grace，该基准用于衡量云计算微服务的性能，以及一套测试内存密集型处理的 Apache Spark 基准，黄仁勋称这是 "foundation for cloud data centers."。

与最新一代 x86 CPU 相比，Grace 在微服务方面平均快 30%，在数据处理方面快 20%，而 "只使用在整个服务器节点上测量的 60%的功率"，（没有指明竞争对手的 CPU）。他说："云服务提供商可以在电力有限的数据中心配备 1.7 倍的 Grace 服务器，每台服务器的吞吐量提高 25%。

虽然 Nvidia 为云计算和科学计算设计了 Grace 超级芯片，但其为 Grace Hopper 超级芯片量身定做的 "处理巨型数据集，如用于推荐系统和大型语言模型的 AI 数据库"。

它非常适合这些数据量大的应用，因为 Grace Hopper 将 Grace CPU 和 Hopper GPU 集合在一个模块上，并通过高速接口将它们连接起来。这种接口 "比 PCI Express 快七倍"，黄仁勋说，这意味着这种超级芯片比通过 PCIe 通信的独立 CPU 和 GPU 更适合这些应用。" This is Nvidia’ s inference platform, one architecture for diverse AI workloads and maximum data center acceleration and elasticity"。



Availability Expands for H100-Based Services And Systems

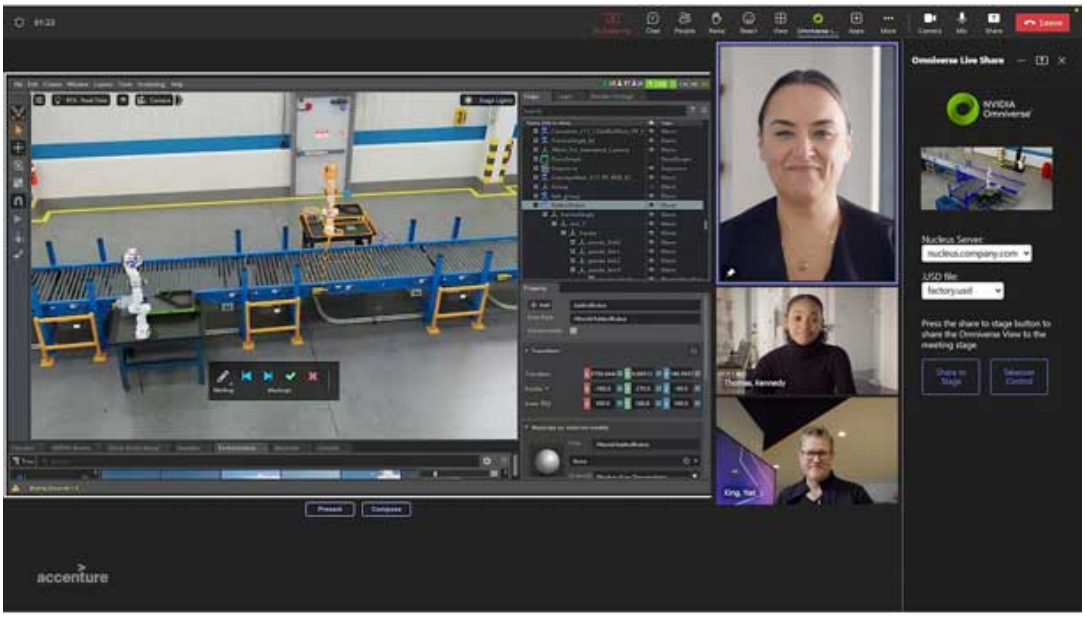
Nvidia 宣布，在其强大的、基于 Hopper 的 H100 数据中心 GPU 上运行的产品和服务正在推广普及，用于加速 AI 训练和推理。

H100 于去年首次发布，是基于安培的 A100（2020 年首次亮相）的继任者。新的 Hopper 架构及其内置 Transformer Engine，H100 与 A100 相比，在大型语言模型上，训练速度快 9 倍，推理速度快 30 倍。

在云计算方面，Nvidia 表示，甲骨文云基础设施已经推出了新的 OCI 计算裸机 GPU 实例，配备了 H100 GPU，供应有限。亚马逊提供 H100 动力服务，可以扩展到 20000 个 H100 GPU。

基于 H100 的云实例目前可从 Cirrascale 和 CoreWeave 获得。Nvidia 表示，谷歌云、Lambda、Paperspace 和 Vultr 都有搭载 H100 的实例。

至于企业内部的服务器，Nvidia 说使用 H100 GPU 的系统现在可以从 Atos、Cisco Systems、Dell Technologies、Gigabyte、Hewlett Packard Enterprise、Lenovo 和 Supermicro 获。Nvidia 的 DGX H100 也已上市，它具有八个用 NVLink 互连的 H100 GPU。



Omniverse Cloud Limited Availability Begins with Microsoft Azure

Nvidia 已经向选定的企业提供了 Omniverse Cloud，这是其用于创建和运行 3-D 互联网应用的新 platform-as-a-service，该公司将其称为 "industrial metaverse"。

微软 Azure 是第一个提供基于订阅的 Omniverse Cloud 的云服务提供商，它提供了对 Nvidia 在其 OVX 基础设施上的全套 Omniverse 软件应用的访问。该芯片设计公司的 OVX 服务器使用其 L40 GPU，该 GPU 为图形密集型应用以及人工智能驱动的二维图像、视频和三维模型的生成进行了优化。

虽然 Nvidia 过去曾将 Omniverse 推广到各行各业，但该公司将 Omniverse Cloud 的推广工作集中在汽车行业，称它可以让设计、工程、制造和营销团队将其工作流程数字化。
Omniverse Cloud 实现的应用包括连接三维设计工具以加快汽车开发，创建工厂的数字双胞胎以模拟生产线的变化，以及对车辆运行进行闭环模拟。
该服务将在今年下半年广泛提供。

BlueField-3 DPU Now in Production, Adopted By Oracle Cloud

Nvidia 宣布，其 BlueField-3 数据处理单元 DPU 现已全面投入生产，并被 Oracle 云基础设施纳入其网络堆栈。

BlueField-3 是 Nvidia 的第三代 DPU，它旨在降低 CPU 的网络、存储和安全工作负载并加速，同时还能实现新的安全和管理程序功能。该公司表示，测试表明，与没有该组件的系统相比，装有 BlueField DPU 的服务器的功耗降低了 24%。

BlueField-3 支持以太网和 InfiniBand 连接，运行速度高达 400 千兆比特/秒，与 BlueField-2 相比，它的计算速度提高了四倍，加密加速速度提高了四倍，存储处理速度提高了两倍，内存带宽提高了四倍。
除了甲骨文云，BlueField-3 还被微软 Azure 和 CoreWeave 使用。Nvidia 表示，Check Point Software、Cisco Systems、Dell Technologies、Juniper Networks、Palo Alto Networks、Red Hat 和 VMware 等公司正在开发使用 BlueField DPU 的解决方案。

New RTX 4000 GPUs For Laptops, Compact GPU For Workstation

Nvidia 发布了六种新的 RTX GPU，用于运行繁重内容创作、设计和 AI 应用的笔记本电脑和紧凑型工作站。

该公司表示，笔记本电脑本月晚些时候将推出新的 RTX 5000、RTX 4000、RTX 3500、RTX 3000 和 RTX 2000 GPU，所有这些 GPU 都使用与 Nvidia 面向消费者的 GeForce RTX 40 系列显卡相同的 Ada Lovelace 代。

据 Nvidia 称，这些 GPU 配备了高达 16GB 的内存来处理大型数据集，与上一代安培的 RTX GPU 相比，它们可以为工作站应用提供高达两倍的单精度浮点性能，高达两倍的光线追踪性能和高达两倍的 AI 训练性能。
Nvidia 正通过新的 RTX 4000 小尺寸 GPU 实现强大的紧凑型桌面工作站，该 GPU 带有来自 Ada Lovelace 架构的相同性能改进。该公司表示，PNY 和 Leadtek 等全球经销商将从 4 月开始以 1240 美元的估计价格提供这种紧凑型工作站 GPU。工作站供应商预计将在今年晚些时候发布配备新 GPU 的系统。

https://mp.weixin.qq.com/s/Rg-9FghZUcmuHb6ff34yfw