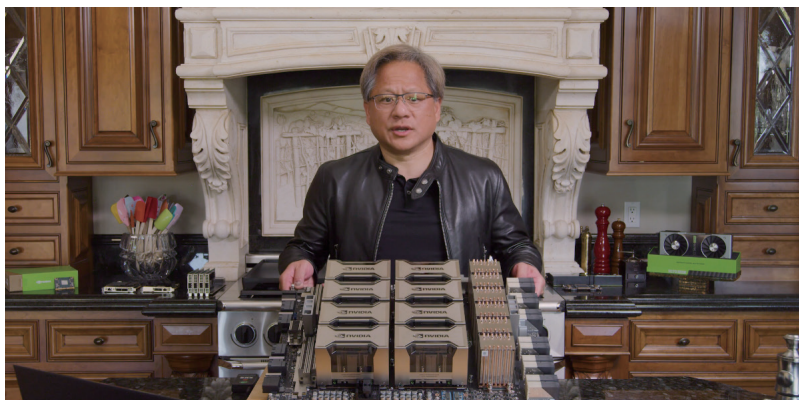


NVIDIA GTC 2020 大会：众多技术新品助力各行业 AI 技术发展

策划 | 《智能制造》编辑部

2020 年 5 月 14 日，NVIDIA 的 GTC 2020 主题演讲正式上线。受新冠肺炎疫情影响，原定于 2020 年 3 月下旬在旧金山举办的 GTC 2020 大会不得不延期，并改在 NVIDIA 创始人兼首席执行官黄仁勋家的厨房中在线举行。黄仁勋依然以一身黑色皮衣出场，只是头发较往常长了不少。在线上演讲开始前的采访中，已在家工作了超过 45 天的黄仁勋笑称，“你们明天一定认不出我的”。



GTC 大会是 NVIDIA 每年最重要的新品发布平台之一。这场主题演讲虽然经历了延期，还从会展中心改在了 CEO 的家里，但不可否认它仍然是 NVIDIA 和全球 GPU、AI 技术应用的年度重头戏。在主题演讲中，黄仁勋谈及了 NVIDIA 发展的各个主要方面，包括图形显卡、数据中心、AI 技术、医疗，甚至还有最新收购进来的网络互联业务等。

NVIDIA 全新 Ampere 数据中心 GPU 全面投产

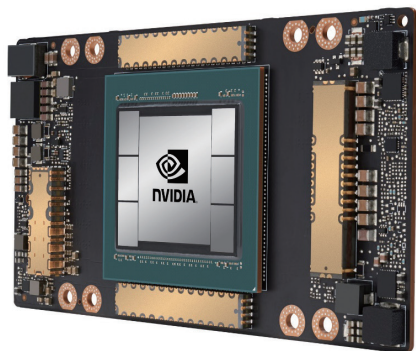
在本次 GTC 2020 宣布的产品中，最重磅的即是这款重达 50 磅的史上最大 GPU——首款基于 NVIDIA Ampere 架构的 NVIDIA A100。

NVIDIA A100 采用了 NVIDIA Ampere 架构的突破性设计，该设计为 NVIDIA 第八代 GPU 提供了迄今为止最大的性能飞跃，集 AI 训练和推理于一身，并且其性能相比于前代产品提升了高达 20 倍。作为一款通用型工作负载加速器，A100 还被设计用于数据分析、科学计算和云图形。

NVIDIA 创始人兼首席执行官黄仁勋表示：“云计算和

AI 发展的强劲势头正在推动数据中心设计的结构性转变，过去的纯 CPU 服务器正在被高效的加速计算基础架构所取代。NVIDIA A100 GPU 作为一个端到端的机器学习加速器，其实现了从数据分析到训练再到推理整个过程较前代产品提速 20 倍的 AI 性能飞跃。这是 NVIDIA 首次在同一平台上实现对横向扩展以及纵向扩展负载的加速。NVIDIA A100 将在提高吞吐量的同时，降低数据中心的成本。”

A100 所采用的全新弹性计算技术能够为每项工作分配适量的计算能力。多实例 GPU 技术可将每个 A100 GPU 分



割为多达七个独立实例来执行推理任务,而第三代NVIDIA NVLink互联技术能够将多个A100 GPU合并成一个巨大的GPU来执行更大规模的训练任务。

众多全球领先的云服务供应商和系统构建商正在计划将A100 GPU集成到其产品中,其中包括:阿里云、AWS、Atos、百度智能云、思科、Dell Technologies、富士通、技嘉科技、Google Cloud、新华三、HPE、浪潮、联想、Microsoft Azure、甲骨文、Quanta/QCT、Supermicro和腾讯云。

A100 的五大突破

NVIDIA A100 GPU的突破性技术设计来源于五大关键性创新。

◎ NVIDIA Ampere 架构: A100的核心是NVIDIA Ampere GPU架构,该架构包含超过540亿个晶体管,这使其成为全球最大的7纳米处理器。

◎ 具有TF32格式的第三代Tensor Core核心: NVIDIA第三代Tensor Core将变得更加灵活、快速且易于使用。其功能经过扩展后加入了专为AI开发的全新TF32格式,它能在无需更改任何代码的情况下,使FP32精度下的AI性能提高多达20倍。此外,Tensor Core核心现在支持FP64精度,相比于前代,其为HPC应用所提供的计算力比之前提高了多达2.5倍。

◎ 多实例GPU: 全新在MIG技术可以将单个A100 GPU分割为多达七个独立的GPU,为不同规模的工作提供不同的计算力,以此实现利用率和投资回报率的最大化。

◎ 第三代NVIDIA NVLink: 使GPU之间的高速连接增加至原来的两倍,实现服务器的高效性能扩展。

◎ 结构化稀疏: 利用AI数学固有的稀疏性,可以使GPU性能提升了一倍。

凭借这些新功能,NVIDIA A100将成为AI训练和推

理以及科学模拟、对话式AI、推荐系统、基因组学、高性能数据分析、地震建模和金融预测等各种高要求工作负载的理想选择。

NVIDIA DGX A100 新系统已搭载 NVIDIA A100, 很快将登陆云端

与NVIDIA A100同时发布的还有NVIDIA DGX A100系统,内置8个由NVIDIA NVLink互联的NVIDIA A100 GPU。目前阿里云、AWS、百度智能云、Google Cloud、Microsoft Azure、甲骨文和腾讯云正在计划提供基于A100的服务。

此外,众多全球领先的系统制造商也将提供各种基于A100的服务器,包括Atos、思科、Dell Technologies、富士通、技嘉科技、新华三、HPE、浪潮、联想、Quanta/QCT和Supermicro。

为帮助合作伙伴加快服务器的开发,NVIDIA开发了HGX A100服务器构建模块,该模块采用了多GPU配置的集成基板形式。

HGX A100的4-GPU版通过NVLink实现GPU之间的完整互联,8-GPU版配置的HGX A100通过NVIDIA NVSwitch实现GPU之间的全带宽通信。采用全新MIG技术的HGX A100能够被分割为56个小型GPU,每个GPU的速度都比NVIDIA T4更快;或者也可以将其用作一台拥有10petaflops AI性能的巨型8-GPU服务器。

DGX A100系统集成了8个全新NVIDIA A100 Tensor Core GPU,具有320GB内存,用以训练最大型的AI数据集,以及最新的高速NVIDIA Mellanox HDR 200Gbps互连。

利用A100的多实例GPU功能,每台DGX A100系统能够被分割为多达56个实例,用于加速多个小型工作负载的处理速度。凭借这些功能,企业可在一个完全集成的软件定义平台上根据自己的需求优化计算力和资源,加快数据分析、训练和推理等各种工作负载的速度。

NVIDIA 打造新一代 DGX SuperPOD, 性能达 700 Petaflops

NVIDIA还展示了新一代DGX SuperPOD。该集群由140台DGX A100系统组成,AI计算能力高达700Petaflops。利用Mellanox HDR 200Gbps InfiniBand互连技术,NVIDIA将140台DGX A100系统结合在一起,构建了DGX SuperPOD AI超级计算机,用于公司内部对

话式 AI、基因组学和自动驾驶等领域的研究。

为帮助客户构建他们自己的由 A100 赋能的数据中心，NVIDIA 发布了全新 DGX SuperPOD 参考架构。该参考架构为客户提供了指引蓝图，使客户能够参照 NVIDIA 所采取的设计原则和最佳实践经验，构建他们自己的基于 DGX A100 的 AI 超级计算集群。

DGXpert 计划与 DGX-Ready 软件

NVIDIA 还推出了 NVIDIA DGXpert 计划，帮助 DGX 客户与 NVIDIA 的 AI 专家建立联系；此外还有 NVIDIA DGX-Ready 软件计划，帮助客户在 AI 工作流程中充分利用各种经过认证的企业级软件。

DGXpert 计划将聚集一批精通 AI 的专家。他们可以帮助指导客户完成从计划、实施到持续优化的整个 AI 部署流程，还可以帮助 DGX A100 客户构建和维护最先进的 AI 基础架构。

NVIDIA DGX-Ready 软件计划可帮助客户快速找到和使用经 NVIDIA 测试的第三方 MLOps 软件，帮助客户提高数据科学生产效率、加速 AI 工作流程并增加 AI 基础架构的可访问性和利用率。该计划的首批 NVIDIA 认证合作伙伴有：Allegro AI、cnvrg.io、Core Scientific、Domino Data Lab、Iguazio 和 Paperspace。

DGX A100 技术规格

◎ 8 个 NVIDIA A100 Tensor Core GPU，AI 性能达 5Petaflops，GPU 内存共 320GB，有着每秒 12.4TB 的带宽。

◎ 6 个采用第三代 NVIDIA NVLink 技术的

NVIDIA NV Switch 互联结构，4.8TB 每秒的双向带宽。

◎ 9 个 Mellanox ConnectX-6 HDR 200Gb/s 网络接口，提供总计每秒 3.6TB 的双向带宽。

◎ Mellanox 网络计算和网络加速引擎，例如 RDMA、GPUDirect，以及 SHARP 技术（Scalable Hierarchical Aggregation and Reduction Protocol），实现最高性能和可扩展性。

◎ 15TB Gen4 NVMe 内存，速度比 Gen3 NVMe SSDs 快 2 倍。

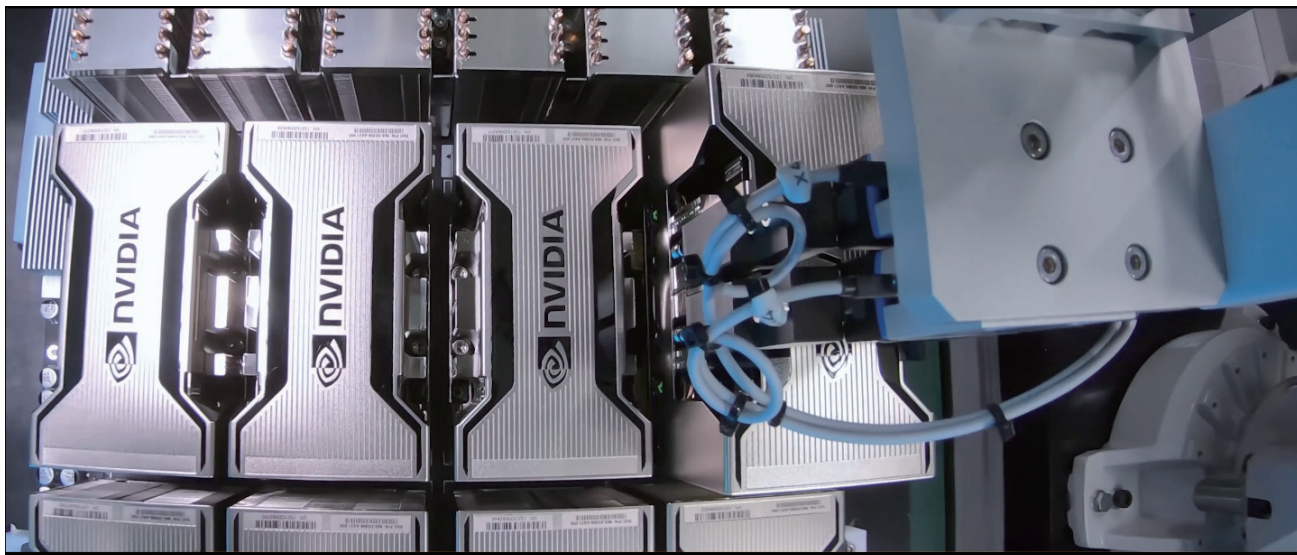
NVIDIA DGX 软件堆栈，其中包括专为 AI 和数据科学工作负载而优化的软件，能够实现性能的最大化，使企业能够更快获得 AI 基础架构投资回报。

一个由 5 台 DGX A100 系统组成的机架可代替一个包括了 AI 训练和推理基础设施的数据中心，而且功耗仅是其 1/20，所占用的空间是其 1/25，成本是其 1/10。

A100 软件优化

针对 A100，NVIDIA 还发布了多个软件堆栈更新，使应用程序开发者能够充分发挥 A100 GPU 创新技术的性能。这些更新包括了 50 多个新版本 CUDA-X 库，可用于加速图形、模拟和 AI；CUDA 11；多模态对话式 AI 服务框架 NVIDIA Jarvis；深度推荐应用框架 NVIDIA Merlin；以及 NVIDIA HPC SDK，其中包括能够帮助 HPC 开发者调试和优化 A100 代码的编译器、库和工具。

目前，基于 NVIDIA Ampere 架构的 NVIDIA A100 已全面投产并向全球客户交付。



NVIDIA HPC SDK 加速 HPC 应用开发

NVIDIA HPC SDK 是一款用于编程加速计算系统的全套集成 SDK 综合套件, 包含编译器和库。HPC 开发者可通过它来编程整个 HPC 平台, 从 GPU 基础到 CPU 以及整体互联。该款 SDK 的首个通用版本计划于 2020 年夏末发布。

NVIDIA HPC SDK C++ 和 Fortran 编译器是首批同时支持 C++ 17 并行算法和 Fortran 内联函数等标准语言结构自动 GPU 加速的编译器。

NVIDIA HPC SDK 为 GPU 加速应用程序带来了以下新功能和性能优化。

◎ NVIDIA Fortran、C 和 C++ 编译器可以使用 GPU 加速数学库, 事实标准 OpenACC 和 OpenMP 并行指令, 以及 CUDA 实现异构 CPU + GPU 服务器并行应用程序的移植、编写和调试。

◎ NVIDIA 数学库为 HPC 基础线性代数和信号处理算法提供了高度优化的嵌入式 GPU 加速。除了为 GPU 加

速提供方便的入口匝道之外, 数学库还为支持的例程提供了更高的性能, 并让用户能够自动获取包括最新 A100 GPU 在内的新型 GPU 架构的支持。

◎ 全新 NVSHMEM 库提供了可以从主机或设备启动, 并可以在 CUDA 内核和 OpenACC 程序中调用的多 GPU 和多节点通信原语。NVSHMEM 为 GPU 内存实施 OpenSHMEM 标准, 并可以通过扩展提高 GPU 的性能。

◎ NVIDIA Nsight 计算和 Nsight 系统是适用于 OpenMP 和 OpenACC 中多核 CPU 应用程序以及 OpenACC 和 CUDA 中 GPU 加速应用程序的性能分析工具 (目前, 并非每个平台都支持所有功能)。

◎ NVIDIA 调试器可以以 OpenACC、OpenMP 和 MPI 高级别第三方调试器的基础, 调试 CUDA 应用程序。

◎ HPC 容器制造商简化了构建和维护 Docker 和 Singularity 容器的过程。

NVIDIA EGX 边缘 AI 平台将实时 AI 带入制造、零售、电信和医疗等行业

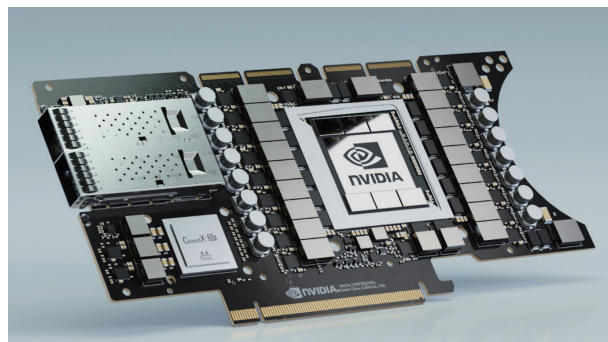
在边缘计算领域, NVIDIA 发布了两款性能强大的 EGX 边缘 AI 平台产品: 适用于大型商业通用服务器上的 EGX A100, 和适用于微型边缘服务器的微型 EGX Jetson Xavier NX。这两款产品能够在边缘端提供安全、高性能的 AI 处理能力。

借助于 NVIDIA EGX™ 边缘 AI 平台, 医院、商店、农场和工厂可以实时处理和保护来自数万亿个边缘传感器的海量数据流。该平台实现了服务器群的远程安全部署、管理和更新。

EGX A100 聚合加速器和 EGX Jetson Xavier NX 微型边缘服务器分别被设计用于满足不同尺寸、成本和性能需求。例如搭载了 EGX A100 的服务器可以管理机场中的数百台摄像头, 而 EGX Jetson Xavier NX 则可用于管理便利店中的少量摄像头。而云原生技术支持能够确保整个 EGX 产品家族都可以使用同一经过优化的 AI 软件轻松构建和部

署 AI 应用。

由 NVIDIA Ampere 架构赋能的 EGX A100



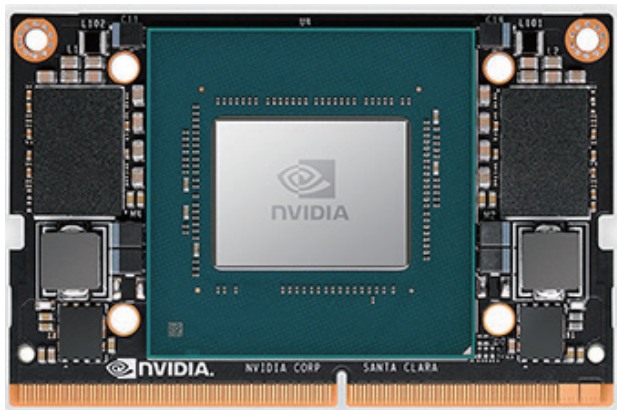
EGX A100 是首个基于 NVIDIA Ampere 架构的边缘 AI 产品。随着 AI 技术的关注点日益向边缘发展, 企业机构可将 EGX A100 添加到其服务器中, 从而实时处理和保护来自边缘传感器的海量流式数据。

EGX A100 结合了 NVIDIA Ampere 架构所具有的突破性计算性能与 NVIDIA Mellanox ConnectX-6 Dx SmartNIC 所具备的网络加速和关键性安全功能, 能将标准型和专用型的边缘服务器转变为极具安全性的云原生 AI 超级计算机。

NVIDIA Ampere 架构是 NVIDIA 的第 8 代 GPU 架构, 为在边缘运行 AI 推理和 5G 应用等各类计算密集型工作负载提供了有史以来最大的性能飞跃, 使 EGX A100 可以实时处理来自摄像头和其他物联网传感器的大量流式数据, 从而更快地获得洞见并提高业务效率。

借助 NVIDIA Mellanox ConnectX-6 Dx 板载网卡, EGX A100 可以每秒接收高达 200 Gb 的数据并将其直接发送到 GPU 内存以进行 AI 或 5G 信号处理。EGX A100 作为一个云原生软件定义加速器, 可以处理对延迟最敏感的 5G 应用, 为单位个行动点 (如: 商店、医院和工厂车间) 的智能实时决策提供了高性能的 AI 和 5G 平台。

体积小, 但性能强大的 EGX Jetson Xavier NX



EGX Jetson Xavier NX 是一款体积小、性能强大的 AI 超级计算机, 适用于微型服务器和边缘智能物联网盒。目前, 其生态圈内的合作伙伴已提供了超过 20 种解决方案。该产品将 NVIDIA Xavier SoC 的性能整合到了一个信用卡大小的模块中。运行 EGX 云原生软件堆栈的 EGX Jetson Xavier NX

可以快速处理来自多个高分辨率传感器的流式数据。

该高效节能的模块能够在 15W 功率下提供最高 21TOPS 的计算力, 或在 10W 功率下提供最高 14TOPS 的计算力。因此, EGX Jetson Xavier NX 为嵌入式边缘计算设备开启了大门——此类设备需要通过提高性能来支持 AI 工作负载, 但同时又受限于尺寸、重量、功耗预算或成本等因素。

经过完整优化的跨 EGX 边缘 AI 平台云原生软件

EGX 边缘 AI 平台的云原生架构使其可以运行容器化的软件, 以支持一系列 GPU 加速的工作负载。

NVIDIA 应用框架包括了用于医疗领域的 Clara、用于电信领域的 Aerial、用于对话式 AI 领域的 Jarvis、用于机器人技术领域的 Isaac, 以及用于零售、智慧城市和交通等领域的 Metropolis。它们可以组合使用或单独使用, 为各种边缘应用场景开辟了新的可能性。

目前整个 NVIDIA EGX 产品家族均支持云原生技术, 因此智能机器制造商和 AI 应用开发者们可以在嵌入式及边缘设备上构建和部署针对机器人技术、智慧城市、医疗以及工业物联网等领域先进的软件定义功能。

来自全球 EGX 生态圈的支持

全球企业计算服务供应商如 Atos、Dell Technologies、富士通、技嘉科技、HPE、IBM、浪潮、联想、Quanta/QCT 和 Supermicro 目前均提供通过 NVIDIA EGX 软件启用的边缘服务器。ADLINK 和 Advantech 等大型服务器和物联网系统制造商也提供此类服务器。

Whiteboard Coordinator、Deep Vision AI、IronYun、Malong 和 SAFR by RealNetworks 等软件供应商可以使用这些服务器以及经过优化的应用框架来构建和部署用于医疗、零售、制造和智慧城市等领域的解决方案。

EGX A100 将于 2020 年年底上市。基于 EGX Jetson Xavier NX 的即用型微型边缘服务器现已上市, 可供希望大批量生产边缘系统的公司选购。

NVIDIA 为全球领先的数据分析平台 Apache Spark 提速

NVIDIA 宣布正在与开源社区合作, 为 Apache Spark 3.0 带来端到端的 GPU 加速。Apache Spark 3.0 是一个热门

的分析引擎, 被全球范围内超过 50 万名数据科学家用于大数据处理。



随着 Spark 3.0 发布的临近，数据科学家和机器学习工程师将能够把革命性的 GPU 加速应用于普遍使用 SQL 数据库操作进行的 ETL（提取、转换和加载）数据处理工作负载中。并且，能够在同一 Spark 集群上处理 AI 模型训练任务，而无需将工作负载作为单独流程在单独的基础架构上运行。这样就可以对整个数据科学的处理流程进行高性能数据分析，对从数据湖到模型训练所涉及的数十、乃至数千 TB 的数据进行加速，而且无需对已被应用于本地和云端 Spark 应用程序的现有代码进行修改。

NVIDIA 企业计算主管 Manuvir Das 表示：“数据科学是当今企业和研究者所面临最大的高性能计算挑战。从 ETL 到训练再到推理，整个 Spark 3.0 方案的原生 GPU 加速为用户提供了最终将大数据潜力与 AI 性能相连所需的性能和规模。”

NVIDIA GPU 加速 Spark 中的 ETL 和数据传输

NVIDIA 正在提供一种用于 Apache Spark 的全新开源 RAPIDS 加速器，帮助数据科学家提高其数据流程的端到端性能。该加速器能够将之前需要在 CPU 上运行的功能，转到 GPU 上运行，以发挥 GPU 的性能。

◎ 在无需更改任何代码的情况下显著提高 Spark SQL 和 DataFrame 的运行性能，从而加速 Spark 中的 ETL 数据流程。

◎ 无需独立的机器学习和深度学习集群，即可在相同的基础架构上加速数据准备和模型训练。

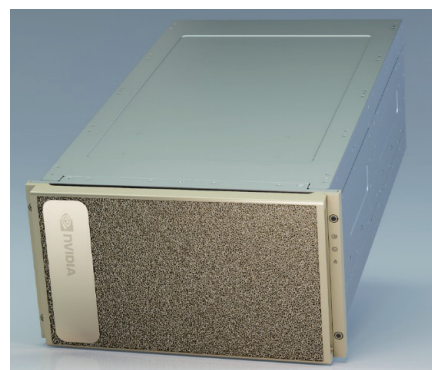
◎ 加速 Spark 分布式集群中各节点之间的数据传输性能。这些资源库使用 UCF Consortium 的开源 UCX（Unified Communication X）框架，并通过使数据直接在 GPU 内存之间移动来最大程度地减少延迟。

NVIDIA 先进的 AI 系统 NVIDIA DGX A100 助力世界对抗 COVID-19

NVIDIA 推出的 NVIDIA DGX A100 是先进的第三代 AI 系统，具有高达 5Petaflops 的 AI 性能，并且首次将整个数据中心的性能和功能集成到一个灵活的平台中。

DGX A100 系统已开始向全球供应。首批订单将送往美国能源部（U.S. Department of Energy, DOE）的阿贡国家实验室（Argonne National Laboratory）。该实验室将运用该集群的 AI 和计算力来更好地研究和应对 COVID-19。

NVIDIA 创始人兼首席执行官黄仁勋表示：“NVIDIA DGX A100 是专为先进 AI 打造的高性能系统。NVIDIA DGX 是首个为端到端机器学习工作流——从数据分析到训练再到推理构建的 AI 系统。借助于全新 DGX 的巨大性能



飞跃，面对呈指数级速度增长的 AI 模型和数据，机器学习工程师们能够保持领先。”

NVIDIA 扩展 NVIDIA Clara，联合全球医疗合作伙伴抗击 COVID-19

在医疗方面，NVIDIA 宣布了 NVIDIA Clara 医疗平台的一些重大拓展，包括性能的提升、主要合作伙伴关系和新的功能，旨在帮助医学界更好地追踪、测试和治疗 COVID-19。

该平台集 AI 和加速计算于一身，可以帮助医疗研究人员、技术解决方案供应商和医院以全新的方式更快地应对疫情。

◎ 打破基因组测序速度记录：为 COVID-19 研究人员提供 90 天 NVIDIA Clara Parabricks 免费使用权限。该软件创造了在 20 分钟内完成整个人类基因组 DNA 序列分析的记录。

◎ 疾病检测 AI 模型：与美国国家卫生研究院（National Institutes of Health）共同开发的 AI 模型已经上线。该模型可帮助研究人员通过胸部 CT 扫描研究 COVID-19 的严重程度，并开发新的工具来更好地研究、评估病毒，并进行感染检测。

◎ 专为智能医院开发的 NVIDIA Clara Guardian：NVIDIA Clara Guardian 采用智能视频分析和自动语音识别技术，使新一代智能医院能够在尽量避免人员接触的同时，监测病人的生命体征。

NVIDIA 医疗业务副总裁 Kimberly Powell 表示：“COVID-19 疫情促进了科技、研究和医疗行业的进一步合作，共同开发新的计算解决方案，从而加快了解该疾病的传播方式、规模和严重性。人类需要将最先进的 AI 和加速计算技术应用于医疗行业的各个领域当中，并将在未来的整个医疗行业中广泛地感受到其所产生的影响。”

全新 AI 模型帮助检测 COVID-19

NVIDIA 还为研究人员提供了另一组用于对抗 COVID-19 的 AI 模型可以帮助研究人员通过胸部 CT 扫描数据检测和研究受感染患者的病情。

这组模型已在 NGC 软件中心的最新版 Clara 影像中上线。这组模型由 NVIDIA 应用研究团队与美国国家卫生研究院（NIH）的临床医生和数据科学家共同合作开发，其所采用的数据来自中国、意大利、日本和美国等 COVID-19 高感染率地区。

构建该组 AI 模型，使用到了用于医学影像的 NVIDIA Clara 应用框架。NVIDIA Clara 包含有相关工具，可以用于特定领域的 AI 训练和部署，这使得 NVIDIA 和美国国家卫生研究院（NIH）得以在不到三周的时间内就开发出这组模型。

目前，NVIDIA 和 Mass Brigham 正在领导着一项新的 COVID-19 跨国联合研究计划，该计划将新冠病毒 AI 模型扩展到 X 光影像，这样就可以在不共享任何患者数据和保护患者隐私的情况下在本地使用这组模型。

NVIDIA Clara Guardian 为一线医院提供帮助

全新的 NVIDIA Clara Guardian 应用框架能通过将普通传感器转换为智能传感器，为医院提供其所急需的公共安全和患者监控 AI 解决方案生态系统。其主要用例包括：自动体温筛查、防护口罩检测、安全性社交隔离和远程患者监测。

该生态系统中的合作伙伴正在使用预先训练的模型和迁移学习技术开发和部署 AI 应用程序，这些 AI 应用程序集视觉、语音和自然语言处理于一身。借助 NVIDIA EGX 边缘 AI 平台，解决方案供应商可以在整个医院环境中安全地部署和管理大量设备。

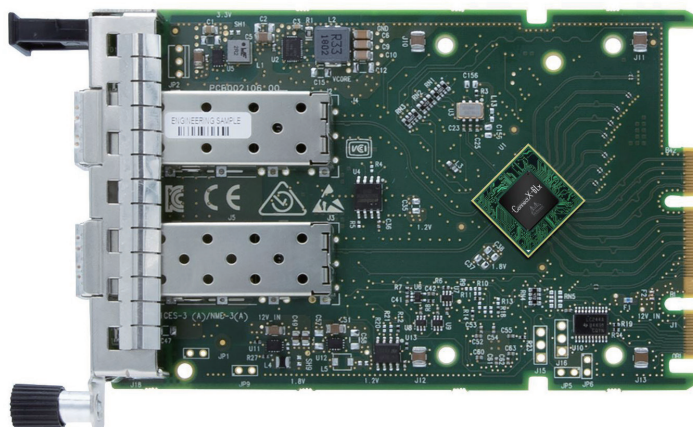
目前，AnyVision、BriefCam、Care.ai、Chooch AI、Deep Vision、Diycam、IntelliSite、码隆科技、Ouva、SafelyYou、SAFR、SmartCow、TeiaCare、Tonbo Imaging 和 Whiteboard Coordinator 等数十家解决方案合作伙伴已

NVIDIA 推出 25G 安全智能网卡（SmartNIC）

在网络连接方面，NVIDIA 推出了一款高度安全且高效的 25/50Gb/s 以太智能网卡（Smart Network Interface Controller，SmartNIC）——NVIDIA Mellanox ConnectX-6 Lx SmartNIC，以满足飞速增长的大

规模企业和云业务的需求。

ConnectX-6 Lx 是 ConnectX 系列网卡的第 11 代产品。这款产品专为满足现代化数据中心的需求而设计，25Gb/s 网络正在成为各种主流应用的标准，例如企业级应用、AI



和实时分析等。

这款全新的 SmartNIC 充分利用了软件定义和硬件加速等先进技术，将更多的安全和网络处理工作从 CPU 卸载到网卡上。

NVIDIA 的 Mellanox 网络事业部软件和 NIC 产品高级副总裁 Amit Krig 表示：“现代云和企业数据中心需要加速网络来应对扩展性和计算方面的挑战。ConnectX-6 Lx SmartNIC 就是面向新一代解耦式的、可重组的数据中心的经济高效、可加速应用的和安全的智能网卡。”

包括 IPsec 在线加密和硬件信任根（Hardware Root of Trust）等安全加速功能，可提升 10 倍性能的连接追踪（Connection Tracking）功能等，实现了整个数据中心的在零信任下的安全保障（Zero Trust security）。领先的 RoCE（RDMA over converged Ethernet）技术、先进的虚拟化和容器化技术，以及 NVMe over Fabrics 存储卸载技术等，将成为高性能可扩展网络的保障。

Mellanox ConnectX-6 Lx 提供：两个 25Gb/s 端口或一个 50Gb/s 端口、从主机到网络之间采用 PCIe Gen 3.0 或 4.0 x8。

同时 Mellanox ConnectX-6 Lx 的安全功能包括硬件信任根（Hardware Root of Trust）、用于有状态 L4 防火墙的连接追踪（Connection Tracking）技术，IPSec 在线加密加速等。

GPU Direct RDMA 加速技术，专为 NVMe over Fabrics（NVMe-oF）存储、可扩展加速计算和高速视频传输等应用加速。

Zero Touch RoCE（ZTR）功能，可以实现了高性能 RoCE 网络的可扩展和易部署，且无需配置交换机。

加速交换和数据包处理（Accelerated switching and packet processing，ASAP2）技术，内置了面向虚拟化和容器化应用的 SR-IOV 和 VirtIO 硬件卸载引擎，可加速新一代软件定义网络，以及对于防火墙服务进行连接追踪。

目前，ConnectX-6 Lx 处于样品阶段，预计将在 2020 年第三季度全面上市。与 ConnectX 系列的所有产品一样，该产品兼容所有 Mellanox SmartNIC 软件协议栈。ConnectX SmartNIC 与 Mellanox Spectrum 系列交换机和 LinkX 系列电缆和收发器一起，可为高性能网络提供最完整的端到端解决方案。

『宝马集团采用 NVIDIA 重新定义工厂物流管理』

2020 年 5 月 14 日，NVIDIA 与宝马集团共同宣布，宝马集团已采用全新 NVIDIA Isaac 机器人平台对其汽车工厂进行优化，利用先进的 AI 计算与可视化技术为其打造物流

机器人。

此次合作重点围绕实施基于 NVIDIA 技术的端到端系统展开，采用基于软件架构开发、运行于 NVIDIA 开放式



Isaac 机器人软件平台的机器人，完成从训练、测试到部署的全过程。宝马集团希望借此合作优化物流工厂的流程，从而更快捷、更高效地生产定制化配置的汽车。开发完成后，该系统将在宝马集团全球范围内的工厂进行部署。

“宝马集团采用 NVIDIA Isaac 机器人平台重塑其工厂物流管理流程是一次具有革命性意义的举动。” NVIDIA 创始人兼首席执行官黄仁勋表示，“宝马集团引领着机器人工厂时代，利用在人工智能和机器人领域取得的突破进展，创造出下一个高度可定制、及时且有序的制造水平。”

宝马集团物流高级副总裁 Jürgen Maidl 表示：“宝马秉承着将选择权交给客户的原则，为不同客户提供针对各款车型，多样化功能的量身定制产品。要想在一条生产线上，生产更高质量、高度定制化的各款车型，需要端到端的先进计算解决方案。与 NVIDIA 的合作使我们又更强的力量推动工厂物流技术的未来发展，并最终提升宝马集团的全球客户的满意度。”

此次合作将采用 NVIDIA DGX AI 系统和 Isaac 仿真技术对机器人进行训练和测试；采用 NVIDIA 光线追踪 GPU 对合成机器零件进行渲染，以此强化训练；并采用基于高性能 NVIDIA Jetson 和 EGX edge 计算机以及 Isaac 软件开发套件（SDK）构建的一组全新 AI 机器人。

宝马集团供应链庞大，汽车工厂要从 4500 个供应商那里接收数百万个零件，单一零件编号数量高达 23 万。随着宝马集团的汽车销售在过去 10 年中翻了一番，达到 250 万辆，这些数字还会持续增长。此外，宝马集团的车辆平均为

客户提供 100 种不同的定制化选择，意味着 99% 的客户订单都各有差别，这些都给工厂物流提出了巨大挑战。

为了优化这种物流层面极高的复杂性，如今自主 AI 物流机器人能够协助当前流程，在同一条生产线上，组装高度定制化的车辆。

Maidl 表示：“大量定制化配置的需求从根本上来讲，在三个基础层面对宝马集团的生产提出了挑战，即计算、物流规划和数据分析。”

为了应对这些挑战，宝马集团选择与 NVIDIA 合作。使用 NVIDIA 的 Isaac 机器人平台开发了五款具有 AI 功能的机器人，来改进物流工作流程，如用于自动运输物料的导航机器人、用于选择和整理零件的操控机器人，这些机器人都采用 NVIDIA Jetson AGX Xavier 和 EGX 边缘计算机。

基于 NVIDIA Isaac SDK 开发的机器人，利用大量强大的深度神经网络，可以处理感知、分割、姿势预估和人体姿势预估等计算任务，以完成感知环境、检测物体、自动导航和移动物体。这些机器人在真实数据和合成数据上进行训练，并使用 NVIDIA GPU 在各种光线和遮挡条件下渲染光线追踪机器部件，以增强真实数据。然后，在 NVIDIA DGX 系统上，基于实际数据和合成数据来训练深度神经网络。接下来，这些机器人在 NVIDIA 的 Isaac 模拟器上进行导航和操控的连续测试，并运行于 NVIDIA Omniverse 虚拟环境中，在该环境中，多位来自不同地区的宝马集团员工都能够同一个仿真环境中展开工作。IM