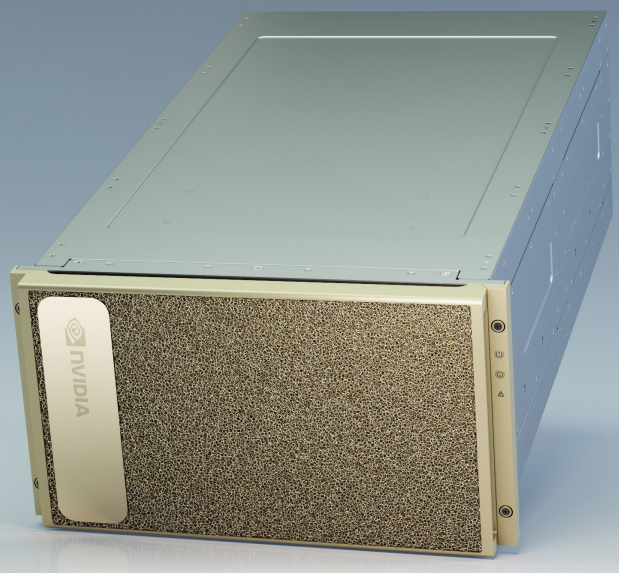
NVIDIA DGX A100

通用的 AI 基础架构系统



扩展企业 AI 的挑战

每家企业都需要利用人工智能 (AI) 实现转型，以在这个充满挑战的时代求得生存，继而实现蓬勃发展。但长期以来，传统方法所采用的计算架构较为缓慢，而且总是分开处理分析、训练和推理工作负载，所以企业需要一种适用于 AI 基础架构的平台对此加以改进。传统方法不仅复杂、成本高、扩展速度受限，而且对现代 AI 束手无策。 因此，企业、开发者、数据科学家和研究人员都需要一个新平台，以便统一处理所有 AI 工作负载、简化基础架构以及提高投资回报率 (ROI)。

适用于各种 AI 工作负载的通用系统

作为一种通用系统，NVIDIA DGX™ A100 可处理各种 AI 工作负载，包括分析、训练和推理。DGX A100 设立了全新计算密度标准，不仅在 6U 外形尺寸下封装了 5 petaFLOPS 的 AI 性能，而且用单个统一系统取代了传统的计算基础架构。此外，凭借 NVIDIA A100 Tensor Core GPU 中的多实例 GPU (MIG) 功能，DGX A100 首次实现了强大算力的精细分配，使得管理员可针对特定工作负载分配大小合适的资源。DGX A100 提供了多达 640GB 的 GPU 显存总量，可将大型训练作业的性能提升高达三倍，并将 MIG 实例的大小翻倍，因而能够处理极为复杂的大型作业，同时轻松应对简单、小型的任务。DGX A100 利用 NGC 中经过优化的软件运行 DGX 软件堆栈，同时将密集计算能力和全面的工作负载灵活性相结合，因此成为单节点部署以及通过 NVIDIA DeepOps 部署大规模 Slurm 和 Kubernetes 集群的理想之选。

直接获得 NVIDIA DGXpert 团队支持

NVIDIA DGX A100 不仅仅是一台服务器，更是一个完整的软硬件平台。它基于巨大的 DGX 试验基地——NVIDIA DGX SATURNV——所积累的知识经验搭建而成，背后更有数千名 NVIDIA DGXpert 的支持。作为精通 AI 的从业者，DGXpert 会提供规范指导和设计专业知识，帮助推动 AI 转型。他们在过去十年中积累了丰富的专业知识和经验，可帮助您从 DGX 投资中尽可能获得更大价值。DGXpert 有助于确保关键应用快速启动并保持平稳运行，从而大幅缩短获得见解的时间。

系统规格

	NVIDIA DGX A100 640GB	NVIDIA DGX A100 320GB
GPU 个数	8 个 NVIDIA A100 80GB GPU	8 个 NVIDIA A100 40 GB GPU
GPU 显存	共 640GB	共 320GB
性能	5 petaFLOPS AI 10 petaOPS INT8	
NVIDIA NVSwitch	6	
系统功率	最大 6.5 千瓦	
CPU	两个 AMD Rome 7742、共 128 个核心、2.25 GHz (基准频率)、3.4 GHz (最大加速频率)	
系统内存	2TB	1TB
网络	8 个单端口 Mellanox ConnectX-6 VPI 200Gb/s HDR InfiniBand 两个双端口 Mellanox ConnectX-6 VPI 10/25/50/100/ 200Gb/s 以太网	8 个单端口 Mellanox ConnectX-6 VPI 200Gb/s HDR InfiniBand 1 个双端口 Mellanox ConnectX-6 VPI 10/25/50/100/ 200Gb/s 以太网
存储	操作系统： 两个 1.92TB M.2 NVMe 驱动器 内部存储： 30TB (8 个 3.84TB) U.2 NVMe 驱动器	操作系统： 两个 1.92TB M.2 NVMe 驱动器 内部存储： 15TB (4 个 3.84TB) U.2 NVMe 驱动器
软件	Ubuntu Linux 操作系统 同时支持： Red Hat Enterprise Linux CentOS	
系统重量	最大 123.16 千克	
包装后的系统重量	最大 163.16 千克	
系统尺寸	高度：264.0 毫米 宽度：最大 482.3 毫米 长度：最大 897.1 毫米	
运行温度范围	5-30 °C	

更快解决问题

NVIDIA DGX A100 配备八个 NVIDIA A100 Tensor Core GPU，可实现出色的加速性能，且已针对 NVIDIA CUDA-X™ 软件和端到端 NVIDIA 数据中心解决方案堆栈进行全面优化。NVIDIA A100 GPU 引入了全新精度 Tensor Float 32 (TF32)，该精度与 FP32 的原理类似，但与前一代相比，可面向 AI 提供高达 20 倍的每秒浮点运算次数 (FLOPS)。最重要的是，实现此类加速无需更改任何代码。此外，在使用含 FP16 的 NVIDIA 自动混合精度时，A100 仅增加一行代码，即可将性能再提升两倍。

A100 80GB GPU 将高带宽显存从 40GB (HBM) 增加一倍至 80GB (HBM2e)，其 GPU 显存带宽超过 2TB/s，比 A100 40GB GPU 增加了 30%，亦达到全球领先水平。DGX A100 还推出了速度比上一代产品高出一倍的全新 NVIDIA NVSwitch™ 以及第三代 NVIDIA® NVLink® 技术，后者可将 GPU 间的直连带宽增加一倍至 600GB/s，几乎相当于 PCIe 4.0 的十倍。这一强大功能可大幅缩短问题解决时间，让用户能够应对此前无法解决的难题。

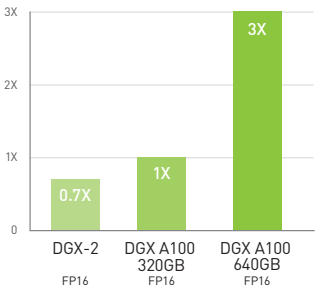
安全性更高的企业 AI 系统

NVIDIA DGX A100 采用多层级架构保护所有主要的软硬件组件，确保 AI 企业处于稳定的安全状态。DGX A100 内置安全机制，覆盖基板管理控制器 (BMC)、CPU 载板、GPU 载板、自加密驱动和安全启动，可帮助 IT 人员专注于 AI 操作，而不必花时间评估和应对安全威胁。

通过 NVIDIA Mellanox 实现卓越的数据中心可扩展性

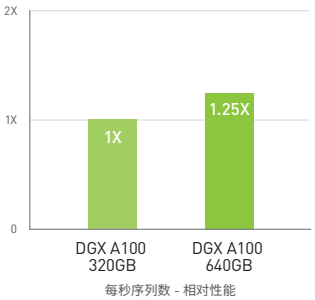
NVIDIA DGX A100 配备所有 DGX 系统中速度领先的 I/O 架构，是 NVIDIA DGX SuperPOD™ 等大型 AI 集群的基础构件，而后者为可扩展的 AI 基础架构描绘了企业蓝图。DGX A100 具有八个用于实现集群的单端口 NVIDIA Mellanox® ConnectX®-6 VPI HDR InfiniBand 适配器，以及多达两个用于存储和网络连接的双端口 ConnectX-6 VPI 以太网适配器，二者的速度均可达到 200Gb/s。借助大规模 GPU 加速计算与精尖网络硬件和软件优化的强强联合，DGX A100 可扩展至数百乃至数千个节点，从而攻克对话式 AI 和大规模图像分类等更艰巨的挑战。

针对大型模型提供高达 3 倍的 AI 训练速度



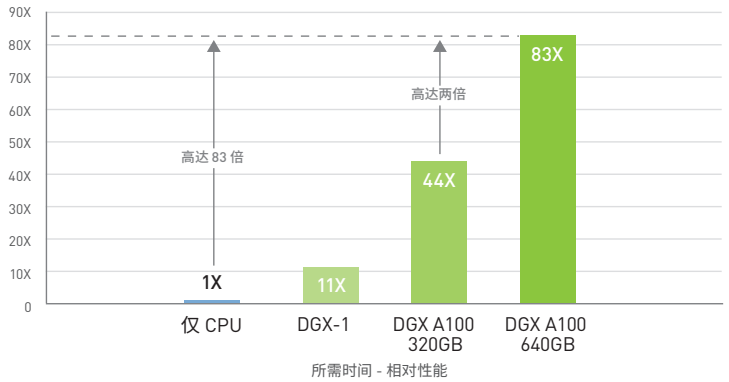
HugeCTR 框架上的 DLRM，精度 = FP16 | 1 个 DGX A100 640GB 批量大小 = 48 | 2 个 DGX A100 320GB 批量大小 = 32 | 1 个 DGX-2 (16 个 V100 32GB) 批量大小 = 32，加速性能通过 GPU 数量实现标准化。

AI 推理吞吐量提升至 1.25 倍



通过 (1/7) MIG 切片测量的 MLPerf 0.7 RNN-T，框架：TensorRT 7.2，数据集 = LibriSpeech，精度 = FP16。

相较于 CPU 速度提升至 83 倍；相较于 DGX A100 320GB，在大数据分析基准测试方面速度提升至两倍



大数据分析基准测试 | 针对 10TB 数据集的 30 次分析零售查询、ETL、ML、NLP | CPU：19 个英特尔至强金牌 6252 2.10 GHz，Hadoop | 16 个 DGX-1 (每个含 8 个 V100 32GB)，RAPIDS/Dask | 12 个 DGX A100 320GB 和 6 个 DGX A100 640GB，RAPIDS/Dask/BlazingSQL，加速性能通过 GPU 数量实现标准化。

携手值得信赖的数据中心领军者，共同打造成熟的基础架构解决方案

通过与领先的存储和网络技术提供商合作，我们提供了一套基础架构解决方案组合，其中融合了 NVIDIA DGX POD™ 参考架构的诸多优点。这些解决方案在我们的 NVIDIA 合作伙伴网络 (NPN) 中作为完全集成且可随时部署的产品提供，旨在简化并加快数据中心 AI 部署。

如需了解有关 NVIDIA DGX A100 的详细信息，请访问 **https://www.nvidia.cn/data-center/dgx-a100/**