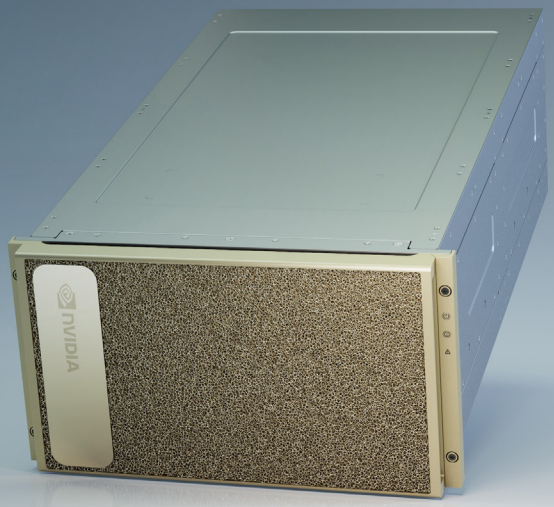
数据表



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

NVIDIA DGX™ A100 是适用于所有 AI 工作负载，包括分析、训练、推理的通用系统。DGX A100 设立了全新计算密度标准，不仅在 6U 外形规格下封装了 5 Petaflop 的 AI 性能，而且用单个统一系统取代了传统的计算基础设施。此外，DGX A100 首次实现了强大算力的精细分配。利用 NVIDIA A100 Tensor Core GPU 中的多实例 GPU (MIG) 功能，管理员可针对特定工作负载分配大小合适的资源。

DGX A100 具有高达 640GB 的总 GPU 显存，可将大规模训练作业的性能提升高达 3 倍，并将 MIG 实例的大小增加一倍，从而从容应对颇为复杂的大任务，以及简单轻松的小任务。DGX A100 运行集成 NVIDIA NGC™ 优化软件的 DGX 软件堆栈，兼具密集算力与全面的工作负载灵活性，因而非常适合处理单节点部署以及使用 NVIDIA Bright Cluster Manager 部署的大规模 Slurm 和 Kubernetes 集群。

高水平的支持能力和专业知识

NVIDIA DGX A100 不仅仅是一台服务器，更是一个完整的软硬件平台。它基于全球最大的 DGX 集群 NVIDIA DGX SATURNV 积累的知识经验而建立，背后有 NVIDIA 数千名 DGXpert 支持。DGXpert 是一个拥有众多 AI 从业者的团队，团队成员在过去十年间积累了丰富的专业知识和经验，可帮助您更大限度地提升 DGX 投资价值。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	
GPU	**8 个 NVIDIA A100 80GB Tensor Core GPU**	
GPU 显存	**共 640GB**	
性能	**5 petaFLOPS AI** **10 petaOPS INT8**	
NVIDIA NVSwitch	**6**	
系统功耗	**最大 6.5 千瓦**	
CPU	**双路 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**	
网络	**8 个单端口 NVIDIA ConnectX-7** **200Gb/s 的 InfiniBand 端口** **2 个双端口 NVIDIA ConnectX-7 VPI** **10/25/50/100/200Gb/s 以太网**	**8 个单端口 NVIDIA ConnectX-6 VPI** **200Gb/s 的 InfiniBand 端口** **2 个双端口 NVIDIA ConnectX-6 VPI** **10/25/50/100/200Gb/s 以太网**
存储	**操作系统：2 个 1.92TB M.2 NVME 驱动器** **内部存储：30TB (8 个 3.84TB) U.2 NVMe 驱动器**	
软件	**Ubuntu Linux 操作系统** **同时支持：** **Red Hat 企业级 Linux** **CentOS**	
系统重量	**最大 123.16 千克**	
包装后系统重量	**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°C 至 30°C**	

更快解决问题

NVIDIA DGX A100 配备 8 个 NVIDIA A100 Tensor Core GPU，可出色完成加速任务，并针对 NVIDIA CUDA-X™ 软件和整套端到端 NVIDIA 数据中心解决方案进行全面优化。NVIDIA A100 GPU 引入 Tensor Float 32 (TF32) 精度，即 TensorFlow 和 PyTorch AI 框架的默认精度格式。TF32 的工作原理与 FP32 类似，但相较于上一代产品，TF32 可提供高达 20 倍的 AI 每秒浮点运算 (FLOPS) 性能。而最重要的是，实现此类加速无需改动任何代码。

A100 80GB GPU 的 GPU 显存带宽比 A100 40GB GPU 增加了 30%，以每秒超过 2 万亿字节的速度 (2TB/s) 达到全球领先水平。此外，与上一代 NVIDIA GPU 相比，A100 GPU 具有超大片内内存，包括 40 MB 的二级缓存，扩大近 7 倍，可更大限度地提升计算性能。DGX A100 还首次推出第三代 NVIDIA® NVLink®，使 GPU 到 GPU 直接带宽提高一倍，直逼每秒 600 千兆字节 (GB/s)，几乎比 PCIe 4.0 高 10 倍。此外，新款 NVIDIA NVSwitch™ 的速度是上一代的 2 倍。这种强大的性能可助力用户更快解决问题，以及应对此前无法解决的难题。

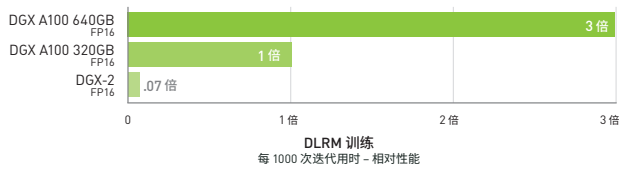
安全性更高的企业 AI 系统

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 载板和自加密驱动，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。

借助 NVIDIA 网络实现数据中心的非凡可扩展性

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（后者为可扩展的 AI 基础设施描绘了企业蓝图）。DGX A100 拥有 8 个用于集群的单端口 NVIDIA ConnectX®-7 InfiniBand 网卡，以及最高 2 个用于存储和网络连接的双端口 ConnectX-7 VPI (InfiniBand 或以太网) 网卡，二者的速度均能达到 200 Gb/s。将 ConnectX-7 与 NVIDIA Quantum-2 InfiniBand 交换机相连，即可用更少的交换机和线缆构建 DGX SuperPOD，从而节省数据中心基础设施的 CAPEX 和 OPEX。借助海量 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 训练吞吐量

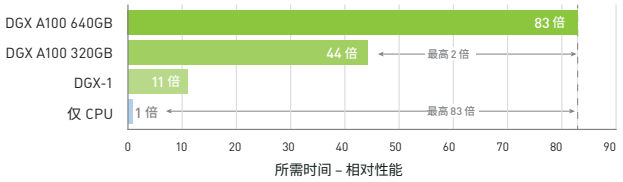


HugeCTR 框架上的 DLRM，精度 = FP16 | 1x DGX A100 640GB 批量大小 = 48 | 2x DGX A100 320GB 批量大小 = 32 | 1x DGX-2 (16x V100 32GB) 批量大小 = 32。加速性能通过 GPU 数量实现标准化。

AI 推理吞吐量提升高达 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 吞吐量提升高达 83 倍；相较于 DGX A100 320GB，在大数据分析基准测试方面吞吐量提升高达两倍



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: 19x 英特尔至强金牌 6252 2.10 GHz，Hadoop | 16x DGX-1 (每个含 8 个 V100 32GB)，RAPIDS/Dask | 12x DGX A100 320GB 和 6x 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。借助 NVIDIA 合作伙伴网络 (NPN)，我们将提供全面集成、可立即部署的解决方案，帮助企业更轻松、更快速地部署数据中心 AI。

了解详情

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/**
如需了解 NVIDIA 技术在电信行业的热门用例，请访问 **https://www.nvidia.cn/industries/telecommunications/**

NVIDIA