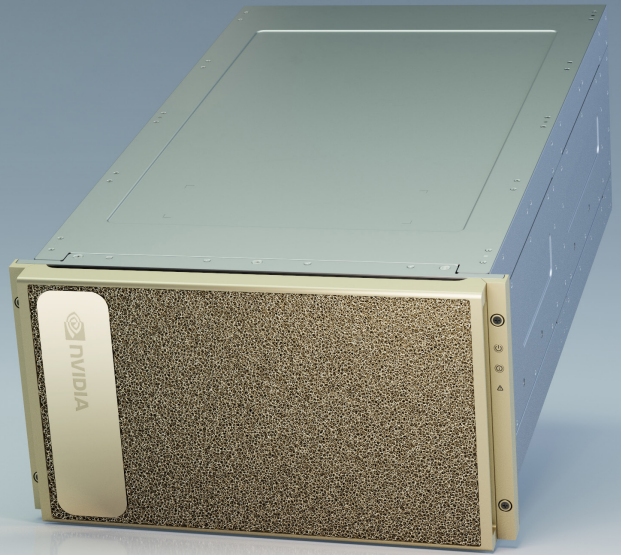




## NVIDIA DGX A100 通用的 AI 基础架构系统



### 扩展企业 AI 的挑战

在这个充满挑战的时代，每家企业都需利用人工智能 (AI) 实现转型，从而生存下来并蓬勃发展。然而，传统方法所采用的计算架构较为缓慢，而且总是分开处理数据分析、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。传统架构复杂、成本高、限制了规模化增长的速度，没有为现代 AI 做好准备。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)。

### 适用于所有 AI 工作负载的通用系统

NVIDIA DGX™ A100 通用系统可处理各种 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，在 6U 外形尺寸下封装了 5 petaFLOPS 的 AI 性能，用单个统一系统取代了传统的计算基础架构。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 功能，管理员可针对特定工作负载分配大小合适的资源，确保能从容应对颇为复杂的大任务，以及简单轻松的小任务。运行 NGC 上优化过的 DGX 软件堆栈，结合密集算力和完整的工作负载灵活性，让 DGX A100 成为适用于单节点部署以及部署了 NVIDIA DeepOps 的大规模 Slurm 和 Kubernetes 集群的理想之选。

### 直接获取 NVIDIA DGXpert 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。作为经验丰富的 AI 从业者，DGXpert 可提供规范性指导和项目设计专长，进而帮助推动 AI 转型。在过去十年间，他们积累了丰富的专业知识和经验，可帮助您更大幅度地提升 DGX 投资价值。DGXpert 有助于确保关键应用快速启动、运行并保持流畅运转，从而大幅缩减获得见解的时间。

#### 系统规格

GPU	8 块 NVIDIA A100 Tensor Core GPU
GPU 显存	共 320 GB
性能	5 petaFLOPS AI 10 petaOPS INT8
NVIDIA NVSwitch	6
系统功耗	最高 6.5 kW
CPU	两块 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	1TB
网络	8 个单端口 Mellanox ConnectX-6 VPI 200 Gb/s 的 HDR InfiniBand 1 个双端口 Mellanox ConnectX-6 VPI 10/25/50/100/200 Gb/s 以太网
存储	操作系统：2 个 1.92TB M.2 NVME 驱动器 内部存储：15 TB (4x 3.84TB) U.2 NVME 驱动器
软件	Ubuntu Linux 操作系统
系统重量	123 千克
包装后的系统重量	143 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 块 NVIDIA A100 Tensor Core GPU，可帮助用户出色地完成加速任务，同时也针对 NVIDIA CUDA-X™ 软件和端到端 NVIDIA 数据中心解决方案堆栈进行了全面优化。NVIDIA A100 GPU 实现了与 FP32 原理相同的全新精度级别 TF32，相较于上一代产品，可提供高达 20 倍 FLOPS 的 AI 性能。而最重要的是，实现此类加速无需改动任何代码。通过 NVIDIA 自动混合精度功能，只需要增加一行代码 A100 就可以提供额外两倍的 FP16 精度性能的提升。同时，A100 GPU 拥有世界领先的显存带宽 (1.6 TB/s)，与上一代产品相比，增幅超过 70%。另外，A100 GPU 有超大片上内存，包括 40 MB 的二级缓存，比上一代产品大近 7 倍，可更大限度地提升计算性能。DGX A100 还推出速度为上一代 2 倍的全新 NVIDIA NVSwitch 和新一代 NVIDIA NVLink™ 技术，后者可将 GPU 之间的直连带宽增加一倍，从而达到 600 GB/s，而这几乎是 PCIe Gen 4 的 10 倍。这种强大的功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

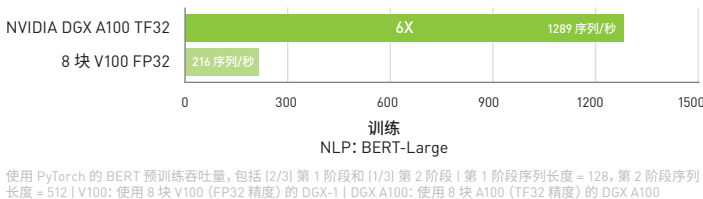
NVIDIA DGX A100 采用多层级架构来保护所有主要的软硬件组件，确保 AI 企业处于稳定的安全状态。DGX A100 内置安全机制，覆盖基板管理控制器 (BMC)、CPU 载板、GPU 载板、自加密驱动和安全启动，可帮助 IT 人员专注于 AI 操作，而不用花时间评估和应对安全威胁。

## 通过 Mellanox 实现卓越的数据中心可扩展性

NVIDIA DGX A100 配备所有 DGX 产品中最快的 I/O 架构，是 NVIDIA DGX SuperPOD™ 等大型 AI 集群的基石，也为可扩展的 AI 基础架构描绘企业蓝图。DGX A100 拥有 8 个用于集群的单端口 Mellanox ConnectX-6 VPI HDR InfiniBand 适配器，以及 1 个用于存储和网络连接的双端口 ConnectX-6 VPI 以太网适配器，二者的速度均能达到 200 Gb/s。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。

如需详细了解 NVIDIA DGX A100，请访问 [www.nvidia.cn/DGXA100](http://www.nvidia.cn/DGXA100)

### DGX A100 提供 6 倍训练性能



### DGX A100 提供 172 倍推理性能



### DGX A100 提供 13 倍数据分析性能



## 携手数据中心领军者 打造经验证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，我们将提供全面集成、可立即部署的解决方案，帮助 IT 人员更快速、更轻松地部署数据中心 AI。