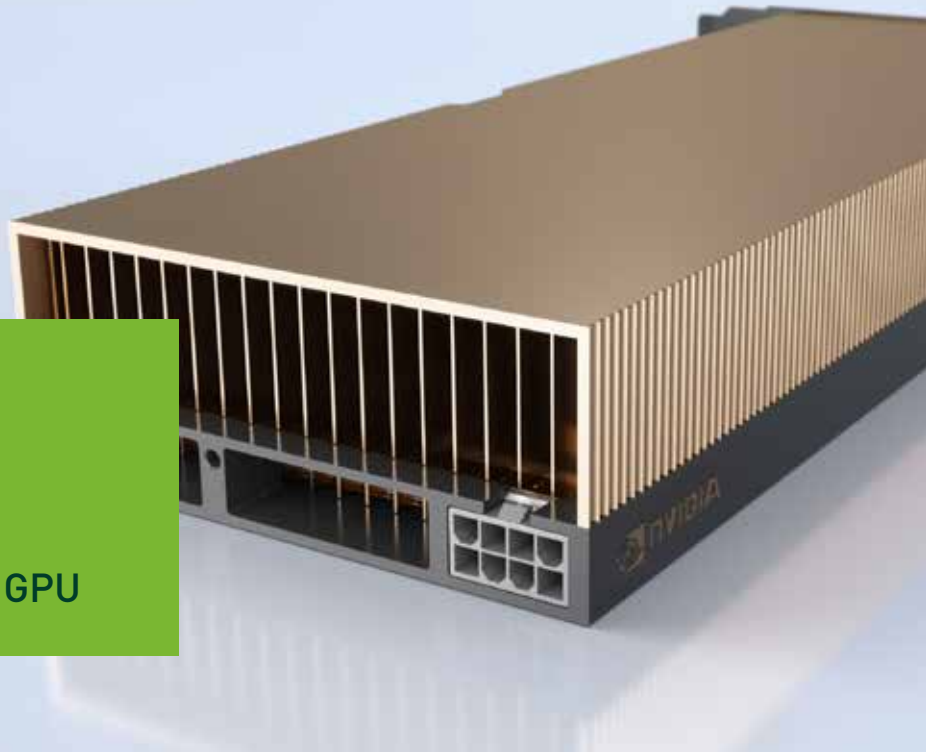
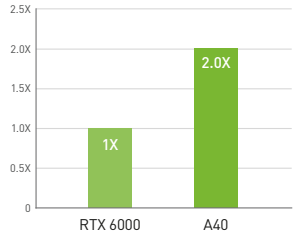


NVIDIA A40
适用于视觉计算的强大数据中心 GPU

NVIDIA A40 可加速数据中心要求严苛的视觉计算工作负载，将最新的 NVIDIA Ampere 架构 RT Core、Tensor Core 和 CUDA® 核心与 48 GB 图形显存相结合。从可以随时随地访问的强大虚拟工作站到专用的渲染节点，NVIDIA A40 将新一代 NVIDIA RTX™ 技术引入数据中心，处理更先进的专业可视化工作负载。

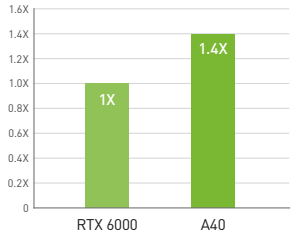
渲染性能提升高达 2 倍[1]

Iray 2020.1



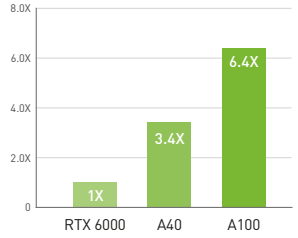
图形性能提升高达 40%[1]

SPECviewperf 2020



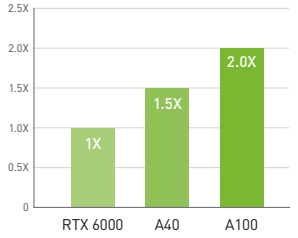
AI 训练性能提升高达 3 倍[2]

BERT 预训练吞吐量



单精度 (FP32) HPC 性能提升高达 50%[2]

NAMD



规格

GPU 架构	**NVIDIA Ampere 架构**
GPU 显存	**带有 ECC 的 48 GB GDDR6**
显存带宽	**696 GB/秒**
互联接口	**NVIDIA® NVLink® 112.5 GB/s (双向)[3] PCIe 4.0 31.5 GB/s (双向)**
基于 NVIDIA Ampere 架构的 CUDA 核心	**10752**
NVIDIA 第二代 RT Core	**84**
NVIDIA 第三代 Tensor Core	**336**
FP32 TFLOPS 峰值 (非 Tensor)	**37.4**
使用 FP16 累加的 FP16 Tensor TFLOPS 峰值	**149.7| 299.4***
TF32 Tensor TFLOPS 峰值	**74.8| 149.6***
RT Core 性能 TFLOPS	**73.1**
使用 FP32 累加的 BF16 Tensor TFLOPS 峰值	**149.7| 299.4***
INT8 Tensor TOPS 峰值 INT 4 Tensor TOPS 峰值	**299.3| 598.6*** **598.7	1197.4***
外形规格	**4.4" (高) x 10.5" (长) 双插槽**
显示端口	**3 个 DisplayPort 1.4**; 支持 NVIDIA Mosaic 和 Quadro® Sync[4]**
最大功耗	**300 瓦**
电源接口	**8 引脚 CPU**
散热解决方案	**被动式**
虚拟 GPU (vGPU) 软件支持	**NVIDIA vPC/vApp、NVIDIA RTX 虚拟工作站、NVIDIA 虚拟计算服务器**
支持的 vGPU 配置文件	**请参阅《虚拟 GPU 许可指南》**
NVENC | NVDEC	**1x| 2x (包括 AV1 解码)**
通过硬件信任根进行安全可靠的引导	**是**
NEBS Ready	**3 级**
计算 API	**CUDA、DirectCompute、OpenCL™、OpenACC®**
图形 API	**DirectX 12.07[5]、Shader Model 5.17[5]、OpenGL 4.68[6]、Vulkan 1.18[6]**
MIG 支持	**否**

* 启用结构化稀疏技术
** 默认情况下，我们会将 A40 配置为用于开展虚拟化工作，并禁用物理显示接口。用户可通过管理软件工具启用显示输出。

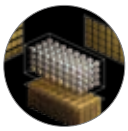
NVIDIA Ampere 架构细览

NVIDIA AMPERE 架构 CUDA 核心

速度提升一倍的单精度浮点 (FP32) 运算处理和改善的能效可显著提高图形和计算工作流程的性能，例如复杂的 3D 计算机辅助设计 (CAD) 和计算机辅助工程 (CAE)。

第二代 RT CORE

凭借高达 2 倍于上一代产品的吞吐量，以及并行运行光线追踪与着色或降噪功能的能力，第二代 RT Core 可大幅加快电影内容的逼真渲染、建筑设计评估以及产品设计的虚拟原型制作等工作负载的运行速度。这项技术还可提升光线追踪动态模糊的渲染速度，从而更快获得结果，并增加视觉准确度。



第三代 TENSOR CORE

Tensor Float 32 (TF32) 精度提供的训练吞吐量高达上一代的 5 倍，而且无需更改代码即可加速 AI 和数据科学模型的训练。从硬件上支持结构化稀疏使推理吞吐量提升一倍。Tensor Core 还为图形处理引入了诸多 AI 功能，例如为选定应用程序带来了深度学习超级采样 (**DLSS**)、AI 降噪和增强编辑等功能。

48 GB GDDR6 显存，支持 NVLINK

超高速 GDDR6 显存可通过 **NVLink**[3] 扩展到高达 96 GB，为数据科学家、工程师和创意专业人士提供所需的大容量显存，让他们能够处理大型数据集以及数据科学和模拟等工作负载。

PCIE EXPRESS 4.0

PCI Express 4.0 提供的带宽比 PCIe Gen 3 多一倍，提高了 CPU 内存的数据传输速度，从而可以更快地处理 AI、数据科学和 3D 设计等数据密集型任务。更快的 PCIe 性能还能加速 GPU 直接显存访问 (DMA) 传输，这在 GPU 与支持 **GPUDirect® for Video** 的设备之间提供了更快的视频数据输入/输出通信速度，从而带来强大的直播解决方案。A40 向后兼容 PCI Express 第 3 代，这提供了部署灵活性。

数据中心效率和安全性

NVIDIA A40 采用节能高效的双插槽设计，能效是前一代的两倍，且可兼容全球 OEM 生产的各种服务器。NVIDIA A40 包含通过硬件信任根技术进行安全可靠的引导，确保固件不会被篡改或损坏。

NVIDIA A40 GPU 可提供先进的视觉计算功能，包括实时光线追踪、AI 加速和多工作负载灵活性，从而加速深度学习、数据科学和基于计算的工作负载。由 NVIDIA A40 和 NVIDIA RTX 虚拟工作站 (vWS) 以及 NVIDIA 虚拟计算服务器软件提供动力支持的虚拟工作站受益于各种行业应用程序和专业软件的广泛测试，可提供极佳的性能和稳定性。

所有深度学习框架

mxnet
PYTORCH
Apache Spark
TensorFlow

适用于专业应用程序的 RTX

Adobe Premiere Pro
SOLIDWORKS
PLM Software SIEMENS NX
AUTODESK ARNOLD
REDSHIFT
AUTODESK VRED
KeyShot
UNREAL ENGINE
blender
octanerender
V-ray

如需详细了解 NVIDIA A40 GPU，请访问 **www.nvidia.com/a40**

1 运行渲染和图形测试的配置如下：2 个至强金牌 6126 2.6GHz (3.7GHz Turbo)。256GB 系统内存。NVIDIA 驱动 461.09。渲染测试：Iray 2020.1，NVIDIA Endeavor 场景的渲染时间。图形测试：SPECviewperf 2020 子测试，4K medical-03 合成 | 2 运行 AI 和 HPC 测试的配置如下：AMD EPYC 7742@2.25GHz (3.4GHz Turbo)。512GB 系统内存。NVIDIA 驱动 460.14。
AI 训练：BERT 预训练吞吐量。PyTorch (2/3) 第 1 阶段和 (1/3) 第 2 阶段。用于 RTX 6000 的精度 FP32 以及用于 A40 和 A100 的 TF32。第 1 阶段的序列长度 = 128。第 2 阶段 = 512。单精度 HPC：NAMD 版本 3.0a7，stmv_nve_cuda；精度=FP32；纳秒/天，CUDA 版本：11.1.74 | 3 只有当应用程序支持 NVLink 技术时，才能通过 NVLink 连接两块 NVIDIA A40 显卡，以将性能翻倍并将显存扩展到 96 GB。请联系您的应用程序提供商以确认是否支持 NVLink。| 4 Quadro Sync II 显卡单独销售。Windows 10 和 Linux 上支持 Mosaic。| 5 GPU 支持 DX 12.0 API，硬件功能级别 12 + 1。| 6 产品基于已发布的 Khronos 规格，有望在可用时通过 Khronos 一致性测试过程。如需了解当前的一致性状况，请访问 www.krhonos.org/conformance

PARTNER LOGO FPO

NVIDIA