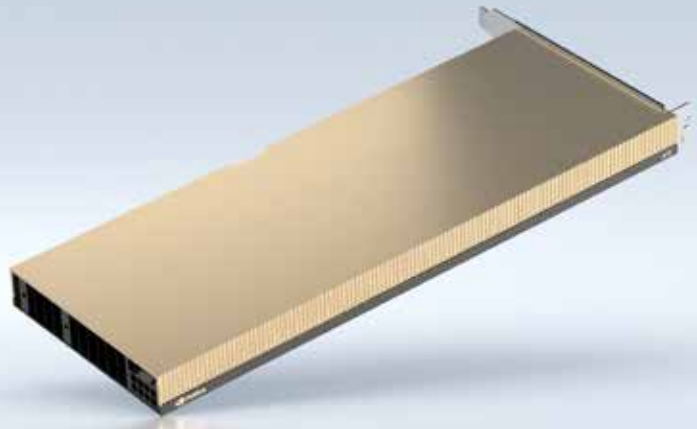
NVIDIA A10

适用于主流企业服务器的 AI 加速的图形和视频

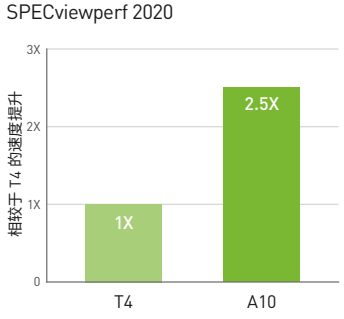


借助强大的 AI 丰富图形和视频应用程序

NVIDIA A10 Tensor Core GPU 与 NVIDIA RTX 虚拟工作站 (vWS) 软件相结合，将主流图形和视频与 AI 服务引入主流企业服务器，为设计师、工程师、艺术家和科学家带来应对当今挑战所需的解决方案。A10 基于最新的 NVIDIA Ampere 架构而构建，将第二代 RT Core、第三代 Tensor Core 和新型流式传输微处理器与 24 GB 的 GDDR6 显存相结合 (皆在 150W 功率范围内)，实现通用的图形、渲染、AI 和计算性能。从可在世界各地访问的虚拟工作站、渲染节点，到运行各种工作负载的数据中心，A10 皆能以单宽、全高、全长 PCIe 外形提供出色性能。

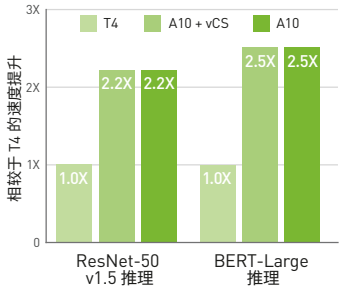
NVIDIA A10 支持作为 NVIDIA-Certified Systems™ 的一部分，在本地数据中心、云和边缘中使用。NVIDIA A10 基于由 NVIDIA NGC™ 目录、CUDA-X™ 库、超过 230 万名开发者和 1800 多个 GPU 优化应用程序组成的丰富的 AI 框架生态系统而构建，帮助企业应对其业务中的关键挑战。

A10 提供比 T4 快 2.5 倍的虚拟工作站性能[1]



A10 提供比 T4 高 2.5 倍的推理性能[2]

NVIDIA vCS 提供等同于裸机的性能



规格

FP32	**31.2 TF**
TF32 Tensor Core	**62.5 TF | 125 TF***
BFLOAT16 Tensor Core	**125 TF |250 TF***
FP16 Tensor Core	**125 TF | 250 TF***
INT8 Tensor Core	**250 TOPS | 500 TOPS***
INT4 Tensor Core	**500 TOPS | 1000 TOPS***
RT Core 数	**72**
编码/解码	**1 个编码器** **2 个解码器** **（+AV1 解码）**
GPU 显存	**24GB GDDR6**
GPU 显存带宽	**600 GB/s**
互联	**PCIe 4.0：64 GB/s**
外形规格	**1 插槽 FHFL**
最大 TDP 功耗	**150W**
vGPU 软件支持	**NVIDIA vPC/vApp、** **NVIDIA RTX™ vWS、** **NVIDIA 虚拟计算服务器 (vCS)**
通过硬件信任根进行安全可靠的引导	**是**
NEBS Ready	**3 级**
电源接口	**PEX 8 针**

*采用稀疏技术

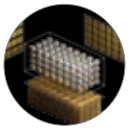
NVIDIA Ampere 架构细览

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。

24GB GDDR6

超高速 GDDR6 显存，为渲染、数据科学、工程模拟和其他 GPU 显存密集型工作负载提供 600 GB/s 带宽。

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 **NVIDIA 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。A10 还向后兼容 PCI Express 3.0，这提供了部署灵活性。

数据中心效率和安全性

NVIDIA A10 采用单插槽、全高、全长节能设计，可兼容全球 OEM 供应商生产的各式服务器。NVIDIA A10 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 A10 Tensor Core GPU 是采用 AI 的主流图形和视频的理想选择。第二代 RT Core 和第三代 Tensor Core 可凭借强大的 AI 在 150W TDP 下为主流服务器丰富图形和视频应用程序。

NVIDIA A10 还可与 NVIDIA 虚拟 GPU (vGPU) 软件结合使用，在易于管理、安全灵活的基础设施 (可进行扩展以满足资源需求) 中加速从图形丰富的 VDI 到高性能虚拟工作站再到 AI 等多个数据中心工作负载。

所有深度学习框架

mxnet

PYTORCH

Spark

TensorFlow

适用于专业应用程序的 RTX

AUTODESK® REVIT®

creo®

CATIA

Rhinoceros®
design, model, present, analyze, realize...

SOLIDWORKS

SIEMENS

如需详细了解 NVIDIA A10 Tensor Core GPU，请访问 **www.nvidia.com/a10**

1 运行测试的服务器配置如下：2 个至强金牌 6154 3.0GHz (3.7GHz Turbo)，NVIDIA RTX vWS 软件，VMware ESXi 7 U2，主机/客户机驱动 461.33。| SPECviewperf 2020 子测试和 HD 3dsmax-07 合成。
2 BERT Large 推理 NVIDIA TensorRT7.2，序列长度 =128，批量大小 =128；NGC 容器：21.02-py3 | ResNet-50 v1.5：NVIDIA TensorRT7.2，INT8 精度批量大小 = 128 NGC 容器：20.12-py3 | 采用 vCS 软件的 NVIDIA A10、VMware ESXi 7 U2 主机/客户机驱动 461.33

NVIDIA