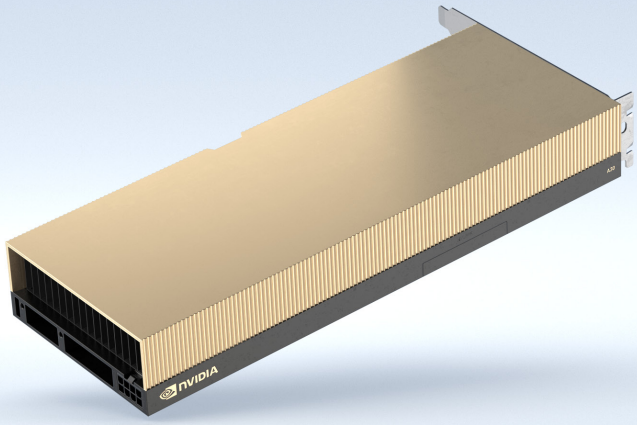
NVIDIA A30
TENSOR CORE GPU
适用于主流企业服务器的多用途计算加速技术



适用于不同企业的 AI 推理和主流计算

NVIDIA A30 Tensor Core GPU 是用途广泛的主流计算 GPU，适用于 AI 推理和主流企业工作负载。这款 GPU 采用 NVIDIA Ampere 架构的 Tensor Core 技术，支持广泛的数学精度，可针对每个工作负载提供单个加速器。

专为大规模 AI 推理而构建的同一计算资源能够通过 TF32 精度快速重新训练 AI 模型，同时还能借助 FP64 Tensor Core 加速高性能计算（HPC）应用。多实例 GPU（MIG）及 FP64 Tensor Core，可在 165W 低功率电路下相结合，实现速度高达 933GB/s 的显存带宽，以上特性均在这一适用于主流服务器的 PCIe 卡上体现。

通过结合使用第三代 Tensor Core 与 MIG 技术，其可在各种工作负载中提供安全的服务质量，所有这些技术都由多功能 GPU 提供支持，从而实现弹性数据中心。A30 在各个规模的工作负载中都具有多用途计算能力，能够尽可能地为主流企业创造价值。

A30 是整个 NVIDIA 数据中心解决方案的一部分，该解决方案由硬件、网络、软件、库以及 NGC™ 中经优化的 AI 模型和应用等构成。作为性能超强的端到端数据中心专用 AI 和 HPC 平台，A30 可助力研究人员交付真实结果，并将解决方案大规模部署到生产环境中。

系统规格

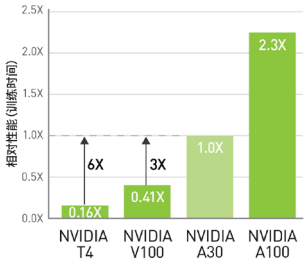
FP64 峰值性能	**5.2 TF**
FP64 Tensor Core 峰值性能	**10.3 TF**
FP32 峰值性能	**10.3 TF**
TF32 Tensor Core	**82 TF | 165 TF***
BFLOAT16 Tensor Core	**165 TF | 330 TF***
FP16 Tensor Core 峰值性能	**165 TF | 330 TF***
INT8 Tensor Core 峰值性能	**330 TOPS | 661 TOPS***
INT4 Tensor Core 峰值性能	**661 TOPS | 1321 TOPS***
媒体引擎	**1 个光流加速器 (OFA)** **1 个 JPEG 解码器 (NVJPEG)** **4 个视频解码器 (NVDEC)**
GPU 显存	**24GB HBM2**
GPU 显存带宽	**933 GB/s**
互联	**第四代 PCIe：64GB/s** **第三代 NVIDIA® NVLINK® 200GB/s****
外形尺寸	**双插槽、全高、全长 (FHFL)**
最大热设计功耗 (TDP)	**165 瓦**
多实例 GPU (MIG)	**4 个 MIG，每个 6GB** **2 个 MIG，每个 12GB** **1 个 MIG，每个 24GB**
虚拟 GPU (vGPU) 软件支持	**适用于 VMware 的 NVIDIA AI Enterprise 软件套件** **NVIDIA 虚拟计算服务器**

* 采用稀疏技术
** NVLink 桥接器最多可连接两个 GPU。

跨工作负载的卓越性能

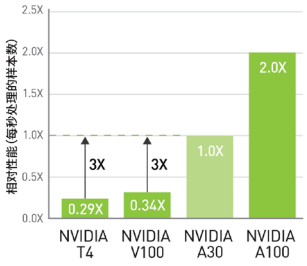
AI 训练 – 吞吐量比 V100 高 3 倍，比 T4 高 6 倍[1]

BERT Large 预训练 (Normalized)



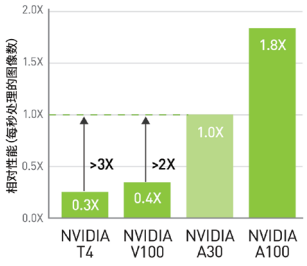
AI 推理 - 实时对话式 AI 的吞吐量比 V100 高 3 倍[2]

BERT Large 推理 (Normalized) 延迟小于 10ms 的吞吐量性能



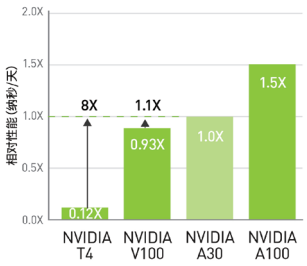
AI 推理 - 在实时图像分类中，吞吐量比 T4 高出 3 倍以上[3]

RN50 v1.5 推理 (Normalized) 延迟小于 7ms 的吞吐量性能

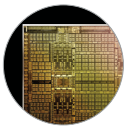


HPC – 吞吐量比 V100 和 T4 分别高 1.1 倍和 8 倍之多[4]

LAMMPS (Normalized)

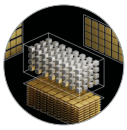


突破性创新



NVIDIA AMPERE 架构

无论是使用 MIG 技术将 A30 GPU 分割为较小的实例，还是使用 NVIDIA NVLink 连接多个 GPU 以加速更大规模的工作负载，A30 均可轻松满足多种规模的加速需求，从小型作业到大型多节点工作负载都无一例外。A30 功能全面，这意味着 IT 经理可借此在主流服务器上充分利用数据中心内的每个 GPU，昼夜不停歇。

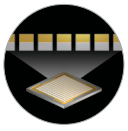


第三代 TENSOR CORE 技术

NVIDIA A30 可提供 165 teraFLOPS (TFLOPS) 的 TF32 精度深度学习性能。相较于 NVIDIA T4 Tensor Core GPU，A30 可将 AI 训练吞吐量提高 20 倍，并将推理性能提高 5 倍以上。A30 可在 HPC 方面提供 10.3 TFLOPS 的性能，比 NVIDIA V100 Tensor Core GPU 高出了近 30%。

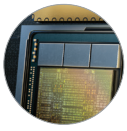
新一代 NVLINK

A30 中采用的 NVIDIA NVLink 可提供两倍于上一代的吞吐量。两个 A30 PCIe GPU 可通过 NVLink 桥接器连接，以提供 330 TFLOPS 的深度学习性能。



多实例 GPU (MIG)

每个 A30 GPU 最多可分割为四个 GPU 实例，这些实例在硬件级别完全独立，并各自拥有高带宽显存、缓存和计算核心。借助 MIG，开发者可为其所有应用实现惊人加速。IT 管理员可为每个作业提供符合其规模的 GPU 加速，进而优化利用率，并让每个用户和应用都能享受 GPU 加速性能。



HBM2 显存

配合高达 24GB 的高带宽显存 (HBM2)，A30 可提供 933GB/s 的 GPU 显存带宽，适用于主流服务器中的多种 AI 和 HPC 工作负载。

结构化稀疏

AI 网络拥有数百万至数十亿个参数。实现准确预测并非要使用所有参数，而且我们还可将某些参数转换为零，以在无损准确性的前提下使模型变得“稀疏”。A30 中的 Tensor Core 可为稀疏模型提供高达两倍的性能提升。稀疏功能不仅更易使 AI 推理受益，同时还能提升模型训练的性能。

为企业提供端到端解决方案

NVIDIA A30 Tensor Core GPU 采用现代数据中心的核心 —— NVIDIA Ampere 架构，是 NVIDIA 数据中心平台不可或缺的一部分。该平台专为深度学习、HPC 及数据分析而构建，并为包括各大深度学习框架在内的 2000 余款应用提供加速。此外，NVIDIA AI Enterprise 是一套端到端云原生 AI 和数据分析软件套件，经认证可在 A30 上运行，适用于结合 VMware vSphere 的基于 hypervisor 的虚拟基础架构。这使您能够在混合云环境中管理和扩展 AI 工作负载。从数据中心到边缘节点均可使用完善的 NVIDIA 平台，不仅能显著提升性能，还能创造众多成本节约机会。

适用于企业的优化软件和服务

NVIDIA CERTIFIED

各种深度学习框架

mxnet PYTORCH APACHE Spark TensorFlow

2000 余款 GPU 加速应用

- Altair nanoFluidX
- Altair ultraFluidX
- AMBER
- ANSYS Fluent
- DS SIMULIA Abaqus
- GAUSSIAN
- GROMACS
- NAMD
- OpenFOAM
- VASP
- WRF

如需详细了解 NVIDIA A30 Tensor Core GPU，请访问 **www.nvidia.cn/data-center/products/a30-gpu**

[1] BERT-Large 预训练（9/10 次）阶段 1 及（1/10 次）阶段 2，阶段 1 的序列长度 = 128，阶段 2 的序列长度 = 512，数据集 = 实时，NGC™ 容器 = 21.03，8 个 GPU：T4（FP32，BS=8，2）| V100 PCIe 16GB（FP32，BS=8，2）| A30（TF32，BS=8，2）| A100 PCIe 40GB（TF32，BS=54，8）| 显示的批量大小分别对应阶段 1 和阶段 2

[2] NVIDIA® TensorRT®，精度 = INT8，序列长度 = 384，NGC 容器 20.12，延迟小于 10ms，数据集 = 合成；1 个 GPU：A100 PCIe 40GB（BS=8）| A30（BS=4）| V100 SXM2 16GB（BS=1）| T4（BS=1）

[3] TensorRT，NGC 容器 20.12，延迟小于 7ms，数据集 = 合成，1 个 GPU：T4（BS=31，INT8）| V100（BS=43，混合精度）| A30（BS=96，INT8）| A100（BS=174，INT8）

[4] 数据集：ReaxFF/C，FP64 | 4 个 GPU：T4，V100 PCIe 16GB，A30

NVIDIA