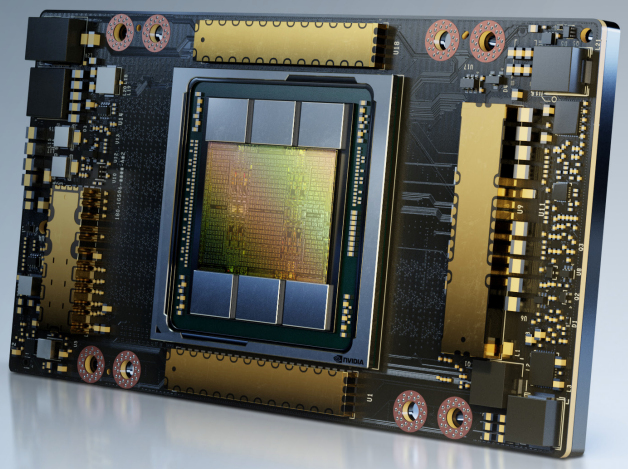
NVIDIA

NVIDIA A100 TENSOR CORE GPU

在各个规模上实现出色加速



适用各种工作负载的出色计算平台

NVIDIA A100 Tensor Core GPU 可针对 AI、数据分析和高性能计算（HPC）应用，在各个规模下实现出色加速，有效助力全球高性能弹性数据中心。作为 NVIDIA 数据中心平台的引擎，相较于前一代 NVIDIA Volta™，A100 可提供高达 20 倍的性能。A100 支持高效扩展，也可划分为七个独立的 GPU 实例，多实例 GPU（MIG）可提供统一平台，助力弹性数据中心动态地适应不断变化的工作负载需求。

NVIDIA A100 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。最新一代 A100 80GB 将 GPU 显存加倍，提供 2TB/s 的全球超快显存带宽，可加速处理超大型模型和海量数据集。

A100 是完整 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 模型和应用等叠加而成。作为性能超强的端到端数据中心专用 AI 和 HPC 平台，它可助力研究人员交付真实的结果，并将解决方案大规模部署到生产环境中。

NVIDIA A100 TENSOR CORE GPU 规格（SXM4 和 PCIE 外形尺寸）

	A100 40GB PCIe	A100 80GB PCIe	A100 40GB SXM	A100 80GB SXM
FP64	9.7 TFLOPS			
FP64 Tensor Core	19.5 TFLOPS			
FP32	19.5 TFLOPS			
Tensor Float 32 （TF32）	156 TFLOPS | 312 TFLOPS*			
BFLOAT16 Tensor Core	312 TFLOPS | 624 TFLOPS*			
FP16 Tensor Core	312 TFLOPS | 624 TFLOPS*			
INT8 Tensor Core	624 TOPS| 1248 TOPS*			
GPU 显存	40GB HBM2	80GB HBM2e	40GB HBM2	80GB HBM2e
GPU 显存带宽	1555GB/s	1935GB/s	1555GB/s	2039GB/s
最大热设计功耗（TDP）	250W	300W	400W	400W
多实例 GPU	最多 7 MIG @ 5GB	最多 7 MIG @ 10GB	最多 7 MIG @ 5GB	最多 7 MIG @ 10GB
外形尺寸	PCIe		SXM	
互联	搭载 2 个 GPU 的 NVIDIA® NVLink® 桥接器：600GB/s** PCIe 4.0：64GB/s		NVLink：600GB/s PCIe 4.0：64GB/s	
服务器选项	合作伙伴及配备 1 至 8 个 GPU 的 NVIDIA 认证系统™		NVIDIA HGX™ A100 合作伙伴和配备 4、8 或 16 个 GPU 的 NVIDIA 认证系统 配备 8 个 GPU 的 NVIDIA DGX™ A100	

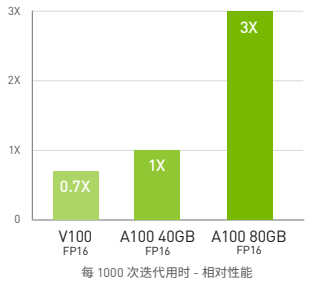
* 采用稀疏技术

** SXM4 GPU 通过 HGX A100 服务器主板连接；PCIe GPU 通过 NVLink 桥接器可桥接多达两个 GPU

跨工作负载的卓越性能

针对大型模型提供高达 3 倍的 AI 训练速度

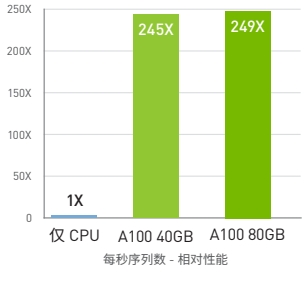
DLRM 训练



HugeCTR 框架上的 DLRM，精度 = FP16 | NVIDIA A100 80GB 批量大小 = 48 | NVIDIA A100 40GB 批量大小 = 32 | NVIDIA V100 32GB 批量大小 = 32。

相较于 CPU 提供高达 249 倍的 AI 推理性能

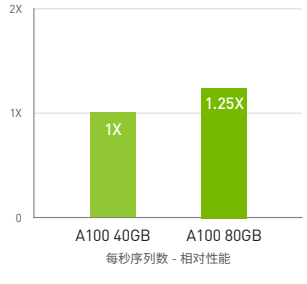
BERT-LARGE 推理



BERT-Large 推理 | 仅 CPU：两个至强金牌 6240 @2.60 GHz，精度 = FP32，批量大小 = 128 | V100：NVIDIA Tensor-RT™ (TRT) 7.2，精度 = INT8，批量大小 = 256 | A100 40GB 和 80GB，批量大小 = 256，精度 = INT8（采用稀疏技术）。

相较于 A100 40GB 提供高达 1.25 倍的 AI 推理性能

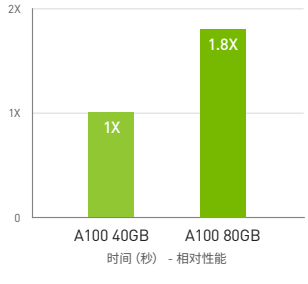
RNN-T 推理：单流



使用 (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。

为 HPC 应用提供高达 1.8 倍的性能

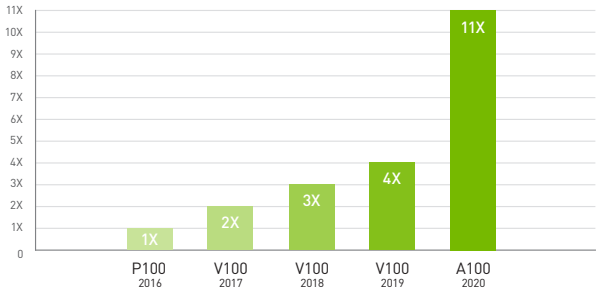
Quantum Espresso



使用 CNT10POR8 数据集测量的 Quantum Espresso，精度 = FP64。

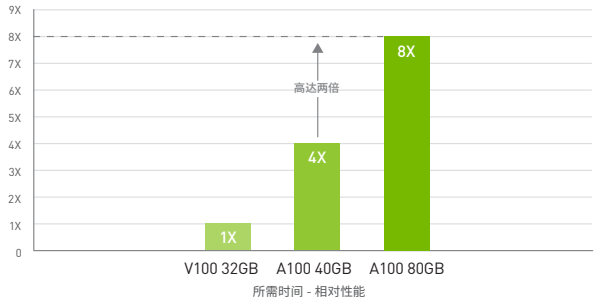
四年来 HPC 性能提升 11 倍

热门 HPC 应用的吞吐量



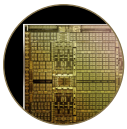
应用加速的几何平均数与 P100 加速性能对比：基准应用：Amber [PME-Cellulose_NVE]、Chroma [szscl21_24_128]、GROMACS [ADH Dodec]、MILC [Apex Medium]、NAMD [stmv_nve_cuda]、PyTorch [BERT-Large Fine Tuner]、Quantum Espresso [AUSURF112-jR]、随机森林 FP32 [make_blobs (160000 x 64 : 10)]、TensorFlow [ResNet-50]、VASP 6 [Si Huge] | 包含双路 CPU 和 4 个 NVIDIA P100、V100 或 A100 GPU 的 GPU 节点。

相较于 A100 40GB，在大数据分析基准测试中速度提升 2 倍



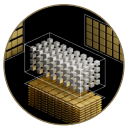
大数据分析基准测试 | GPU-BDB 衍生自 TPCx-BB 基准测试，适用于内部性能测试。GPU-BDB 测试结果并未与 TPCx-BB 的相比较 | 针对 10TB 数据集的 30 次分析零售查询、ETL、ML、NLP | V100 32GB，RAPIDS/Dask | A100 40GB 和 A100 80GB，RAPIDS/Dask/BlazingSQL

突破性的创新



NVIDIA AMPERE 架构

无论是使用 MIG 将 A100 GPU 分割为较小的实例，还是使用 NVLink 连接多个 GPU 来加速大规模工作负载，A100 均可轻松满足不同规模的加速需求，从小型作业到大型多节点工作负载无一例外。A100 功能全面，这意味着 IT 经理可借此全天候充分利用数据中心内的每个 GPU。

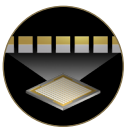


第三代 TENSOR CORE 技术

NVIDIA A100 的深度学习运算能力可达 312 teraFLOPS (TFLOPS)。其深度学习训练的 Tensor 每秒浮点运算次数 (FLOPS) 和推理的 Tensor 每秒万亿次运算次数 (TOPS) 皆为 NVIDIA Volta™ GPU 的 20 倍。

新一代 NVLINK

A100 中采用的 NVIDIA NVLink 可提供两倍于上一代的吞吐量。与 NVIDIA NVSwitch™ 结合使用时，此技术可将多达 16 个 A100 GPU 互联，并将速度提升至 600GB/s，从而在单个服务器上实现出色的应用性能。NVLink 技术可应用在 A100 中：SXM GPU 通过 HGX A100 服务器主板连接，PCIe GPU 通过 NVLink 桥接器可桥接多达 2 个 GPU。



多实例 GPU (MIG)

一个 A100 GPU 最多可分割成七个 GPU 实例，这些实例在硬件级别完全独立，并独自拥有高带宽显存、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 管理员也可为每个作业提供符合其规模的 GPU 加速，进而优化 GPU 利用率，并让每个用户和应用都能使用 GPU 实例。

高带宽显存 (HBM2E)

凭借 80GB 的高带宽显存 (HBM2e)，A100 成为世界首款将显存带宽提升至超过 2TB/s 的 GPU，并将动态随机存取存储器 (DRAM) 的利用效率提高至 95%。A100 提供的显存带宽是上一代产品的 1.7 倍。

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”。A100 中的 Tensor Core 可令稀疏模型的性能获得高达两倍的提升。稀疏功能不仅更容易使 AI 推理受益，同时还能提升模型的训练性能。

NVIDIA A100 Tensor Core GPU 是 NVIDIA 数据中心平台的旗舰产品，可用于深度学习、HPC 和数据分析。该平台可为 2000 余款应用和各大深度学习框架提供加速。A100 适用于桌面、服务器以及云服务，不仅能显著提升性能，还能创造众多节约成本的机会。

面向企业的优化软件和服务

NVIDIA CERTIFIED

各个深度学习框架

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 A100 Tensor Core GPU，
请访问 **https://www.nvidia.cn/data-center/a100/**

NVIDIA