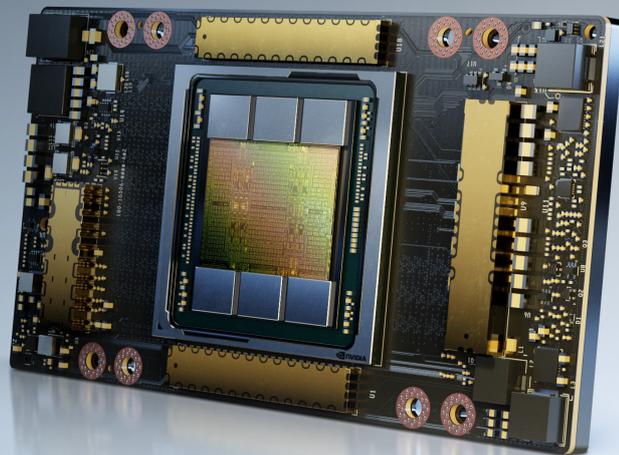




NVIDIA A100 TENSOR CORE GPU

在各种规模上实现出色的加速



适用于各种工作负载的强大计算平台

NVIDIA A100 Tensor Core GPU 可在各个规模下为 AI、数据分析和高性能计算 (HPC) 应用提供出色的加速性能，为全球的高性能弹性数据中心提供强劲助力。作为 NVIDIA 数据中心平台的引擎，与前一代 NVIDIA Volta™ 相比，A100 可使性能提升高达 20 倍。A100 可高效扩展，也可借助多实例 GPU (MIG) 技术划分为 7 个独立的 GPU 实例，从而提供统一的平台，助力弹性数据中心根据不断变化的工作负载需求动态进行调整。

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™ 目录中经优化的 AI 模型和应用等叠加而成。作为适用于数据中心且功能强大的端到端 AI 和 HPC 平台，A100 可助力研究人员获得真实的结果，并能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 外形规格)

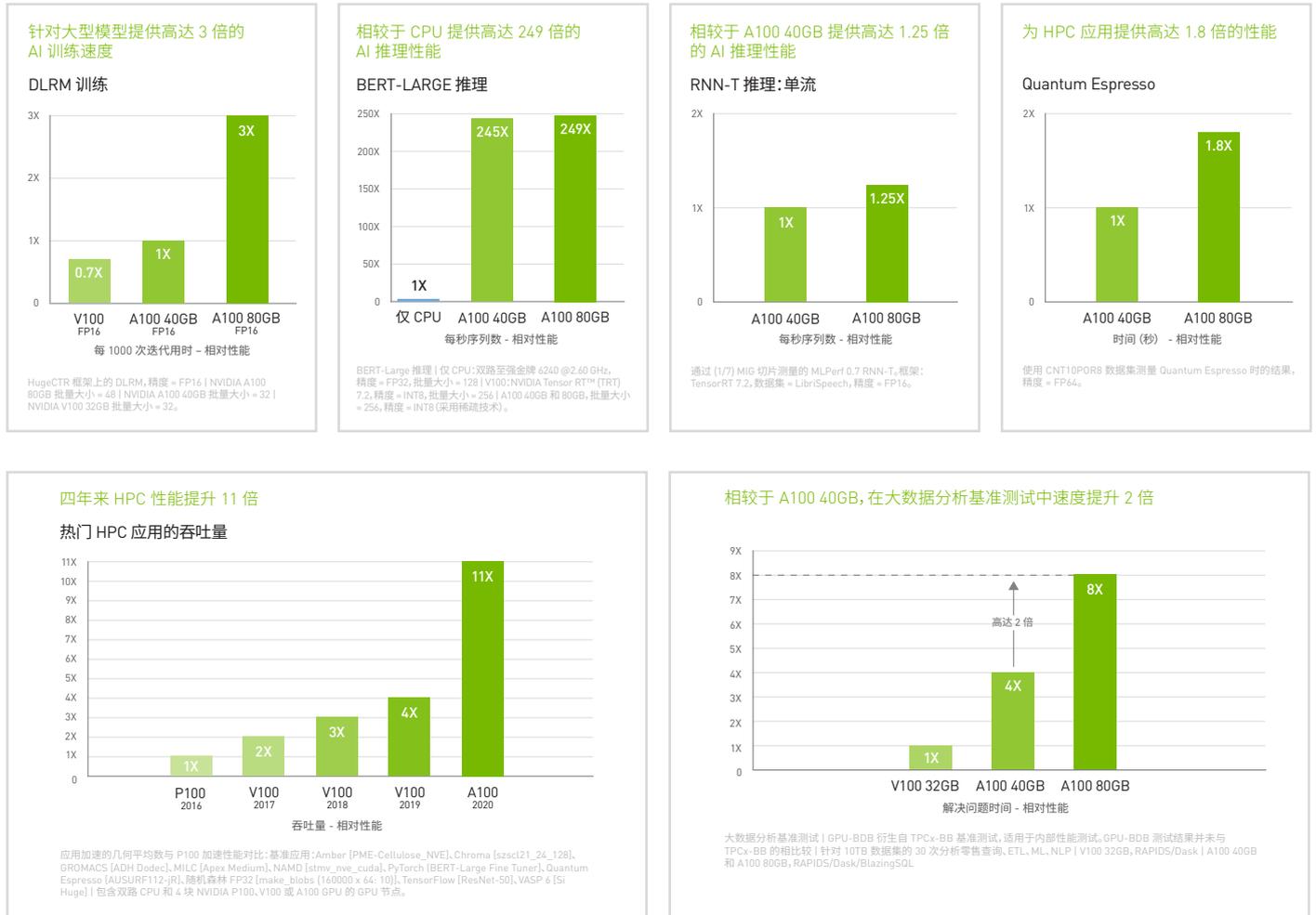
	A100 80GB PCIe	A100 80GB SXM
FP64	9.7 TFLOPS	
FP64 Tensor Core	19.5 TFLOPS	
FP32	19.5 TFLOPS	
Tensor Float 32 (TF32)	156 TFLOPS 312 TFLOPS*	
BFLOAT16 Tensor Core	312 TFLOPS 624 TFLOPS*	
FP16 Tensor Core	312 TFLOPS 624 TFLOPS*	
INT8 Tensor Core	624 TOPS 1248 TOPS*	
GPU 显存	80GB HBM2e	80GB HBM2e
GPU 显存带宽	1935GB/s	2039GB/s
最大热设计功耗 (TDP)	300 瓦	400 瓦 ***
多实例 GPU	最多 7 个 MIG 每个 10GB	最多 7 个 MIG 每个 10GB
外形规格	PCIe 双插槽风冷式或单插槽 液冷式	SXM
互连技术	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	搭载 1 至 8 个 GPU 的 合作伙伴认证系统 和 NVIDIA 认证系统 (NVIDIA-Certified Systems™)	搭载 4 个、8 个或 16 个 GPU 的 NVIDIA HGX™ A100 合作伙伴认证系统 和 NVIDIA 认证系统 搭载 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

*** 400 瓦 TDP (适用于标准配置)。HGX A100-80 GB 自定义散热解决方案 (CTS) SKU 可支持高达 500 瓦的 TDP

跨工作负载的卓越性能



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 经理可以全天候地充分利用数据中心内的每个 GPU。

第三代 TENSOR CORE 技术

NVIDIA A100 的深度学习运算性能可达 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 互连，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 桥接器可桥接多达两块 GPU。

多实例 GPU (MIG) 技术

一个 A100 GPU 最多可划分为 7 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 提供超过 2TB/s 的超快 GPU 显存带宽，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 可为稀疏模型提供高达 2 倍的性能提升。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、高性能计算 (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，更可以节约成本。

面向企业的优化软件和服务



各类深度学习框架

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

准备好开始使用了吗？

如需详细了解 NVIDIA A100 Tensor Core GPU，请访问
www.nvidia.cn/data-center/a100/

