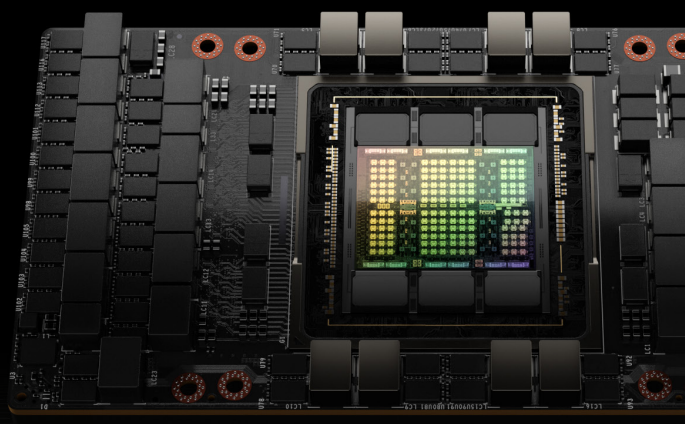
NVIDIA

NVIDIA H100 TENSOR CORE GPU

助力各个数据中心实现卓越的性能、可扩展性和安全性



加速计算的数量级飞跃

NVIDIA H100 Tensor Core GPU 可助力各种工作负载实现卓越的性能、可扩展性和安全性。借助 NVIDIA® NVLink® Switch 系统，可连接多达 256 个 H100 GPU 来加速百亿亿次级（Exascale）工作负载，并可通过专用的 Transformer 引擎来为万亿参数的语言模型提供支持。H100 利用 NVIDIA Hopper™ 架构中的突破性创新技术提供先进的对话式 AI，与上一代产品相比，可使大型语言模型的速度提升 30 倍。

安全地加速从企业级到百亿亿次级（Exascale）规模的工作负载

NVIDIA H100 GPU 配备第四代 Tensor Core 和 Transformer 引擎（FP8 精度），可使大型语言模型的训练速度提升高达 9 倍，推理速度提升惊人的 30 倍，从而进一步拓展了 NVIDIA 在 AI 领域的市场领先地位。对于高性能计算（HPC）应用，H100 可使 FP64 的每秒浮点运算次数（FLOPS）提升至 3 倍，并可添加动态编程（DPX）指令，使性能提升高达 7 倍。借助第二代多实例 GPU（MIG）技术、内置的 NVIDIA 机密计算和 NVIDIA NVLink Switch 系统，H100 可安全地加速从企业级到百亿亿次级（Exascale）规模的数据中心的各种工作负载。

H100 是完整的 NVIDIA 数据中心解决方案的一部分，该解决方案包含以下方面的基础模组：硬件、网络、软件、库以及 NVIDIA NGC™ 目录中经优化的 AI 模型和应用。作为适用于数据中心且功能强大的端到端 AI 和 HPC 平台，H100 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。

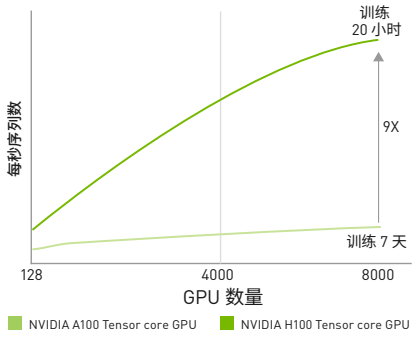
规格

	H100 SXM	H100 PCIe
FP64	**30 TFLOPS**	**24 TFLOPS**
FP64 Tensor Core	**60 TFLOPS**	**48 TFLOPS**
FP32	**60 TFLOPS**	**48 TFLOPS**
TF32 Tensor Core	**1000 TFLOPS***	**800 TFLOPS***
BFLOAT16 Tensor Core	**2000 TFLOPS***	**1600 TFLOPS***
FP16 Tensor Core	**2000 TFLOPS***	**1600 TFLOPS***
FP8 Tensor Core	**4000 TFLOPS***	**3200 TFLOPS***
INT8 Tensor Core	**4000 TOPS***	**3200 TOPS***
GPU 显存	**80GB**	**80GB**
GPU 显存带宽	**3TB/s**	**2TB/s**
解码器	**7 NVDEC 7 JPEG**	**7 NVDEC 7 JPEG**
最大热设计功耗（TDP）	**700 瓦**	**350 瓦**
多实例 GPU	**最多 7 个 MIG，每个 10GB**	
外形规格	**SXM**	**PCIe 双插槽 风冷式**
互连技术	**NVLink：900GB/s PCIe 5.0：128GB/s**	**NVLink：600GB/s PCIe 5.0：128GB/s**
服务器选项	**搭载 4 个或 8 个 GPU 的 NVIDIA HGX™ H100 合作伙伴认证系统和 NVIDIA 认证系统（NVIDIA-Certified Systems™） 搭载 8 个 GPU 的 NVIDIA DGX™ H100**	**搭载 1 至 8 个 GPU 的合作伙伴认证系统及 NVIDIA 认证系统**

* 采用稀疏技术显示。在不采用稀疏技术的情况下，规格降低一半。

可使大型模型的 AI 训练速度提升高达 9 倍

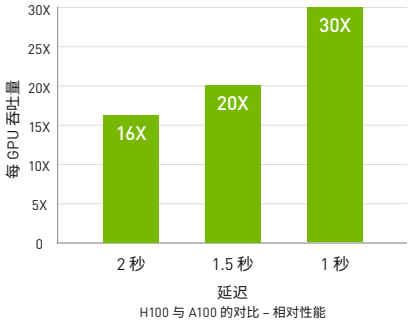
多专家模型（3950 亿个参数）



预计的性能可能会有所变化。利用 1T token 数据集训练具有 3950 亿个参数的多专家模型（MoE）Transformer Switch-XXL 变体 | A100 集群：HDR IB 网络 | H100 集群：NVLink Switch 系统、NDR IB

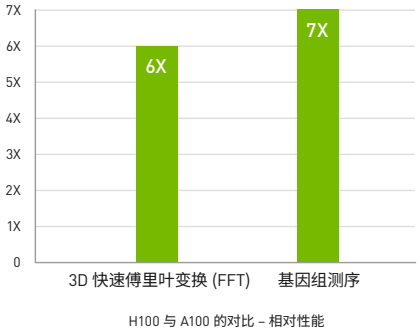
可使大型模型的 AI 推理性能提升高达 30 倍

Megatron 聊天机器人推理（5300 亿个参数）



预计的性能可能会有所变化。基于 Megatron 530B 参数模型的聊天机器人推理，输入序列长度 = 128，输出序列长度 = 20 | A100 集群：HDR IB 网络 | H100 集群：NVLink Switch 系统、NDR IB

可使 HPC 应用的性能提升高达 7 倍

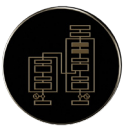


预计的性能可能会有所变化。3D FFT (4K^3) 吞吐量 | A100 集群：HDR IB 网络 | H100 集群：NVLink Switch 系统、NDR IB | 基因组测序 (Smith-Waterman) | 1 个 A100 | 1 个 H100

NVIDIA Hopper 的技术突破

超先进的芯片

H100 由 800 亿个晶体管构建而成，这些晶体管采用专为 NVIDIA 加速计算需求设计的领先的台积电 4N 工艺，造就了这一非常先进的芯片。H100 能够显著提升 AI、HPC、显存带宽、互连和数据中心级通信的速度。



TRANSFORMER 引擎

Transformer 引擎采用软件和 Hopper Tensor Core 技术打造，该技术旨在加速通过重要 AI 模型基础模组（即 Transformer）构建模型的训练工作。Hopper Tensor Core 能够应用混合的 FP8 和 FP16 精度，以大幅加速 Transformer 模型的 AI 计算。

NVLINK SWITCH 系统

NVLink Switch 系统可以跨多个服务器，以每个 GPU 900GB/s 的双向带宽扩展多 GPU 输入 / 输出 (IO)，比 PCIe 5.0 的带宽高 7 倍。NVLink Switch 系统支持由多达 256 个 H100 组成的集群，且带宽比 NVIDIA Ampere 架构上的 InfiniBand HDR 高 9 倍。

NVIDIA 机密计算

NVIDIA 机密计算是 Hopper 架构的内置安全功能，该功能使 NVIDIA H100 成为率先推出的一款具有机密计算功能的加速器。用户可以在获取 H100 GPU 出色加速功能的同时，保护其使用中的数据和应用的机密性和完整性。

第二代多实例 GPU (MIG) 技术

Hopper 架构的第二代 MIG 技术在虚拟化环境中支持多租户、多用户配置，可将 GPU 安全地划分为大小合适的独立实例，以更大限度地为多达 7 倍的安全租户提升服务质量 (QoS)。

DPX 指令

Hopper 架构引入了 DPX 指令，与 CPU 相比将动态编程算法速度提高了 40 倍，与 NVIDIA 前一代 Ampere 架构 GPU 相比，则提高了 7 倍。这大幅加快了疾病诊断、实时路由优化以及图形分析的速度。

NVIDIA H100 CNX 融合加速器

NVIDIA H100 CNX 将 NVIDIA H100 的强大功能与 **NVIDIA ConnectX®-7** 智能网卡（SmartNIC）的先进网络功能融合到一个独特平台上。这种融合为 GPU 驱动的 IO 密集型工作负载提供出色的性能，例如企业数据中心的分布式 AI 训练和边缘的 5G 处理。**详细请了解 NVIDIA H100 CNX**。

企业就绪

NVIDIA H100 Tensor Core GPU 采用全球 AI 基础架构的新引擎，即 NVIDIA Hopper 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 2700 余款应用提供加速。此外，NVIDIA AI Enterprise 还是一套端到端原生云 AI 和数据分析软件套件，经认证可在 H100 上运行，适用于结合 VMware vSphere 的基于服务器虚拟化平台的虚拟基础架构。这使得在混合云环境中管理和扩展 AI 工作负载成为可能。从数据中心到边缘节点均可使用完整的 NVIDIA 平台，不仅能显著提升性能，还能创造众多节约成本的机会。

适用于企业的优化软件和服务

NVIDIA CERTIFIED

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	Simcenter STAR-CCM+

准备好开始使用了吗？

如需详细了解 NVIDIA H100 Tensor Core GPU，请访问 **www.nvidia.cn/data-center/h100/**

NVIDIA