供货英伟达NVIDIA显卡H100

产品名称	供货英伟达NVIDIA显卡H100
公司名称	深圳市创宏佳科技有限公司
价格	.00/件
规格参数	品牌:NVIDIA 型号:H100（80G） 类型:GPU
公司地址	深圳市龙岗区坂田街道金洲嘉丽园2#楼二单元0504
联系电话	021-16601807362 16601807362

产品详情

NVIDIA H100 Tensor Core GPU

为各类数据中心提供出色的性能、可扩展性和安全性。

加速计算的数量级飞跃

通过 NVIDIA H100 Tensor Core GPU，在每个工作负载中实现出色性能、可扩展性和安全性。使用 NVIDIA NVLink Switch 系统，可连接多达 256 个 H100 来加速百亿亿级 (Exascale) 工作负载，另外可通过专用的 Transformer 引擎来处理万亿参数语言模型。与上一代产品相比，H100 的综合技术创新可以将大型语言模型的速度提高 30 倍，从而提供业界lingxian的对话式 AI。

准备好迎接企业 AI 了吗？

企业采用 AI 现已成为主流，企业组织需要端到端的 AI 就绪型基础架构，加快自身迈向新时代的步伐。

适用于主流服务器的 H100 随附五年期 NVIDIA AI Enterprise 软件套件订阅（包括企业支持），能够以强大的性能简化 AI 的采用。这可确保组织能够访问构建 H100 加速的 AI 工作流所需的 AI 框架和工具，例如 AI 聊天机器人、推荐引擎和视觉 AI 等。

NVIDIA H100 TENSOR CORE GPU 规格 （SXM4 和 PCIE 外形规格）

外形规格	H100 SXM4	H100 PCIE

FP64	34 teraFLOPS	26 teraFLOPS
FP64 Tensor Core	67 teraFLOPS	51 teraFLOPS
FP32	67 teraFLOPS	51 teraFLOPS
TF32 Tensor Core	989 teraFLOPS*	756teraFLOPS*
BFLOAT16 Tensor Core	1979 teraFLOPS*	1,513 teraFLOPS*
FP16 Tensor Core	1979 teraFLOPS*	1,513 teraFLOPS*
FP8 Tensor Core	3958 teraFLOPS*	3026 teraFLOPS*
INT8 Tensor Core	3958 TOPS*	3026 TOPS*
GPU 显存	80GB	80GB
GPU 显存带宽	3.35TB/s	2TB/s
解码器	7 NVDEC/7 JPEG	7 NVDEC/7 JPEG
最大热设计功率 (TDP)	高达 700 瓦（可配置）	300-350 瓦（可配置）
多实例 GPU	最多 7 个 MIG @每个 10GB	
外形规格	SXM	PCIe双插槽风冷式

安全地加速从企业级到百亿亿次级规模的工作负载

实时深度学习推理：AI 正在利用一系列广泛的神经网络解决范围同样广泛的一系列商业挑战。出色的 AI 推理加速器不仅要提供非凡性能，还要利用通用性加速这些神经网络。

H100 进一步扩展了 NVIDIA 在推理领域的shichanglingxian地位，其多项先进技术可将推理速度提高 30 倍，并提供超低的延迟。第四代 Tensor Core 可加速所有精度（包括 FP64、TF32、FP32、FP16 和 INT8）。Transformer 引擎可结合使用 FP8 和 FP16
精度，减少内存占用并提高性能，同时仍能保持大型语言模型的准确性。

超大模型的 AI 推理性能提升高达 30 倍

HPC 应用的性能提升高达 7 倍

百亿亿次级高性能计算：NVIDIA 数据中心平台性能持续提升，超越摩尔定律。H100 的全新突破性 AI 性能进一步加强了 HPC+AI
的力量，加速科学家和研究人员的探索，让他们全身心投入工作，解决世界面临的重大挑战。

H100 使双精度 Tensor Core 的每秒浮点运算 (FLOPS) 提升 3 倍，为 HPC 提供 60 teraFLOPS 的 FP64 浮点运算。融合 AI 的高性能计算应用可以利用 H100 的 TF32 精度实现 1 petaFLOP
的吞吐量，从而在不更改代码的情况下，实现单精度矩阵乘法运算。

H100 还采用 DPX 指令，其性能比 NVIDIA A100 Tensor Core GPU 高 7 倍，在动态编程算法（例如，用于 DNA 序列比对 Smith-Waterman）上比仅使用传统双路 CPU 的服务器快 40 倍。

加速数据分析：在 AI 应用开发过程中，数据分析通常会消耗大部分时间。原因在于，大型数据集分散在多台服务器上，由仅配备商用 CPU
服务器组成横向扩展式的解决方案缺乏可扩展的计算性能，从而陷入困境。

搭载 H100 的加速服务器可以提供相应的计算能力，并利用 NVLink 和 NVSwitch 每个 GPU 3 TB/s 的显存带宽和可扩展性，凭借高性能应对数据分析以及通过扩展支持庞大的数据集。通过结合使用 NVIDIA Quantum-2 InfiniBand、Magnum IO 软件、GPU 加速的 Spark 3.0 和 NVIDIA RAPIDS，NVIDIA 数据中心平台能够以出色的性能和效率加速这些大型工作负载。

为企业提高资源利用率：IT 经理设法更大限度地提高数据中心计算资源的利用率（峰值和平均值）。他们通常会通过动态重新配置来合理调整计算资源，从而满足正在处理的工作负载的需求。

H100 中的第二代多实例 GPU (MIG) 技术通过安全地将每个 GPU 分为 7 个独立实例，更大限度地提高每个 GPU 的利用率。凭借机密计算支持，H100 可实现端到端多租户的安全使用，非常适合云服务提供商 (CSP) 环境。

使用支持 MIG 技术的 H100，基础架构管理员可对其 GPU 加速的基础架构作标准化处理，同时能够灵活地为 GPU 资源配置更精细的粒度，从而安全地为开发者提供正确的加速计算量，并确保其所有 GPU 资源得到充分利用。

内置机密计算：当今的机密计算解决方案基于 CPU，对于 AI 和 HPC 等计算密集型工作负载来说，这些解决方案远远无法满足需求。NVIDIA 机密计算是 NVIDIA Hopper 架构的内置安全功能，该功能使 H100 成为 NVIDIA 率先推出的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，保护使用中的数据和应用的机密性和完整性。它创建了基于硬件的可信执行环境 (TEE)，用于保护并隔离在单个 H100 GPU、节点内多个 H100 GPU 或单个 MIG 实例上运行的整个工作负载。在 TEE 内，GPU 加速应用的运行可以保持不变，且不必对其进行分区。用户可以将适用于 AI 和 HPC 的 NVIDIA 软件的强大功能与 NVIDIA 机密计算提供的硬件信任根的安全性相结合。

为大规模 AI 和高性能计算提供出色的性能：Hopper Tensor Core GPU 将为 NVIDIA Grace Hopper CPU+GPU 架构提供支持，该架构专为 TB 级加速计算而构建，可为大型 AI 和 HPC 提供 10 倍的性能。NVIDIA Grace CPU 利用 Arm 架构的灵活性来创建 CPU 和服务器架构，该架构是专门针对加速计算而从头开始设计的。Hopper GPU 与 Grace CPU 搭配，使用 NVIDIA 超快速的芯片间互连技术，可提供 900GB/s 的带宽，比 PCIe 5.0 快 7 倍。与当今运行最快的服务器相比，这种创新设计将 GPU 的聚合系统显存带宽提高 30 倍，并且会将运行数万亿字节数据的应用性能提高 10 倍。