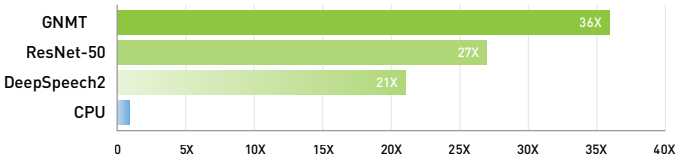
NVIDIA

NVIDIA T4
TENSOR 核心 GPU

助力 AI 训练和推理横向扩展

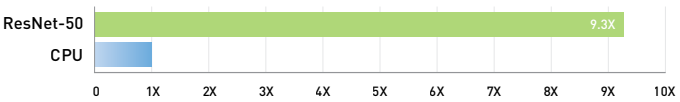
利用全球性能超强劲的扩展加速器 NVIDIA® T4 GPU 打造动力澎湃的服务器。它的 70 瓦半高设计由 NVIDIA Turing™ Tensor 核心提供动力支持，具有革命性的多精度推理性能，可加速各种当今热门的应用程序。这款先进的 GPU 封装在外形小巧的 70 瓦低能耗 PCIe 中，且针对服务器横向扩展进行了优化，专为提供杰出的 AI 性能而打造。

推理性能



一个 NVIDIA T4 GPU 与配双路至强 Gold 6140 CPU 的服务器进行对比

训练性能



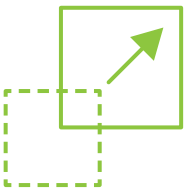
两个 NVIDIA T4 GPU 与配双路至强 Gold 6140 CPU 的服务器进行对比

规格

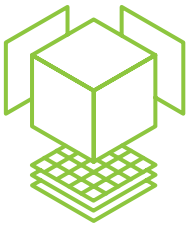
GPU 架构	**NVIDIA Turing**
NVIDIA Turing Tensor 核心数量	**320**
NVIDIA CUDA® 核心数量	**2560**
单精度	**8.1 TFLOPS**
混合精度 (FP16/FP32)	**65 TFLOPS**
INT8	**130 TOPS**
INT4	**260 TOPS**
GPU 显存	**16 GB GDDR6 300 GB/s**
ECC	**支持**
互联带宽	**32 GB / 秒**
系统接口	**x16 PCIe Gen3**
外形尺寸	**PCIe 半高卡**
散热解决方案	**被动式**
计算 API	**CUDA NVIDIA TensorRT™ ONNX**

性能横向扩展推动数据中心加速

小巧的 70 瓦外形设计让 T4 针对可扩展服务器进行了优化，能效相比 CPU 提高了 50 倍之多，大大降低了运营成本。过去两年，NVIDIA 推理平台的性能提高了 10 多倍，仍然是极具能效的分布式 AI 训练和推理解决方案。



NVIDIA T4 数据中心 GPU 是完美适用于分布式计算环境的通用加速器。革命性的多精度性能可加速深度学习以及机器学习训练和推理、视频转码和虚拟桌面。T4 支持所有 AI 框架和网络类型，性能强劲，效率卓越，可最大限度提高大规模部署的效用。



Turing Tensor 核心技术具有多精度计算特性，实现了从 FP32、FP16 到 INT8 以及 INT4 精度的突破性 AI 性能。与 CPU 相比，它的训练性能高达 9.3 倍，推理性能高达 36 倍。

如需详细了解 NVIDIA T4，请访问 **www.nvidia.cn/T4**