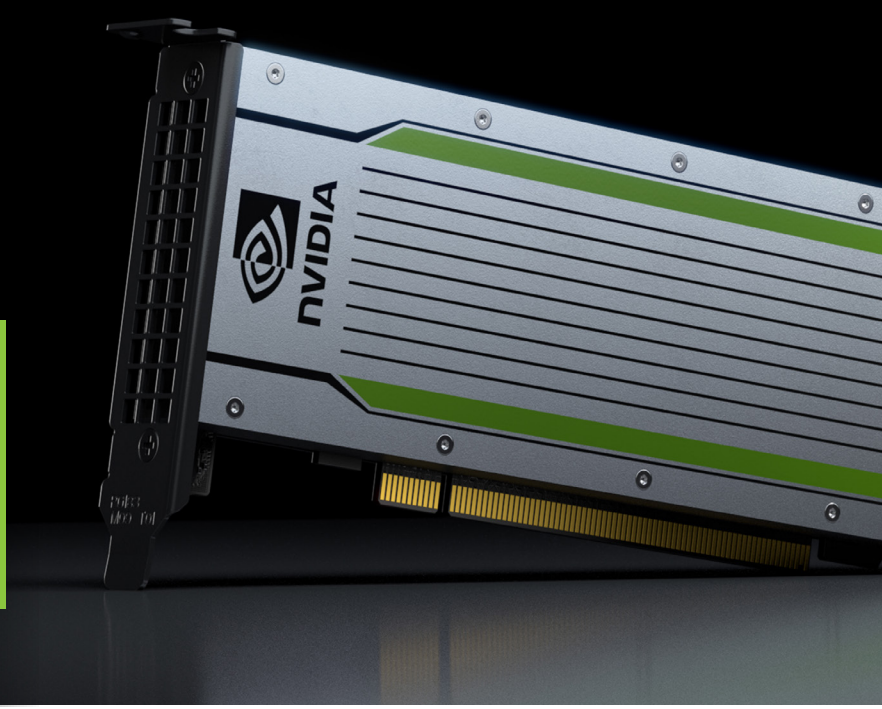
NVIDIA

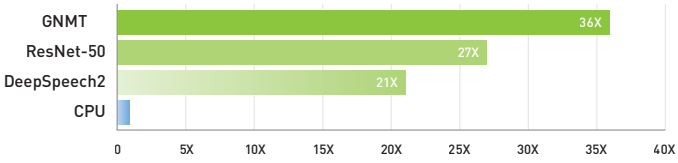


NVIDIA T4
TENSOR 核心 GPU

驅動橫向擴充 AI 訓練和推論

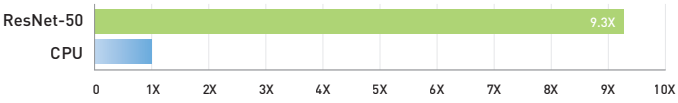
利用世上效能最高的橫向擴充加速器 — NVIDIA® T4 GPU 加快任何伺服器 。 其半高 70 瓦設計搭載 NVIDIA Turing™ Tensor 核心 ， 提供顛覆性的多精度效能 ， 加快各種現代應用 。 這款先進的 GPU 採用高能源效率 70 瓦小尺寸 PCIe 封裝 ， 針對橫向擴充伺服器進行最佳化 ， 專門用於提供最先進的 AI 。

推論效能

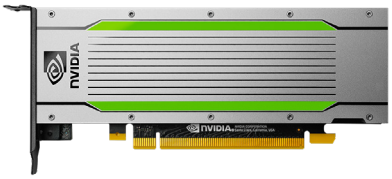


一個 NVIDIA T4 GPU 與搭載雙插槽 Xeon Gold 6140 CPU 之伺服器的比較 。

訓練效能



兩個 NVIDIA T4 GPU 與搭載雙插槽 Xeon Gold 6140 CPU 之伺服器的比較 。

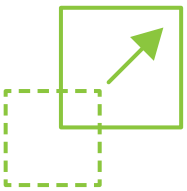


規格

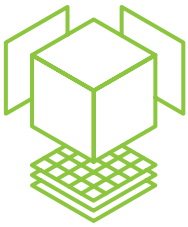
GPU 架構	**NVIDIA Turing**
NVIDIA Turing Tensor 核心	**320**
NVIDIA CUDA® 核心	**2,560**
單精度	**8.1 TFLOPS**
混合精度（FP16/FP32）	**65 TFLOPS**
INT8	**130 TOPS**
INT4	**260 TOPS**
GPU 記憶體	**16 GB GDDR6 300 GB/s**
ECC	**有**
互連頻寬	**32 GB/ 秒**
系統介面	**x16 PCIe Gen3**
尺寸	**半高 PCIe**
散熱解決方案	**被動式**
運算 API	**CUDA, NVIDIA TensorRT™, ONNX**

推動資料中心加速的橫向擴充效能

小尺寸 70 瓦（W）設計使 T4 適用於橫向擴充伺服器 ， 能源效率比 CPU 高 50 倍 ， 大幅降低營運成本 。 在過去兩年內 ， NVIDIA 的推論平台將效率提升超過 10 倍 ， 依然是能源效率最高的分散式 AI 訓練和推論解決方案 。



NVIDIA T4 資料中心 GPU 是理想的通用加速器 ， 適用於分散式運算環境 。 顛覆性的多精度效能可加快深度學習和機器學習訓練與推論 、 視訊轉碼 ， 以及虛擬桌面 。 T4 支援所有 AI 架構和網路類型 ， 提供驚人的效能和效率 ， 將大規模部署的效用最大化 。



Turing Tensor 核心技術支援適用於 AI 的多精度運算 ， 提供從 FP32 至 FP16 至 INT8 的突破性效能 ， 以及 INT4 精度 。 其訓練和推論效能分別比 CPU 高 9.3 倍和 36 倍 。

欲深入瞭解 NVIDIA T4 ， 請上網站 **www.nvidia.com/T4**

ZERONE 零壹科技

NVIDIA