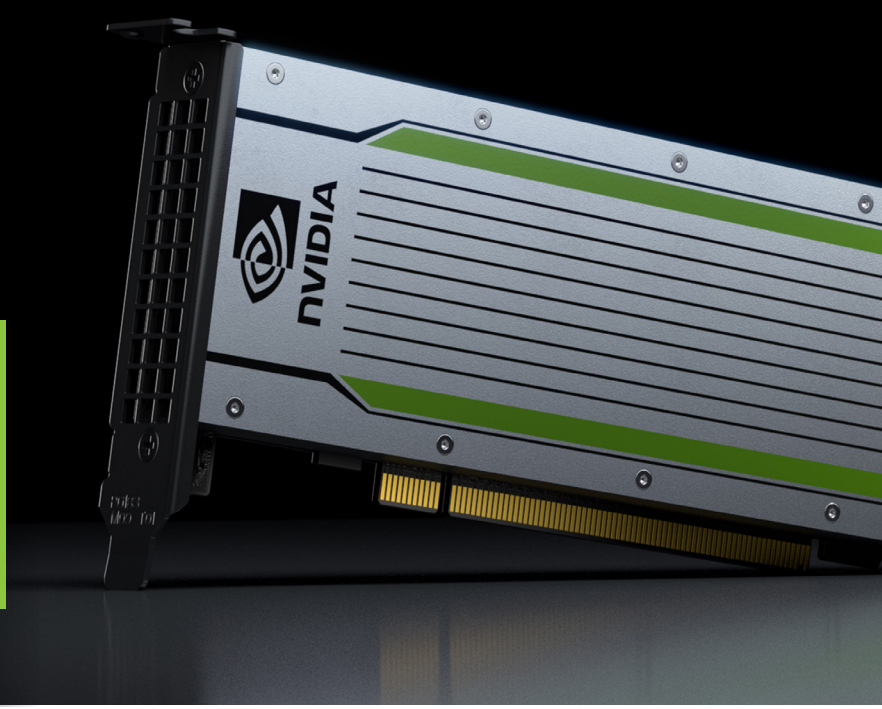
NVIDIA

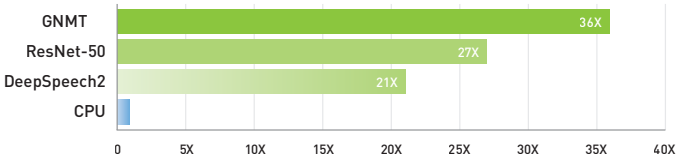
NVIDIA T4
TENSOR コア GPU



AI の大規模なトレーニングと推論を加速

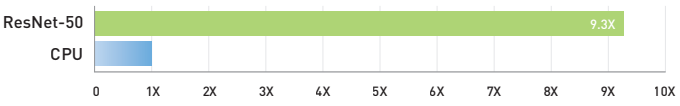
NVIDIA® T4 GPU は、スケールアウト環境のパフォーマンスを飛躍的に向上させる世界最高クラスのアクセラレーターであり、あらゆるサーバーを高速化できます。幅広い最新アプリケーションを加速できるよう、NVIDIA Turing™ Tensor コアを搭載し、革新的な多精度コンピューティングを実現しました。エネルギー効率に優れ、70 ワットの小型 PCIe フォームファクターに収まったこの最先端の GPU は、スケールアウト サーバーに最適化され、新進の AI ワークロードのために設計されています。

推論性能

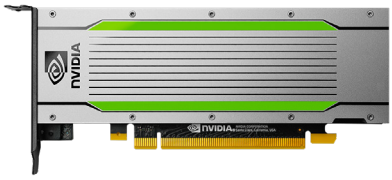


NVIDIA T4 GPU 1 基とデュアル ソケット Xeon Gold 6140 CPU 搭載サーバーの比較

トレーニング性能



NVIDIA T4 GPU 2 基とデュアル ソケット Xeon Gold 6140 CPU 搭載サーバーの比較

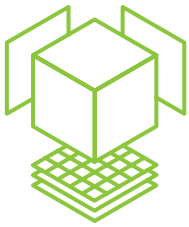


仕様

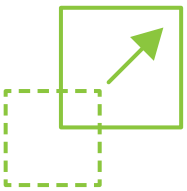
GPU アーキテクチャ	**NVIDIA Turing**
NVIDIA Turing Tensor コア	**320**
NVIDIA CUDA® コア	**2,560**
単精度演算性能	**8.1 TFLOPS**
混合精度演算性能 (FP16/FP32)	**65 TFLOPS**
INT8 精度演算性能	**130 TOPS**
INT4 精度演算性能	**260 TOPS**
GPU メモリ	**16 GB GDDR6 300 GB/秒**
ECC	**対応**
相互接続帯域幅	**32 GB/秒**
システム インターフェイス	**x16 PCIe Gen3**
フォーム ファクター	**ロープロファイル PCIe**
冷却方式	**パッシブ**
計算 API	**CUDA、NVIDIA TensorRT™、ONNX**

データセンター アクセラレーションをさらに強化する圧倒的パフォーマンス

小型のフォーム ファクターと 70 W の電力設計によって T4 は スケールアウト サーバー向けに 最適化されており、CPU の 50 倍 という驚異的なエネルギー効率が 実現されるため、運用コストを 劇的に削減できます。この 2 年間で NVIDIA の推論プラットフォームの 効率性は 10 倍以上向上し、分散 AI トレーニングおよび推論のための 最もエネルギー効率に優れた ソリューションであり続 けています。



Turing Tensor コア テクノロジと 多精度コンピューティングを採用 したことで、FP32 から FP16、 INT8、さらには INT4 の精度での 演算に画期的なパフォーマンスが 発揮され、さまざまな AI ワークロードに対応できます。 トレーニング性能は CPU と比べて 最大 9.3 倍、推論性能は最大 36 倍 に向上しています。



NVIDIA T4 データセンター GPUは、 分散コンピューティング環境に 理想的なユニバーサル アクセラレーターです。革新的な 多精度コンピューティングにより、 ディープラーニングや機械学習の トレーニングや推論、ビデオ コード変換、仮想デスクトップを 高速化します。あらゆる種類の AI フレームワークやネットワークが サポートされ、劇的に パフォーマンスと効率が向上 されることで、大規模な環境を 最大限に有効活用 できるようになります。

NVIDIA T4 の詳細については、**www.nvidia.com/T4** をご覧ください。