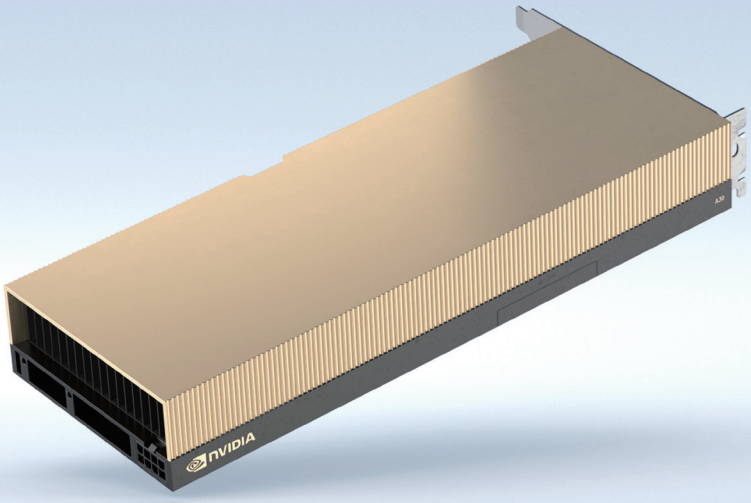
NVIDIA A30
Tensor コアGPU
メインストリーム エンタープライズ サーバー向けの汎用的な コンピューティング アクセラレーション



あらゆる企業を対象としたAI推論および メインストリーム コンピューティング

NVIDIA A30 TensorコアGPUは、AI推論およびメインストリーム エンタープライズ ワークロードのための最も汎用的なメインストリーム コンピューティングGPUです。NVIDIA Ampereアーキテクチャ Tensorコア テクノロジにより、幅広い計算精度をサポートし、単一アクセラレータであらゆるワークロードを高速化します。

大規模なAI推論向けに構築された同一のコンピューティング リソースにて、AIモデルをTF32で素早く再トレーニングし、HPC（ハイパフォーマンス コンピューティング）アプリケーションをFP64 Tensorコアを用いて高速化できます。Multi-Instance GPU（MIG）とFP64 Tensorコアは、毎秒933ギガバイト（GB/s）の高速なメモリ帯域幅と組み合わされ、165Wの低消費電力でメインストリーム サーバーに最適なPCIeカード上で動作します。

第3世代 TensorコアとMIGを組み合わせることで、多様なワークロードに対して保証されたサービス品質を提供し、柔軟性のあるデータセンターを実現する汎用的なGPUの能力を活用できます。大規模なワークロードから小規模なワークロードまで対応可能な汎用性の高いA30のコンピューティング能力は、メインストリームの企業に最大限の価値をもたらします。

A30は、ハードウェア、ネットワーク、ソフトウェア、ライブラリ、NGC™の最適化されたAIモデルとアプリケーションにわたるビルディングブロックで構成される完全なNVIDIAデータセンター ソリューションの一部です。A30はデータセンター向けの最も強力なエンドツーエンドのAI/HPCプラットフォームとして機能し、研究者は現実の成果を出し、ソリューションを大規模な運用環境へ展開できます。

システム仕様

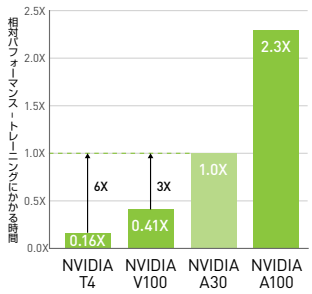
ピークFP64	**5.2 TF**
ピークFP64 Tensorコア	**10.3 TF**
ピークFP32	**10.3 TF**
TF32 Tensorコア	**82 TF | 165 TF***
BFLOAT16 Tensorコア	**165 TF | 330 TF***
ピークFP16 Tensorコア	**165 TF | 330 TF***
ピークINT8 Tensorコア	**330 TOPS | 661 TOPS***
ピークINT4 Tensorコア	**661 TOPS | 1321 TOPS***
メディアエンジン	**1 optical flow accelerator（OFA） 1 JPEGデコーダー（NVJPEG） 4ビデオ デコーダー（NVDEC）**
GPUメモリ	**24GB HBM2**
GPUメモリ帯域幅	**933GB/s**
相互接続	**PCIe Gen4: 64GB/s 第3世代NVIDIA® NVLINK® 200GB/s****
フォームファクター	**デュアルスロット、フルハイト、フルレングス（FHFL）**
最大熱設計電力（TDP）	**165W**
Multi-Instance GPU（MIG）	**6GBのMIGが4つ 12GBのMIGが2つ 24GBのMIGが1つ**
仮想GPU（vGPU）ソフトウェア サポート	**VMware 向けNVIDIA AI Enterprise NVIDIA 仮想コンピュート サーバー**

* スパース性あり

あらゆるワークロードに優れたパフォーマンス

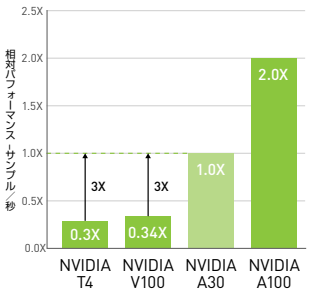
AIトレーニング – V100の最大3倍、T4の最大6倍のスループット

BERT Large Pre-Training（正規化済み）



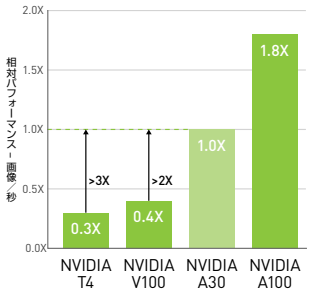
AI推論 – リアルタイム対話型AI[2]でV100の最大3倍のスループット

BERT Large推論（正規化済み）10ms以下のレイテンシでのスループット



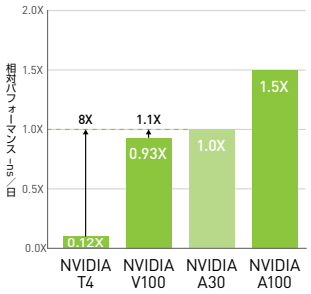
AI推論 - リアルタイム画像分類[3]でT4の3倍以上のスループット

RN50 v1.5推論（正規化済み）7ms以下のレイテンシでのスループット



HPC - V100の最大1.1倍、T4[4]の最大8倍のスループット

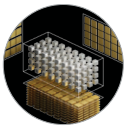
LAMMPS（正規化済み）



画期的なイノベーション

NVIDIA AMPERE アーキテクチャ

MIGでA30 GPUを小さなインスタンスに分割する場合でも、NVIDIA NVLinkで複数のGPUを接続して、より大きなワークロードを高速化する場合でも、A30は最小のジョブから最大級のマルチノード ワークロードに至る多様な規模のアクセラレーションニーズに容易に対応できます。A30 の汎用性により、ITマネージャーはメインストリーム サーバーで、データセンター内のすべてのGPUの利用率を一日中最大化できます。



第3世代 TENSORコア

NVIDIA A30は165 teraFLOPS（TFLOPS）のTF32ディープラーニング性能を発揮します。これはNVIDIA T4 Tensor コアGPUと比較して、AIトレーニングのスループットが20倍以上、推論性能が5倍以上です。HPCでは、A30はNVIDIA V100 TensorコアGPUと比較して30%近く高い 10.3 TFLOPSの性能を提供します。

次世代 NVLINK

A30のNVIDIA NVLinkは、前世代のNVLINKの2倍のスループットを提供します。2基のA30 PCIe GPUをNVLinkブリッジで接続することができ、330 TFLOPSのディープラーニング性能を発揮します。

Multi-Instance GPU（MIG）

A30は最大4つのGPUインスタンスに分割することができ、それぞれが高帯域幅のメモリ、キャッシュ、コンピューティングコアを持ち、ハードウェアレベルで完全に独立しています。MIGにより、開発者はあらゆるアプリケーションで画期的な高速化を実現できます。また、IT管理者はすべてのジョブに対して適切なサイズのGPUアクセラレーションを提供することで、利用率を最適化し、すべてのユーザーやアプリケーションへアクセスを拡大できます。

HBM2

最大24GBの高帯域メモリ（HBM2）を搭載したA30は、メインストリーム サーバーでの多様なAIやHPCワークロードに最適な933 GB/sのGPUメモリ帯域を提供します。

スパース構造

AIネットワークには数百万から数十億のパラメータが存在します。こうしたパラメータのすべてが正確な予測に必要ではなく、一部をゼロに変換することで、精度を損なうことなくモデルを「スパース」にすることができます。 A30のTensorコアは、スパースモデルに対して最大2倍高い性能を発揮します。このスパース機能はAI推論にメリットをもたらしますが、モデルトレーニングの性能を向上させることもできます。

エンタープライズ向けのエンドツーエンドソリューション

最新のデータセンターの心臓部であるNVIDIA Ampereアーキテクチャを搭載したNVIDIA A30 TensorコアGPUは、NVIDIA データセンター プラットフォームに不可欠な要素です。ディープラーニング、HPC、データ分析のために構築されたこのプラットフォームは、あらゆる主要なディープラーニング フレームワークを含む2,000以上のアプリケーションを高速化します。さらに、クラウドネイティブなエンドツーエンドのAIおよびデータ分析ソフトウェア スイートであるNVIDIA AI Enterpriseが、VMware vSphereを用いたハイパーバイザー ベースの仮想インフラストラクチャのA30上で動作することが認証されています。これにより、ハイブリッドクラウド環境でAIワークロードの管理やスケーリングが可能になります。この完全なNVIDIAプラットフォームは、データセンターからエッジに至るあらゆるところで利用でき、劇的な性能向上とコスト削減の機会の両方を提供します。

エンタープライズ向けに最適化された ソフトウェアとサービス

NVIDIA CERTIFIED

あらゆるディープラーニング フレームワーク

mxnet　PYTORCH　Apache Spark　TensorFlow

2,000以上のGPU高速化アプリケーション

- Altair nanoFluidX
- Altair ultraFluidX
- AMBER
- ANSYS Fluent
- DS SIMULIA Abaqus
- GAUSSIAN
- GROMACS
- NAMD
- OpenFOAM
- VASP
- WRF

NVIDIA A30 Tensorコア GPUに関する詳細：**www.nvidia.com/a30**

[1] BERT-Large Pre-Training（9/10エポック）フェーズ1および（1/10エポック）フェーズ2、シーケンス長：フェーズ1=128およびフェーズ2=512、データセット= real、NGC™コンテナー= 21.03、8x GPU: T4（FP32、BS=8、2）、V100 PCIE 16GB（FP32、BS=8、2）、A30（TF32、BS=8、2）、A100 PCIE 40GB（TF32、BS=54、8）。示されているバッチ サイズはそれぞれフェーズ1用とフェーズ2用

[2] NVIDIA® TensorRT®、精度= INT8、シーケンス長= 384、NGCコンテナー=20.12、レイテンシ 10ミリ秒未満、データセット=合成、1x GPU: A100 PCIe 40GB（BS=8）| A30（BS=4）| V100 SXM2 16GB（BS=1）| T4（BS=1）

[3] TensorRT、NGCコンテナー=20.12、レイテンシ　7ミリ秒未満、データセット=合成、1x　GPU:　T4　（BS=31、INT8）　|　V100（BS=43、混合精度）| A30（BS=96、INT8）| A100（BS=174、INT8）

[4] データセット：ReaxFF/C、FP64 | 4x GPU: T4、V100 PCIE 16GB、A30

NVIDIA