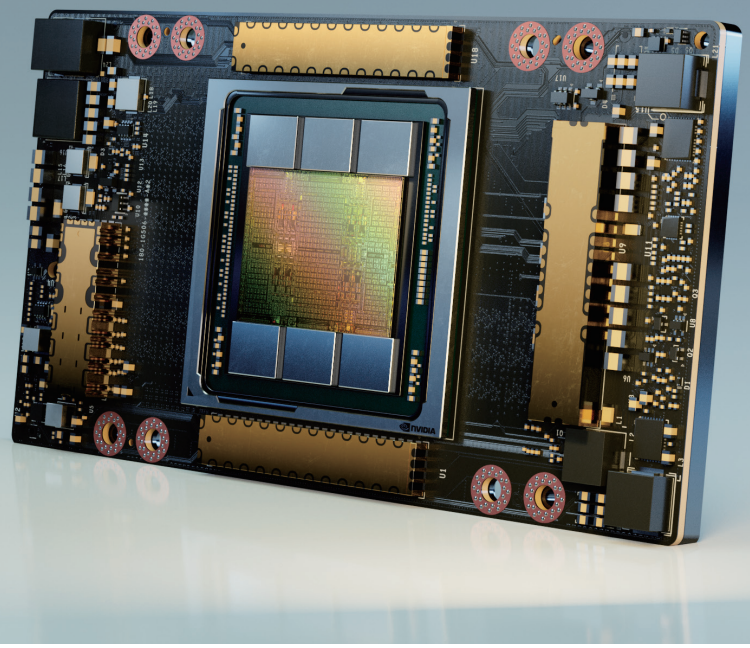
NVIDIA

# NVIDIA A100 TENSORコアGPU

## あらゆるスケールで 前例のないスケール



## あらゆるワークロード向けの 最もパワフルなコンピュートプラットフォーム

NVIDIA® A100 TensorコアGPUは、あらゆるスケールで前例のない加速を実現し、AI、データ分析、ハイパフォーマンスコンピューティング（HPC）アプリケーション向けの世界で最も高性能でエラスティックなデータセンターを強化するものです。NVIDIAデータセンタープラットフォームのエンジンとして、A100は前世代のNVIDIA Volta™よりも最大で20倍高いパフォーマンスを提供します。A100は、効率的にスケールアップしたり、7個のGPUインスタンスに分割したりできるので、マルチインスタンスGPU（MIG）により、エラスティックなデータセンターが変化するワークロードの要求に動的に適応できる統合プラットフォームを実現します。

NVIDIA A100 Tensorコアテクノロジはさまざまな計算精度をサポートしており、1基のアクセラレータであらゆるワークロードに対応します。最新世代のA100 80GBは、GPUメモリが2倍になっており、毎秒2テラバイト（TB/秒）という世界最速のメモリ帯域幅を実現し、大規模なモデルやデータセットに対して解を得るまでの時間を短縮します。

A100は、ハードウェア、ネットワーキング、ソフトウェア、ライブラリ、およびNGC™の最適化されたAIのモデルとアプリケーションにまたがるビルディングブロックによって構築される完全なNVIDIAデータセンターソリューションの一部です。データセンター向けとして最もパワフルなAIおよびHPC向けのエンドツーエンドのプラットフォームであり、研究者は現実世界で成果をもたらし、ソリューションを大規模な運用環境に展開できます。

### システム仕様

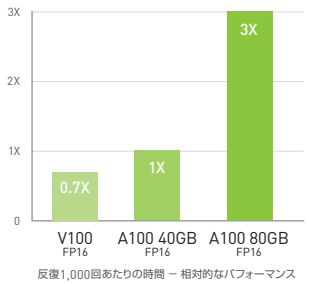
| | NVLink用 NVIDIA A100 | | PCIe用 NVIDIA A100 |
|---|---|---|---|
| ピークFP64 | 9.7 TF | | 9.7 TF |
| ピークFP64 Tensorコア | 19.5 TF | | 19.5 TF |
| ピークFP32 | 19.5 TF | | 19.5 TF |
| ピークFP32 Tensorコア | 156 TF \| 312 TF* | | 156 TF \| 312 TF* |
| ピークBFLOAT16 Tensorコア | 312 TF \| 624 TF* | | 312 TF \| 624 TF* |
| ピークFP16 Tensorコア | 312 TF \| 624 TF* | | 312 TF \| 624 TF* |
| ピークINT8 Tensorコア | 624 TOPS \| 1,248 TOPS* | | 624 TOPS \| 1,248 TOPS* |
| ピークINT4 Tensorコア | 1,248 TOPS \| 2,496 TOPS* | | 1,248 TOPS \| 2,496 TOPS* |
| GPUメモリ | 40GB | 80GB | 40GB |
| GPUメモリ帯域幅 | 1,555 GB/s | 2,039 GB/s | 1,555 GB/s |
| 相互接続 | NVIDIA NVLink 600 GB/s PCIe Gen4 64 GB/s | | NVIDIA NVLink 600 GB/s PCIe Gen4 64 GB/s |
| マルチインスタンスGPU | 最大7MIG（10GB時）のさまざまなインスタンスサイズ | | 最大7MIG（5GB時）のさまざまなインスタンスサイズ |
| フォームファクタ | 4/8 SXM（NVIDIA HGX™ A100上） | | PCIe |
| 最大TDP電力 | 400 W | 400 W | 250 W |
| 主要アプリケーション実効性能 | 100% | | 90% |

* スパース行列演算機能付き

## さまざまなワークロードでの驚異的なパフォーマンス

大規模モデルでのAIトレーニングを最大3倍高速化

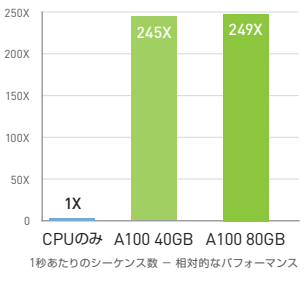
DLRMトレーニング




DLRM on HugeCTR フレームワーク、精度=FP16 | NVIDIA A100 80GBバッチサイズ=48 | NVIDIA A100 40GBバッチサイズ=32 | NVIDIA V100 32GBバッチサイズ=32

CPUと比べAI推論パフォーマンスが最大249倍高速化

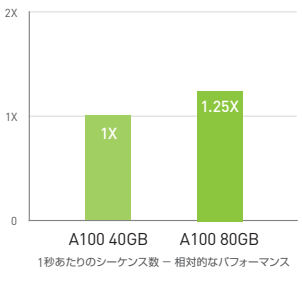
BERT大規模推論




BERT大規模推論 | CPUのみ:Dual Xeon Gold 6240 @2.60GHz、精度=FP32、バッチサイズ=128 | V100:NVIDIA Tensor-RT™ (TRT) 7.2、精度=INT8、バッチサイズ=256 | A100 40GBおよび80GB、バッチサイズ=256、精度=INT8(スパース行列演算機能付き)

A100のAI推論パフォーマンスが最大1.25倍高速化

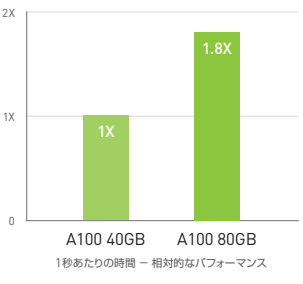
RNN-T推論：単一ストリーム




MLPerf 0.7 RNN-T:(1/7)MIGスライスで測定フレームワーク:TensorRT 7.2、データセット=LibriSpeech、精度=FP16

HPCアプリケーション向けのパフォーマンスが最大1.8倍高速化

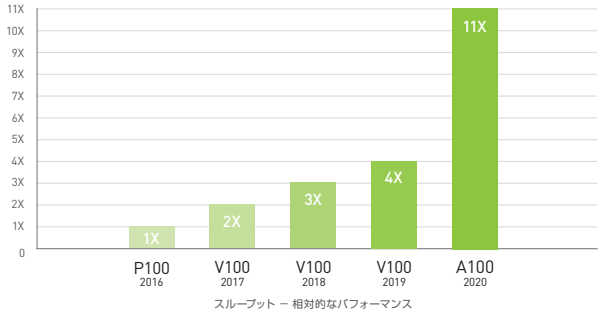
Quantum Expresso




Quantum Espresso(CNT10POR8データセットを用いて測定)、精度=FP64

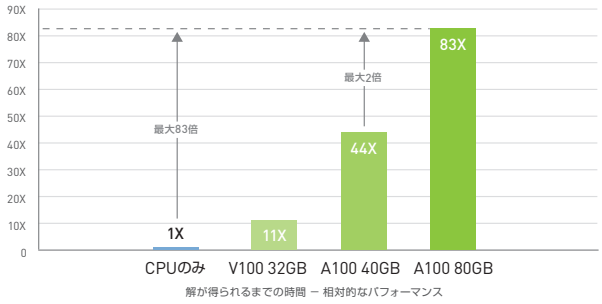
4年間でHPCパフォーマンスが11倍以上向上

上位HPCアプリケーションのスループット



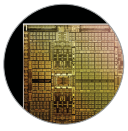

P100に対するアプリケーションの幾何平均の高速化:ベンチマークアプリケーション:Amber [PME-Cellulose_NVE]、Chroma [szscl21_24_128]、GROMACS [ADH Dodec]、MILC [Apex Medium]、NAMD [stmv_nve_cuda]、PyTorch (BERT-Large Fine Tuner)、Quantum Espresso [AUSURF112-jR]、Random Forest FP32 [make_blobs (160000 x 64:10)]、TensorFlow [ResNet-50]、VASP 6 [Si Huge] | 4倍のNVIDIA P100、V100、またはA100 GPUとデュアルソケットCPUを搭載したGPUノード

ビッグデータ分析ベンチマークでCPUより最大83倍、A100 40GBより2倍高速



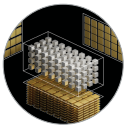

ビッグデータ分析ベンチマーク | 10TBデータセットで分析リテールクエリ30個、ETL、ML、NLP | CPU:Intel Xeon Gold 6252 2.10 GHz、Hadoop | V100 32GB、RAPIDS/Dask | A100 40GBおよびA100 80GB、RAPIDS/Dask/BlazingSQL

## 画期的なイノベーション



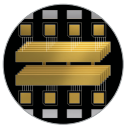
### NVIDIA AMPERE アーキテクチャ

MIGでA100 GPUを小さなインスタンスに分割する場合も、またはNVIDIA NVLink®で複数のGPUを接続して大規模なワークロードを高速化する場合も、A100は、最小のジョブから最大のマルチノードワークロードに至るまで、さまざまな規模の加速ニーズに対応することができます。A100は汎用性が高く、IT管理者がデータセンター内のすべてのGPUを一日中最大限に活用できます。



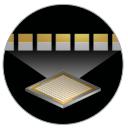
### 第3世代TENSORコア

NVIDIA A100は、312teraFLOPS(TFLOPS)のディープラーニングパフォーマンスを提供します。これは、NVIDIA Volta GPUと比較して、ディープラーニングトレーニングでは20倍のTensor FLOPS、ディープラーニング推論では20倍のTensor TOPSです。



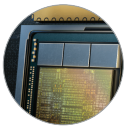
### 次世代NVLINK

A100のNVIDIA NVLinkは、前世代よりも2倍高いスループットをもたらします。NVIDIA NVSwitch™と組み合わせれば、最大16基のA100 GPUを毎秒最大600ギガバイト(GB/秒)で相互接続可能となり、単一サーバーで実現可能なアプリケーションパフォーマンスを最大限まで引き出します。NVLinkは、HGX A100サーバーボードを介したA100 SXM GPU、および最大2GPUのNVLinkブリッジを介したPCIe GPUで利用できます。



### マルチインスタンスGPU (MIG)

A100 GPUは、7つのGPUインスタンスに分割し、それぞれに独立した高帯域幅メモリ、キャッシュ、およびコンピューティングコアを割り当て、ハードウェアレベルで完全に分離できます。MIGによって、開発者がすべてのアプリケーションについて画期的なアクセラレーションを利用できるようになり、IT管理者がすべてのジョブに適切な規模のGPUアクセラレーションを提供できるので、使用率を最適化し、すべてのユーザーとアプリケーションへのアクセスを拡張することが可能です。



### HBM2E

最大80ギガバイト(GB)の高帯域幅メモリ(HBM2e)により、A100は毎秒2TBを超える世界初のGPUメモリ帯域幅と、95%という従来よりも高いダイナミックランダムアクセスメモリ(DRAM)使用効率を実現します。A100では、メモリ帯域幅が前世代の1.7倍になります。

### 構造化スパース性

AIネットワークには、数百万個から数十億個のパラメーターがあります。すべてのパラメーターが正確な予測に必要なわけではないので、一部をゼロに変換し、精度を損なうことなくモデルに「スパース」を設けることができます。A100のTensorコアは、スパースモデルについては最大2倍のパフォーマンスを提供できます。このスパース機能はAI推論に役立ちますが、モデルトレーニングのパフォーマンスを向上させることもできます。

NVIDIA A100 TensorコアGPUは、ディープラーニング、HPC、およびデータ分析用のNVIDIAデータセンタープラットフォームの主力製品です。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め、1,800を超えるアプリケーションを高速化します。A100は、デスクトップからサーバーやクラウドサービスに至るまで、どこでも使うことができ、飛躍的なパフォーマンス向上とコスト削減の両方を実現する機会を提供します。

**あらゆるディープラーニングフレームワーク**

Caffe2
Microsoft Cognitive Toolkit
mxnet
PYTORCH
Apache Spark
TensorFlow
theano

**1800を超えるGPUアクセラレーテッドアプリケーション**

- AMBER
- ANSYS Fluent
- GAUSSIAN
- GROMACS
- LS-DYNA
- NAMD
- OpenFOAM
- Simulia Abaqus
- VASP
- WRF

NVIDIA A100 TensorコアGPUについて詳しくは、**www.nvidia.com/a100** をご覧ください。

© 2020 NVIDIA Corporation. All rights reserved. NVIDIA、NVIDIAのロゴ、CUDA、DGX、HGX、HGX A100、NVLink、NVSwitch、OpenACC、TensorRT、Voltaは、NVIDIA Corporationの米国およびその他の国における商標および/または登録商標です。OpenCLは、Apple Inc.の商標であり、Khronos Group Inc.のライセンス供与のもとで使用しています。その他のすべての商標および著作権は、それぞれの所有者に帰属します。11月20日

NVIDIA