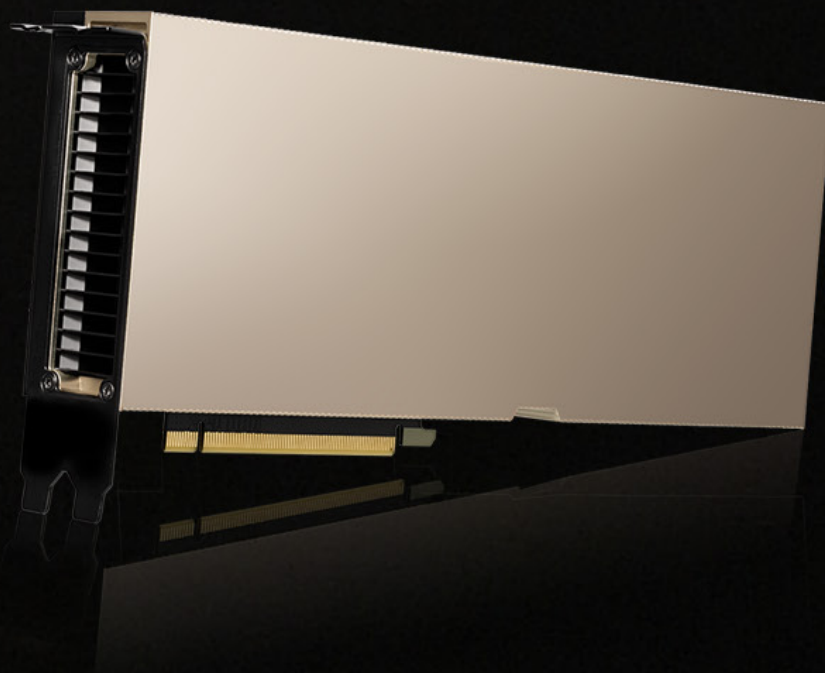


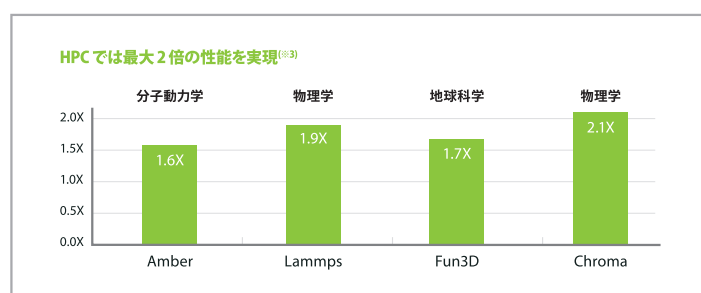
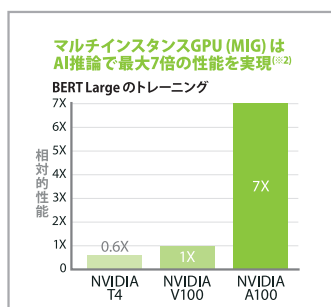
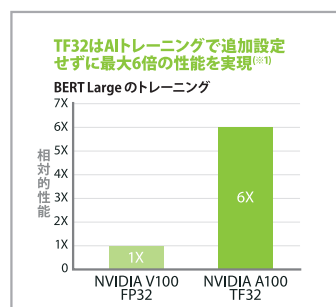


NVIDIA A100 TENSOR コア GPU



規模を問わず、前例のない加速

NVIDIA A100 TensorコアGPUは、AI、データ分析、HPCにあらゆる規模で前例のない高速化を実現し、世界で最も困難なコンピューティングの課題に取り組むことができます。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はあらゆるサイズのワークロードを加速するために、数千規模のGPUまで効率的にスケールアップすることや、新しいNVIDIA Multi-Instance GPU(MIG)テクノロジーを使用して7つの独立したGPUインスタンスに分割することができます。A100の第3世代のTensorコアテクノロジーは、多様なワークロードに対してより多くの計算精度の高速化を提供し、洞察を得るまでの時間と市場投入までの時間を短縮します。



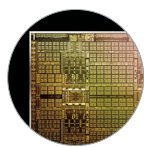
システム仕様 (ピーク性能)

PCI Express向けNVIDIA A100	
GPU アーキテクチャ	NVIDIA Ampere
倍精度演算性能	FP64: 9.7 TFLOPS FP64 Tensor Core: 19.5 TFLOPS
単精度演算性能	FP32: 19.5 TFLOPS Tensor Float 32 (TF32): 156 TFLOPS 312 TFLOPS*
半精度演算性能	312 TFLOPS 624 TFLOPS*
Bfloat16 演算性能	312 TFLOPS 624 TFLOPS*
整数演算性能	INT8: 624 TOPS 1,248 TOPS* INT4: 1,248 TOPS 2,496 TOPS*
GPU メモリ	40 GB HBM2
メモリ帯域幅	1,555 GB / 秒
誤り訂正符号	対応
インターコネクト インターフェース	PCI Express Gen4: 64 GB / 秒 NVIDIA NVLink: 600 GB / 秒
フォーム ファクタ	PCI Express
Multi-Instance GPU(MIG)	さまざまなインスタンス サイズ、 最大7つのGPUインスタンス@5GB
最大消費電力	250W
主要アプリケーション 実効性能	90%
冷却装置	パッシブ冷却
Compute API	CUDA®, DirectCompute, OpenCL™、 OpenACC®

* 構造化疎性有効時

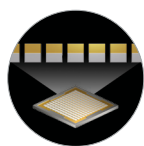
画期的なイノベーション

NVIDIA AMPEREアーキテクチャ



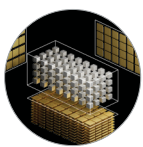
A100は、大小のワークロードを高速化します。MIGを使用してA100 GPUを小さなインスタンスに分割したり、NVLinkを使用して複数のGPUを接続して大規模なワークロードを高速化したりする場合でも、A100は最小のジョブから最大のマルチノードワークロードまで、さまざまな規模の高速化ニーズに容易に対応することができます。A100の汎用性により、IT管理者がデータセンター内のすべてのGPUを常に最大限に活用できます。

MULTI-INSTANCE GPU (MIG)



A100 GPUは、最大7つのGPUインスタンスに分割することができます。それぞれにハードウェアレベルで完全に独立した、高帯域幅メモリ、キャッシュ、コンピュートコアを備えています。MIGにより、開発者はあらゆるアプリケーションに対して画期的な高速化を実現でき、IT管理者はすべてのジョブに適したサイズのGPUアクセラレーションを提供し、利用率を最適化し、すべてのユーザとアプリケーションへのアクセスを拡大することができます。

第3世代 TENSOR コア



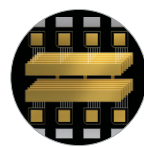
A100は312テラフロップス (TFLOPS) のディープラーニング性能を発揮します。これは、NVIDIA Volta™ GPUと比較すると、ディープラーニングのトレーニングと推論ともに20倍のTensor FLOPSを実現します。

HBM2



40ギガバイト (GB) の高帯域幅メモリ (HBM2) を備えたA100は、1.6TB/秒のメモリ帯域を提供し、ダイナミックランダムアクセスメモリ (DRAM) の利用効率を95%に高めます。A100は、前世代と比べて1.7倍のメモリ帯域を実現しています。

次世代 NVLINK



A100のNVIDIA NVLinkは、前世代と比較して2倍のスループットを実現します。NVIDIA NVSwitch™と組み合わせることで、最大16基のA100 GPUを最大600ギガバイト/秒 (GB/秒) で相互接続することができ、1台のサーバーでのアプリケーション性能を最大化します。NVLinkは、HGX A100サーバーボードを使用したA100 SXM GPU、およびNVLink Bridgeを使用して最大2基のPCIe GPUで利用できます。

構造化疎性



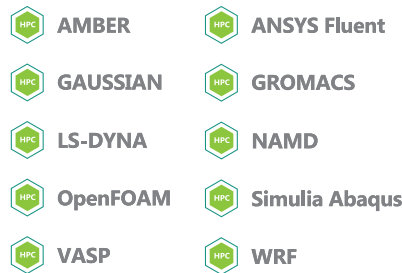
AIネットワークは数百万～数十億のパラメータを持つ大規模なものですが、これらのパラメータのすべてが精度の高い予測に必要なわけではなく、一部のパラメータを0に置き換えてモデルを「スパース化」しても精度に問題が生じることはありません。A100のTensorコアは、疎性モデルで最大2倍の性能を発揮します。疎性機能により、AI推論が容易に高速化すると同時に、モデルのトレーニングの性能も向上します。

NVIDIA A100 TensorコアGPU は、ディープラーニング、HPC、データ分析のためのNVIDIAデータセンタープラットフォームのフラッグシップ製品です。このプラットフォームは、700以上のHPCアプリケーションとあらゆる主要なディープラーニングフレームワークを高速化します。デスクトップから、サーバー、クラウドサービスまであらゆる場所で利用でき、劇的な性能向上とコスト削減を実現します。

あらゆるディープラーニングフレームワークに対応



GPUで高速化する700以上のアプリケーション



NVIDIA A100 TensorコアGPUの詳細
www.nvidia.com/ja-jp/data-center/a100

NVIDIA A100 製品ページ
www.elsa-jp.co.jp/products/detail/nvidia-a100/



※1: フェーズ1 (2/3)とフェーズ2 (1/3)から成るPyTorchを使用したBERTの事前トレーニングでのスループット | フェーズ1のシーケンス長=128、フェーズ2のシーケンス長=512 | V100: NVIDIA V100 TensorコアGPUを8基搭載したNVIDIA DGX-1™サーバーで、FP32精度を使用 | A100: A100を8基搭載したNVIDIA DGX™ A100サーバーで、TF32精度を使用
※2: BERT Largeの推論 | NVIDIA T4 TensorコアGPU: NVIDIA TensorRT™ (TRT) 7.1、精度=INT8、バッチサイズ256 | V100: TRT 7.1、精度FP16、バッチサイズ256 | MIGにより1GPUあたり5GBの7つのインスタンスに分割したA100; プリプロダクションでのTRT、バッチサイズ94、精度INT8 (スパース化)
※3: 使用したV100は1基のV100 SXM2。使用したA100は1基のA100 SXM4。AMBERはPME-Celluloseに基づく。LAMMPSはAtomic Fluid LJ-2.5を使用。FUND3Dはdpwを使用。Chromalはszscd21_24_128を使用。

お問い合わせ先

株式会社 エルザ ジャパン
<https://www.elsa-jp.co.jp/>

〒105-0014 東京都港区芝3丁目42番10号 三田UTビル
TEL: 03-5765-7391 FAX: 03-5765-7235

© 2020 NVIDIA Corporation. 無断転用禁止。NVIDIA、NVIDIA のロゴ、CUDA、DGX、HGX、A100、NVLink、NVSwitch、OpenACC、TensorRT およびVolta は、米国およびその他の国における NVIDIA Corporation の商標および/または登録商標です。OpenCL は Apple Inc. の商標であり、Khronos Group Inc. の許可を得て使用しています。その他の商標および著作権は全て各社の所有物です。ELSA (エルザ) は、テクノロジー・ジョイント株式会社の登録商標です。本カタログの掲載内容は2020年6月現在の情報です。

