Wisteria/BDEC-01 を構成する NVIDIA GPU 及びネットワーク製品の概要

エヌビディア合同会社
平野幸彦[1]
岩谷正樹[2]

1. はじめに

AI、HPC、データ分析のワークロードはますます複雑化と多様化が進み、さらなる GPU コンピューティング能力、マルチ GPU 接続の強化、強力なインターコネクト、これらをサポートする包括的なソフトウェアスタックの充実が求められています。NVIDIA は、NVIDIA Ampere GPU アーキテクチャをベースにした新たな NVIDIA A100 Tensor コア GPU と最新の CUDA ソフトウェア、NVIDIA Mellanox InfiniBand を組み合わせることで、増大し続けるコンピューティングへの期待と課題に応えます。

2. NVIDIA A100 Tensor コア GPU

（1）概 要

A100 GPU はコアアーキテクチャに多くの改良が加わり、前世代 GPU の V100 と比較しても AI、HPC、データ分析などのワークロード処理能力が大幅に高速化されています。新規スパース性機能が利用できる場合には、演算速度をさらに 2 倍にまで高めることができます。また、高帯域幅の HBM2 メモリと、大容量化かつ高速化された L2 キャッシュにより、演算を担う CUDA コアや Tensor コアに効率的にデータを転送することができます。計算精度としても、FP64、FP32、BFLOAT16、FP16、INT8、INT4 などの様々なデータタイプをサポートし、あらゆるワークロードに対応します。

新規の第 3 世代 NVLink と PCIe Gen 4 は、マルチ GPU システム構成を高速化し、ハイパースケールなデータセンターを支援します。さらに、新たな機能であるマルチインスタンス GPU (MIG) は 1 つの GPU を最大 7 つの GPU インスタンスに物理的に分割することを可能にし、次々変化するワークロードの要求に動的に適応できる柔軟性の高い統合プラットフォームを実現します。

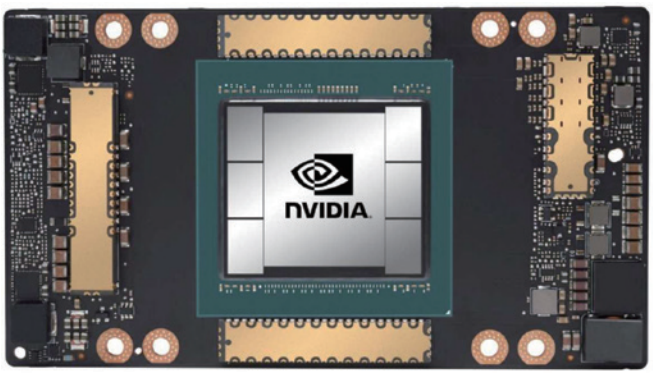
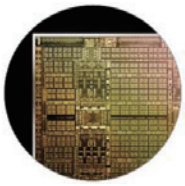


図 1　SXM4 モジュールに搭載された NVIDIA A100 Tensor コア GPU

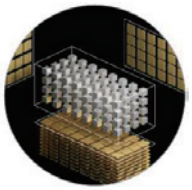
[1] プリンシパル ソリューションアーキテクト
[2] HPC/AI ネットワークプロダクトマーケティング ディレクター

（2）NVIDIA Ampere アーキテクチャの画期的なイノベーション



NVIDIA Ampere アーキテクチャ

マルチインスタンス GPU 機能(MIG) による A100 GPU の最大 7 つのインスタンスへの分割や、NVIDIA NVLink による複数 GPU の接続により、最小ジョブから大規模なマルチノードワークロードに至るまで、さまざまな規模の高速化ニーズに対応することができます。

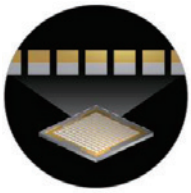


第 3 世代 Tensor コア

Tensor コアは、行列の融合積和演算 (FMA) を実行する専用の演算コアであり、AI のトレーニングと推論に画期的なパフォーマンスを提供します。A100 の Tensor コアは FP64 にも対応し、HPC でこれまでにない倍精度の演算能力を実現します。

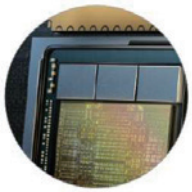
次世代 NVLINK

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。



マルチインスタンス GPU（MIG）

A100 GPU は最大 7 つのインスタンスにハードウェアレベルで分割することができ、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られます。MIG により様々なサイズのワークロードに柔軟に対応することができ、GPU の使用率を最適化できます。



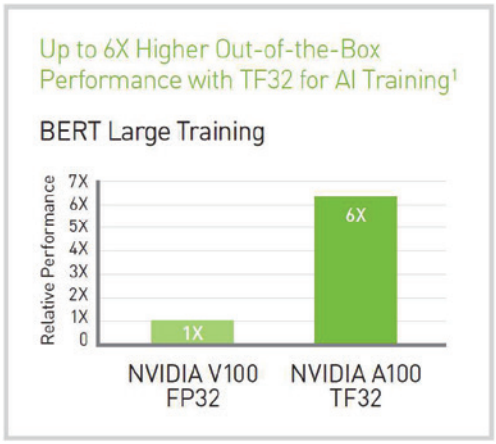
HBM2

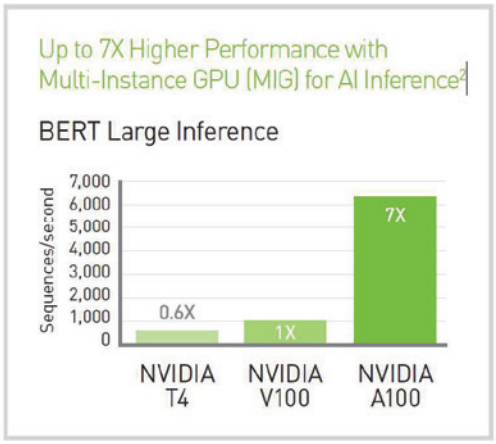
40 ギガバイト(GB) の高帯域幅メモリ（HBM2）の採用により、A100 は前世代の 1.7 倍となる毎秒 1.5TB のメモリ帯域幅を達成し、95%の高使用効率を実現します。

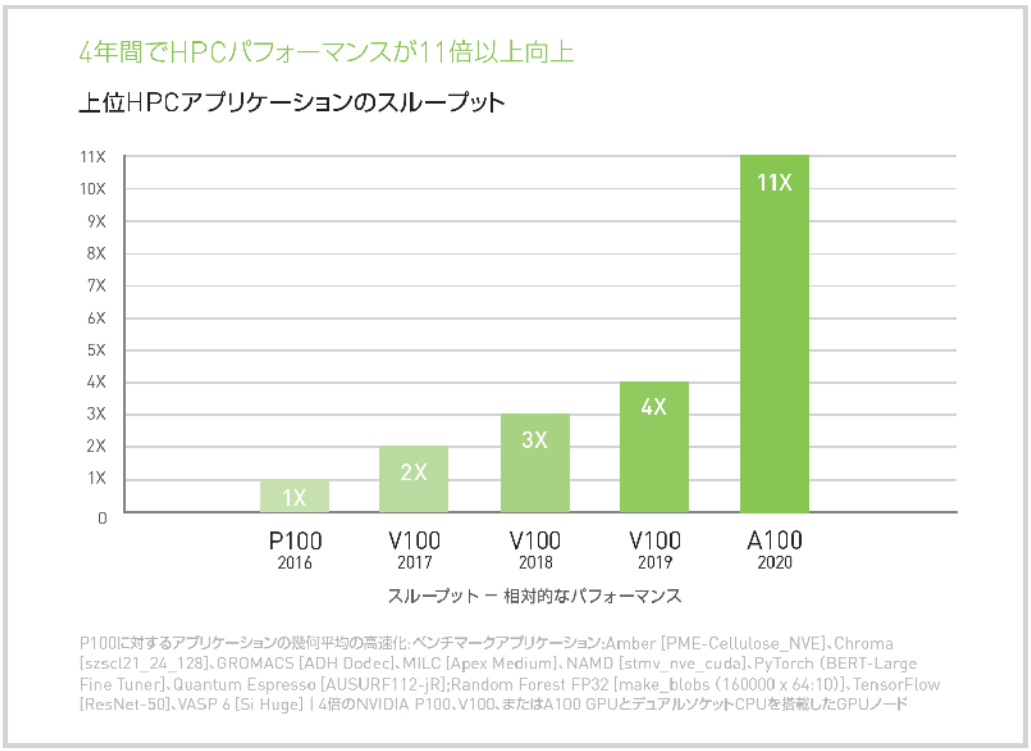
構造化スパース性

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が、全てのパラメーターが正確な予測に必要なわけではなく、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、AI 推論やモデル学習の性能向上に役立ちます。

（3）前世代 GPU と比較した NVIDIA A100 の性能向上例







参考情報

NVIDIA A100 Tensor コア GPU :

https://www.nvidia.com/ja-jp/data-center/a100/

NVIDIA A100 データシート:

https://www.nvidia.com/content/dam/en-zz/ja/Solutions/Data-Center/documents/tensor-core-gpu-nvidia-20201208_r2.pdf

NVIDIA AMPERE アーキテクチャ ホワイトペーパー :

https://www.nvidia.com/content/dam/en-zz/ja/Solutions/Data-Center/documents/nvidia-ampere-architecture-whitepaper-jp.pdf

2. NVIDIA Mellanox InfiniBand

（1）概 要

InfiniBand は、1999 年に最初に発表されたインターコネクト規格の業界標準で、コンピューティングおよびデータ集約型アプリケーションのデファクトスタンダードとなっており、 AI、HPC およびハイパースケールのクラウド環境で幅広く採用され続けています。2020 年 10 月の TOP500 スーパーコンピュータリストにおいては、その 上位 100 位に入るシステムの 60% で InfiniBand HDR が採用されており、国内の大学、研究所の HPC データセンターにおいても多数採用されています。

InfiniBand 技術は、次の 4 つの主要な基本機能に基づいています：

- InfiniBand は、 OSI 階層のトランスポート層から物理層までを一括で提供し、ネットワーク内でネットワーク機能の実行及び管理ができる非常にスマートなエンドポイントネットワークです。ネットワーク機能は、 CPU や GPU へ負荷をかけることなく実行されるため、 CPU または GPU の実アプリケーションに対する処理時間を増やすことが可能となります。InfiniBand の ASIC は、 CPU/GPU、 OS、 Kernel をバイパスしてホスト側のメモリに直接アクセスし通信する機能をサポートしているため、 RDMA 通信、 GPUDirect RDMA および GPUDirect Storage など、非常に効果的かつ効率的な方法で CPU、 GPU、Storage 間で通信を行うことが可能です。
- InfiniBand は、最初からスケールすることを考慮された完全なソフトウェア デファインド ネットワーク (SDN) に基づいたスイッチングネットワークです。ネットワーク経路、ネットワークに接続されたデバイス（スイッチ、エンドノード）を集中管理し、オペレーティングシステムをスイッチ内で実行するため、全てのスイッチに専用の外部管理機構を必要としません（Ethernet スイッチの場合、ネットワーク全体管理をする場合には必要になります)。 これにより、 InfiniBand は Ethernet や他の独自のネットワークと比較して、コストパフォーマンスに優れたネットワークファブリックになります。また、In-Network Computing などの独自の技術革新を可能にし、ネットワーク経由で転送されるデータの処理が可能です。重要な例として、 Scalable Hierarchical Aggregation and Reduction Protocol (SHARP) ™ 技術があります。この技術により、シミュレーション解析、 MPI による並列計算、及び深層学習アプリケーションフレームワークのパフォーマンスが大幅に向上することが実証されています。
- InfiniBand は、ネットワークを 1 つのサブネットマネージャが集中的に管理、制御、運用することが可能です。 1 つのサブネットで設定をすることにより、多種多様なネットワークトポロジーに対応し、対象となるアプリケーションに応じてデータセンターネットワークをカスタマイズおよび最適化することが可能となります。InfiniBand は、Ethernet のようにネットワークのさまざまな部分に対して個別のスイッチ構成を作成する必要はなく、複数の複雑なネットワークアルゴリズムを処理する必要もありません。

InfiniBand は、 Ethernet よりもより通信パフォーマンスを向上させるとともに、Ethernet よりも OPEX を削減することが可能です。

- InfiniBand は、InfiniBand Trade Association (IBTA) による国際標準化されたネットワーク規格であり、下位及び後方互換性を保証します。また、管理ソフトウェアはオープン API を使用したオープンソースを使用しています。

上記の 4 点から、 InfiniBand は、非常にコスト効率が高く、管理が容易で最高の通信パフォーマンスを達成出来るネットワークであると言えます。

(2) NVIDIA Mellanox InfiniBand HDR

HDR 200Gb／秒 InfiniBand 製品ラインは、 2019 年から生産されており、IBTA HDR InfiniBand 仕様をサポートし、世界中の多くの主要なスーパーコンピュータ、ディープラーニングプラットフォーム、クラウドデータセンターで使用されています。

ConnectX-6 InfiniBand HCA

ConnectX-6 InfiniBand HCA は、200Gb／秒の単方向リンク速度 (または 400Gb ／秒の相方向) で、最大 2 ポートの独立したポートを提供します。エンドツーエンドで低いアプリケーションレイテンシーを保ち、毎秒 2 億 1,500 万件のメッセージ処理が可能です。ConnectX-6 は PCIe Gen3、 Gen4 をサポートしています。

NVIDIA Mellanox Quantum HDR 200 Gb／秒 InfiniBand スイッチ

NVIDIA Mellanox Quantum HDR InfiniBand スイッチは、レイヤ 2 とレイヤ 3 に対応しており、レイテンシーは 130ns (HDR モード) と非常に小さく、SHARP v2、 Congestion Control、 Adaptive Routing、 SHIELD 機能をサポートします。 1U サイズのスイッチで、 200Gb／秒 のスループット性能を持つポートを 40 ポート備えています。

(3) Aquarius のインターコネクト

WisteriaBDEC-01 の データ･学習ノード群である Aquarius は､各ノードが転送速度 200Gb／秒 の InfiniBand HDR の 4 リンクを用いたフルバイセクションバンド幅を持つ相互接続網にて結合されています。これにより、演算加速装置 1 基当たり 10 GB／秒以上の高速なインジェクションバンド幅を有し、高速な通信性能を得ることが可能となります。

(4) Odyssey と Aquarius の相互接続

WisteriaBDEC-01 は、A64FX アーキテクチャを採用した FX1000 にて構成される Odyssey と、x86 アーキテクチャを持つ汎用 CPU と GPU によって構成された Aquarius との間を InfiniBand EDR を用いた 2.0TB／秒のネットワークバンド幅で接続しており、NVIDIA は異機種で構成された両システム間の高速通信を提供し、「計算・データ・学習」融合を目指す本シ

ステムの実現をサポートしています。

Odyssey と Aquarius を跨いで MPI による通信を用いたプログラムを実行することはできませんが、東京大学情報基盤センターにおいて開発中である両システム間ライブラリや制御機構を含んだソフトウェアである h3-Open-BDEC[3] により、両システムのより深い融合が目指されています。特に、ヘテロジニアス環境下での異なるコンポーネント間のデータ受け渡しのためのライブラリである h3-Open-SYS/WaitIO には、異なる機種の計算機間での MPI 通信を実現するための機能が含まれており、WisteriaBDEC-01 では、InfiniBand が提供する高い性能を活用した本ライブラリを用いた通信が行われています。

NVIDIA Mellanox InfiniBand は、本ライブラリをはじめとする様々なソフトウェアにより構成される h3-Open-BDEC が目指す「計算＋データ＋学習」の融合を目指したヘテロジニアスシステムの能力を最大限に引き出し、最小の計算量・消費電力での計算を実行するためのプラットフォームとして貢献しています。

[3] http://nkl.cc.u-tokyo.ac.jp/h3-Open-BDEC/