NVIDIA、HGX-2 を発表、HPC と AI コンピューティングを単一のアーキテクチャに融合

クラウドサーバー プラットフォーム HGX-2 が混合精度のワークロードを加速；
2 ペタフロップスの処理能力により AI 用途での記録的なパフォーマンスを実現

台北―GTC Taiwan―2018 年 5 月 30 日―NVIDIA は本日、コンピューティング プラットフォーム NVIDIA HGX-2™ を発表しました。これは、人工知能（AI）とハイパフォーマンス コンピューティング (HPC) の両方に向けた単一のコンピューティング プラットフォームとして初となる製品です。

混合精度の計算能力を備えたクラウドサーバー プラットフォーム HGX-2 は、比類のない柔軟性を提供し、コンピューティングの未来を後押しします。HGX-2 により、科学技術計算およびシミュレーションに向けた、倍精度浮動小数点 (FP64) や単精度浮動小数点 (FP32) での高精度な演算が可能となります。また、AI のトレーニングや推論のためには、半精度浮動小数点 (FP16) や整数 (Int8) を用いることもできます。HPC と AI を組み合わせたアプリケーションが増加の一途をたどっている状況ですが、HGX-2 のかつてない多用途性により、こうしたアプリケーションからの要求も満たされることになります。

NVIDIA HGX-2 プラットフォームベースのシステムを市場に導入するため、今日の業界をリードする数多くのコンピューター メーカーが、互いに連携しながら計画を立てました。

NVIDIA の創業者兼 CEO であるジェンスン・フアン (Jensen Huang) は、本日から開催されている GPU テクノロジ カンファレンス Taiwan で講演を行い、その中で次のように述べています。「コンピューティングの世界は変わりました。CPU のスケーリングが減速しているにもかかわらず、今やコンピューティングへの需要はとどまるところを知らないのです。Tensor コア GPU を備えた NVIDIA の HGX-2 は、この業界に強力で多用途のコンピューティング プラットフォームを提供します。このプラットフォームは、HPC と AI を融合し、世界における重要課題を解決するものです。」

HPC と AI に向けた最先端のシステムを作り上げるメーカーにとって、HGX-2 は「構成要素」の役割を果たすものです。HGX-2 は AI のトレーニングに関するベンチマーク ResNet-50 において、1 秒あたり 15,500 点の画像速度を実現し、トレーニングにおける記録を打ち立てました。また HGX-2 は、CPU のみで構成されたサーバーであれば 300 台まで置き換えることが可能です。

HGX-2 には インターコネクト ファブリック である NVIDIA NVSwitch™ などの画期的な機能が搭載されています。この機能は、16 基の NVIDIA Tesla® V100 Tensor コア GPU をシームレスにつなぎ合わ

せ、2 ペタフロップスの AI パフォーマンスを実現する 1 つの巨大な GPU として動作させるというものです。HGX-2 を用いて作られた初のシステムが、先日発表された NVIDIA DGX-2™ でした。

HGX-2 は、元となった NVIDIA HGX-1 が Computex 2017 に出展されてから 1 年を経て発表されました。リファレンス アーキテクチャ HGX-1 は世界有数のサーバー メーカーや、大規模なデータセンターを運用している企業に広く採用されました。こういった企業の中には、Amazon Web Services、Facebook、Microsoft が含まれています。

OEM、ODM によるシステムも年内に予定

業界をリードするサーバー メーカーである Lenovo、QCT、Supermicro、Wiwynn の 4 社は、HGX-2 をベースにした自社製のシステムを、年内に上市する計画を発表しました。

さらに、世界トップレベルの ODM メーカーである Foxconn、Inventec、Quanta、Wistron の 4 社が HGX-2 ベースのシステムを設計中です。これらもやはり年内上市が予定されており、世界最大級のクラウド データセンターでの利用が、その目的とされています。

NVIDIA の GPU アクセラレーテッド サーバー プラットフォーム製品群

HGX-2 は NVIDIA GPU アクセラレーテッド サーバー プラットフォーム製品群の 1 つです。このエコシステムは、AI、HPC、アクセラレーテッド コンピューティングによる広範なワークロードを最適なパフォーマンスで処理できる条件を満たした、サーバー クラスの製品によって構成されています。

大手サーバー メーカーからのサポートを受けたこのプラットフォームは、GPU と CPU の最適な混合および相互接続を提供することで、データセンターのサーバー エコシステムと協調します。これによって、多様なトレーニング (HGX-T2)、推論 (HGX-I2)、スーパーコンピューティング (SCX) といったアプリケーションが実現します。お客様は、サーバー プラットフォームを個別に選択することで、アクセラレーテッド コンピューティングによる複合的なワークロードに対応し、クラス内最高レベルのパフォーマンスを達成できるようになります。

広範な産業をサポート

トップクラスの OEM および ODM メーカーが、HGX-2 への強力な支持を表明しています。

「Foxconn は長年、ハイパースケール コンピューティング ソリューションへの注力を続けており、お客様からの評価も獲得することができました。Foxconn は、NVIDIA と共同で HGX-2 プロジェクトに取り組め

ることを嬉しく思っています。HGX-2 は、人工知能／ディープラーニングに対する爆発的な需要を満たすソリューションの中で、もっとも期待できるものであると言えるでしょう」

— エド・ウー (Ed Wu) 氏、Foxconn 社コーポレート エグゼクティブ バイスプレジデント兼 Ingrasys 社会長

「Inventec には、高いパフォーマンスとスケーラビリティを兼ね備えた、堅牢で革新的な設計のサーバーを、世界有数のデータセンターを運用するお客様に提供してきたという確かな実績があります。Inventec は、今後の製品設計へ HGX-2 を早急に組み込むことで、世界中の企業が利用できる最も強力な AI ソリューションを、自社の製品ラインナップに取り入れることになるでしょう」
— エヴァン・チェン (Evan Chien) 氏、IEC White Box Product Center 代表兼 Inventec 社中国事業部門ディレクター

「NVIDIA の HGX-2 は、AI や HPC に集中したワークロードに対し 2 ペタフロップスものパフォーマンスを提供可能にする設計により、この分野で超えねばならないハードルを上げました。規模を拡大しパフォーマンスも最高にしたいというお客様からのニーズは高まっていますが、サーバーの構成要素である HGX-2 を用いれば、こういったニーズを満たすことができる新システムも迅速に開発可能となるでしょう」

— ポール・ジュ (Paul Ju) 氏、Lenovo DCG バイスプレジデント兼ゼネラル マネージャー

「クラウドの実現におけるリーダー的企業として、Quanta は様々な革新的ユース ケースに向けた次世代クラウド ソリューションの開発に力を入れています。AI アプリケーションの大幅な増加を受け、Quanta は NVIDIA と緊密に連携し、お客様が最新かつ最高の GPU テクノロジから利益を得られることを保証しています。HGX-2 発売時のパートナーとして、AI クラウドを実現するこの重要な企業と共同で GPU 演算製品ラインナップを拡大できることに、私たちは興奮を覚えているところです」
— マイク・ヤン (Mike Yang) 氏、Quanta Computer 社バイスプレジデント兼 QCT 社社長

「AI モデルのサイズは急速に拡大中であり、トレーニングのために数週間を要することもあります。こうした状況への対応に役立てるため、Supermicro は HGX-2 プラットフォームをベースにしたクラウド サーバーを開発中です。HGX-2 システムにより、複雑なモデルのトレーニングでも効率的にできるようになるでしょう」

— チャールズ・リアン (Charles Liang) 氏、Supermicro 社社長兼 CEO

「NVIDIA のパートナーとして共に働けることを、大変光栄に思います。今日、最新のテクノロジ環境において、AI クラウド コンピューティングへの需要が台頭しています。HGX-2 システムの高いパフォーマンスとモジュール方式による柔軟性は、研究目的、科学用途から政府による利用まで、様々なコンピューティング領域に間違いなく大きく貢献するでしょう。」

― ジェフ・リン (Jeff Lin) 氏、Wistron 社 Enterprise Business Group 代表

「Wiwynn はハイパースケール データセンターとクラウド インフラストラクチャ ソリューションの提供を専門とする企業です。NVIDIA、そしてサーバー構成要素である HGX-2 とのコラボレーションにより、AI や HPC による計算能力集約型のワークロードに向けた、2 ペタフロップスのコンピューティングをお客様に提供できるでしょう」

― スティーブン・リュー (Steven Lu) 氏、Wiwynn 社バイスプレジデント

NVIDIA について

NVIDIA が 1999 年に開発した GPU は、PC ゲーム市場の成長に拍車をかけ、現代のコンピューターグラフィックスを再定義し、並列コンピューティングを一変させました。最近では、GPU ディープラーニングが最新の AI、つまりコンピューティングの新時代の火付け役となり、世界を認知して理解できるコンピューター、ロボット、自動運転車の脳の役割を GPU が果たすまでになりました。今日、NVIDIA は「AI コンピューティングカンパニー」として知名度を上げています。詳しい情報は、http://www.nvidia.co.jp/ をご覧ください。

NVIDIA についての最新情報:

公式ブログ NVIDIA blog、Facebook、Google+、Twitter、LinkedIn、Instagram、NVIDIA に関する動画 YouTube、画像 Flickr

本件に関するお問い合わせ先:

エヌビディア　広報/マーケティングコミュニケーションズ
中村かおり　Email アドレス：knakamura@nvidia.com　TEL: 03-6743-8712
吉川香葉子　Email アドレス：kyoshikawa@nvidia.com TEL: 080-8891-3352

エヌビディア広報事務局
株式会社イニシャル　東山・石井・大迫
Email アドレス：　nvidia@vectorinc.co.jp
Tel：　03-5572-7306　Fax：　03-5572-6065

クラウドサーバー プラットフォーム NVIDIA HGX-2 のメリット、影響、パフォーマンスおよび能力、HGX-2 が柔軟性と多用途性を提供してコンピューティングの未来および HPC と AI を統合するアプリケーションの要求を満たす性能を後押しすること、コンピューター メーカーが市場に HGX-2 プラットフォームをベースにしたシステムを導入する計画、コンピューティングへの需要がとどまるところを知らないにもかかわらず CPU のスケーリング速度が低下していること、HGX-2 が業界に強力で多用途なプラットフォームを提供して世界における重要課題を解決すること、HGX-2 がメーカーにとって構成要素の役割を果たし HPC と AI に

向けた最先端のシステムを作り出しかつ CPU のみで構成されるサーバーであれば 300 台まで置き換える性能を持つこと、HGX-1 が世界有数のサーバー メーカーおよびデータセンターを運用する企業により広く採用されていること、リーダー的なサーバー メーカーおよびトップレベルの ODM メーカーによる HGX-2 をベースとするシステムを年内に市場へ導入する計画、NVIDIA GPU アクセラレーテッド サーバー プラットフォームのメリット、パフォーマンス、性能、HGX-2 が人工知能／ディープラーニングに対する需要を満たすための最も期待できるソリューションであること、HGX-2 が Inventec の今後の設計に組み込まれ Inventec の製品ラインナップに利用できる最も強力な AI ソリューションを取り入れること、顧客の高まるニーズを満たすため Lenovo のシステムを手助けする HGX-2 の性能、HGX-2 システムが複雑なモデルの効率的なトレーニングを実現すること、AI クラウド コンピューティングに対する需要が今日最新のテクノロジ環境において台頭していること、様々なコンピューティング領域に多大な貢献を成す HGX-2 の性能、Wiwynn が AI および HPC のワークロードに対し 2 ペタフロップスのコンピューティングを顧客に提供するのを HGX-2 が実現することなど、本プレスリリースにおける一定の記載は将来の見通しに関する記述であり、予測とは著しく異なる結果を生ずる可能性があるリスクと不確実性を伴っています。