# NVIDIA、世界をリードするディープラーニング コンピューティング プラットフォームを強化
# 6 か月でパフォーマンスを 10 倍に

メモリが 2 倍になった最新 *Tesla V100 32GB GPU*、
革新的な *NVSwitch* ファブリック、包括的なソフトウェア スタックでパフォーマンスの向上を実現
*DGX-2* が初の 2 ペタフロップ ディープラーニング システムに

**カリフォルニア州サンノゼ — GPU テクノロジ カンファレンス — (2018 年 3 月 27 日)** — NVIDIA は本日、世界をリードする同社のディープラーニング コンピューティング プラットフォームに対する一連の重要な機能強化を発表しました。これにより、ディープラーニング ワークロードにおいて、6 か月前の前世代に比べ 10 倍のパフォーマンスを実現します。

すべての主要クラウド サービス プロバイダーとサーバー メーカーに採用されている NVIDIA プラットフォームへの主な機能強化には、もっともパワフルなデータセンター GPU である NVIDIA® Tesla® V100 のメモリが 2 倍になったほか、革新的な最新 GPU インターコネクト ファブリックである NVIDIA NVSwitch™ が含まれます。NVSwitch によって、最大 16 基の Tesla V100 GPU を使って 1 秒間に 2.4 テラバイトという記録的な速度で同時通信できるようになります。また、完全に最適化された最新のソフトウェア スタックも発表しました。

他にも、NVIDIA は、2 ペタフロップスもの計算能力を提供できる初の単一サーバーである NVIDIA DGX-2™ によって、ディープラーニング コンピューティングにおける重要なブレークスルーを打ち出しました。DGX-2 は 15 ラックを占める 300 台のサーバーに匹敵するディープラーニング処理能力を備えつつ、サイズはその 60 分の 1 で、電力効率は 18 倍です。。

NVIDIA の創業者兼 CEO であるジェンスン・フアン (Jensen Huang) は、GTC 2018 でこのニュースを発表した際、次のように述べています。「このディープラーニングの驚異的な機能強化は、今後さらに発表される機能のほんの一部でしかありません。これらの機能強化の多くは、またたく間に世界標準となった NVIDIA のディープラーニング プラットフォームに基づいています。当社はムーアの法則を大きく上回るペースでプラットフォームのパフォーマンスを劇的に向上させており、医療、輸送、科学探査、その他の数え切れない分野での変革を後押しするブレークスルーをもたらします。」

**Tesla V100 のメモリが 2 倍に**

Tesla V100 GPU は、世界の先端を行く研究者に広く採用されています。そしてこのたび、メモリインテンシブなディープラーニング ワークロードやハイパフォーマンス コンピューティング ワークロードを処理できるよう、メモリが 2 倍になりました。

32 GB のメモリが搭載された Tesla V100 GPU を使用すれば、データ サイエンティストは、より深くより大規模なディープラーニング モデルのトレーニングをこれまで以上に正確に行えるようになります。また、メモリに制約のある HPC アプリケーションでも、以前の 16 GB 版と比べてパフォーマンスを最大 50% 向上させることができます。

Tesla V100 32GB GPU は、NVIDIA DGX システム ポートフォリオ全体ですぐに利用できます。また、主要コンピューター メーカーである Cray、Hewlett Packard Enterprise、IBM、Lenovo、Supermicro、および Tyan が、第 2 四半期中に新しい Tesla V100 32GB 搭載システムのロールアウトを開始すると発表しました。他にも、Oracle Cloud Infrastructure が、今年後半にクラウドで Tesla V100 32GB の提供を開始する計画を明らかにしています。

**NVSwitch: 革新的なインターコネクト ファブリック**

NVSwitch は、最高の PCIe スイッチの 5 倍の帯域幅を実現できます。これにより、開発者はより多くの GPU が相互にハイパーコネクトされたシステムを構築して、従来のシステムの限界を破り、はるかに大規模なデータセットを実行できるようになるでしょう。また、ニューラルネットワークの並列トレーニングのモデル化など、より大規模で複雑なワークロードを扱えるようになります。

NVSwitch は、NVIDIA が開発した初の高速インターコネクト テクノロジである NVIDIA NVLink™ によって提供されるイノベーションを拡張します。NVSwitch を使用すると、システム設計者は、NVLink ベースの GPU の任意のトポロジを柔軟に接続できる、より一層高度なシステムを構築することが可能になります。

**最先端の GPU アクセラレーテッド ディープラーニングと HPC ソフトウェア スタック**

NVIDIA のディープラーニングと HPC ソフトウェア スタックの更新は、無償で NVIDIA の開発者コミュニティに提供されます。このコミュニティの総登録ユーザー数は、1 年前の 48 万人から、現在では 82 万人を超えるまでになっています。

その更新には、NVIDIA CUDA®、TensorRT、NCCL と cuDNN、およびロボット工学向けの新しい Isaac ソフトウェア開発者キットが含まれます。加えて、大手クラウド サービス プロバイダーとの緊密な連

携により、すべての主要なディープラーニング フレームワークが NVIDIA の GPU コンピューティング プラットフォームを十分に活用できるよう絶えず最適化されます。

**NVIDIA DGX-2: 世界初の 2 ペタフロップ システム**

NVIDIA の最新 DGX-2 システムが 2 テラフロップのマイルストーンを達成しました。これは、コンピューティング スタックのあらゆるレベルで NVIDIA が開発した幅広い業界最先端テクノロジによって実現したものです。

DGX-2 は NVSwitch をサポートする初のシステムで、NVSwitch によってシステムの 16 基すべての GPU がユニファイド メモリ領域を共有できるようになります。開発者は、最大規模のデータセットときわめて複雑なディープラーニング モデルに対処できるディープラーニング トレーニング能力を手にすることができます。

NVIDIA の完全に最適化された最新のディープラーニング ソフトウェア スイートと統合された DGX-2 は、ディープラーニングの研究とコンピューティングの限界を押し広げようとするデータ サイエンティストのために構築されました。

DGX-2 は、最先端のニューラル機械翻訳モデル FAIRSeq のトレーニングを 2 日未満で終えることができます。これは、昨年 9 月に発表された Volta 搭載の DGX-1 から 10 倍のパフォーマンス向上になります。

**Tesla V100 32GB の業界サポート**

Microsoft の音声と言語のテクニカル フェロー兼責任者であるゼドン・ファン (Xuedong Huang) 氏は、次のように述べています。「Microsoft と NVIDIA は、AI テクノロジにおける長年の協業において多大な進歩を遂げました。たとえば最近では、中国語-英語翻訳でのブレークスルーが挙げられます。新しい Tesla V100 32GB GPU によって、両社はより大規模で複雑な AI モデルのトレーニングを短期間で行えるようになるでしょう。その結果、音声認識と機械翻訳におけるモデルの精度を人間の能力にまで高め、Cortana、Bing、Microsoft Translator などのサービスの機能強化を図ることができるはずです。」

イスラエルの SAP イノベーション センター担当バイス プレジデントであるマイケル・ケメルマカー (Michael Kemelmakher) 氏は、次のように述べています。「当社は SAP Brand Impact アプリケーション用に DGX-1 と新しい Tesla V100 32GB を組み合わせて評価しました。SAP Brand Impact アプリケーションは、動画内のブランドの露出頻度をほぼリアルタイムで自動的に分析します。メモリが追加されたことで、より大規模な ResNet-152 モデルの、より高解像度の画像を処理できるようになり、エラー率が平

均 40% 低下しました。これにより、精度の高い監査可能なサービスをタイムリーかつ大規模に提供できるようになりました。」

**NVIDIA DGX 製品ポートフォリオ**

DGX-2 は、データ サイエンティストが新しいディープラーニング モデルやイノベーションの開発、テスト、展開、拡張をすばやく行えるようにするため設計された 3 つのシステムから成る NVIDIA DGX 製品ポートフォリオの最新の追加コンポーネントです。

16 基の GPU を搭載した DGX-2 はそのラインナップの最上位に位置し、Tesla V100 GPU を 8 基搭載した NVIDIA DGX-1 システムと、コンパクトなデスクサイド デザインで 4 つの Tesla V100 GPU を搭載した世界初のパーソナル ディープラーニング スーパーコンピューターである DGX Station™ に加わります。これらのシステムにより、データ サイエンティストは、デスクで行っている複雑な実験から、最大規模のディープラーニングの問題へと研究を拡大し、そのライフワークを実現できるでしょう。

詳細な技術仕様や注文フォームなど、詳しくは https://nvda.ws/2IRilLe をご覧ください。

**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

NVIDIA Tesla V100 GPU、NVIDIA NVSwitch、最新のソフトウェア スタック、NVIDIA DGX-2、NVIDIA DGX-1、および NVIDIA DGX Station の利点、影響、性能、機能に関する記述、ディープラーニングの進化とその結果実現するブレークスルーの意味、利点、影響に関する記述、主要なコンピューター メーカーによる Tesla V100 システムの今後の利用に関する記述、および NVIDIA DGX-2 の可用性に関する記述を含め (ただし、これらに限定されません)、本プレス リリースに記載されている記述の中には、将来予測的なものが含まれており、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。かかるリスクと不確実性は、世界的な経済環境、サードパーティに依存する製品の製造・組立・梱包・試験、技術開発および競合による影響、新しい製品やテクノロジの開発あるいは既存の製品やテクノロジの改良、当社製品やパートナー企業の製品の市場への浸透、デザイン・製造あるいはソフトウェアの欠陥、ユーザーの嗜好および需要の変化、業界標準やインターフェイスの変更、システム統合時に当社製品および技術の予期せぬパフォーマンスにより生じる損失などを含み、その他のリスクの詳細に関しては、2018 年 1 月 28 日を末日とする会計期間の Form 10-K レポートなど、米証券取引委員会 (SEC) に提出されている NVIDIA の報告書に適宜記載されます。SEC への提出書類は写しが NVIDIA のウェブサイトに掲載されており、NVIDIA から無償で入手することができます。これらの将来予測的な記述は発表日時点の見解に基づくものであって将来的な業績を保証するものではなく、法律による定めがある場合を除き、今後発生する事態や環境の変化に応じてこれらの記述を更新する義務を NVIDIA は一切負いません。