技術概要

GPU 対応データセンターのスケーリングに関する検討事項

最先端の AI データセンターでディープラーニング ワークロードを実行するための新基準とベスト プラクティス

NVIDIA

目次

概要

今日のエンタープライズおよびハイパースケールのデータセンターは、大量のデータで高付加な計算を行うディープ ニューラル ネットワーク (DNN) や人工知能 (AI) を使用したワークロードを念頭に構築されることが増えています。要求されるコンピューティング レベルがきわめて高く、GPU がもたらすメリットは多大です。大規模な並列処理能力を備え、高いメモリ帯域幅に最適化されている GPU を使用することで、AI クラスの行列積と分析によってすばやくインサイトを獲得できます。水冷システムやホット アイル コンテインメントなどの高度な冷却技術と、高密度、高電力のラックによって GPU サーバーをサポートすれば、フロア スペースを大幅に削減できます。また、高付加なワークロードを高い効率性とパフォーマンスで実行するため、全体的な電力も抑えられます。このドキュメントでは、ラック レイアウト、システムおよびネットワーク アーキテクチャ、ストレージなどの電源、冷却装置に焦点を当て、「GPU 対応」のデータセンター構築に関するベスト プラクティスを紹介します。ディープラーニング用 NVIDIA® DGX-1™ システムおよび NVIDIA® Tesla® V100 GPU アクセラレータでの高負荷ワークロードのシナリオに基づいて、コストを最小限に抑える方法や、高負荷ワークロードのスケーリング時にデータセンターを NVIDIA GPU に最適化するためのヒントを紹介します。

AI データセンターの新基準

今日の一般的なデータセンターでは、CPU ソケット数が 1 から 2 のサーバーで汎用のワークロードを実行しています。しかし、よりすばやくインサイトを得るため、人工知能 (AI) ワークロードを使用した新しいコンピューティングの基準が普及しつつあります。このようなワークロードで高負荷なディープ ニューラル ネットワーク (DNN) を使用するためには、GPU テクノロジに対応した新型のサーバーを使用してパフォーマンスを大幅に向上させる必要があります。このドキュメントでは、効率とパフォーマンスの最適化のために、GPU ベースのテクノロジを活用したデータセンターの設計、展開、管理、監視を中心に説明します。

AI やディープラーニング ワークロードには、DNN トレーニングと DNN 推論の 2 つの運用モードがあります。GPU ベースのサーバーで DNN ワークロードを実行すると多くのメリットがあります。たとえば、サーバーあたりおよびワット数あたりの性能が大幅に向上し、ラック数が減ることでデータセンター全体の消費電力も減少します。

高密度 GPU サーバー 1 台の性能は CPU ベース サーバー数十台分に匹敵します。以下のグラフ[1]は、GPU と CPU で同等の標準ワークロードに必要なサーバー ラック数を示しています。グラフ 1 は AI 研究のワークロードの例です。27 台の NVIDIA DGX-1 ラック (666 KW) のパフォーマンスは、CPU のみのシステムで構成された 478 台のラック (12,054 KW) と同等です。グラフ 2 は AI バッチ生産の例で、34 台の NVIDIA DGX-1 ラック (656 KW) が 1,602 台の CPU のみのラック (34,944 KW) と同等です。グラフ 3 は混合ワークロードの例で、30 台の NVIDIA DGX-1 ラック (648 KW) が 1,119 台の CPU のみのサーバー (24,752 KW) と同等です。AI/ディープラーニングおよび HPC のワークロードが同じ量だとすれば、CPU のみのデータセンターと比較して、GPU 対応データセンターの専有面積は 1/40、電力は 1/20 に抑えられます。

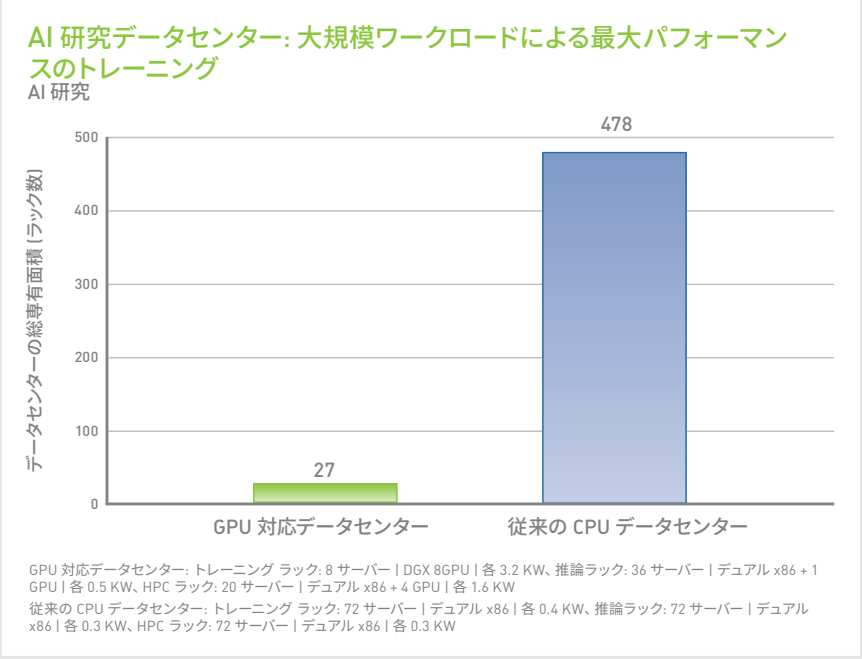


図 1

グラフ 1: 高密度のコンピューティング リソースを使用して AI トレーニングとアルゴリズム開発を集中的に行う研究用 AI GPU 対応データセンターの例。データ準備と AI 推論に特化したリソースを提供します。

1. NVIDIA パフォーマンス ラボ

グラフ 2: 大規模な生産環境において主に AI 推論に焦点を当てた生産 AI 推論 GPU 対応データセンターの例。データ準備と AI トレーニングに特化したリソースを提供します。

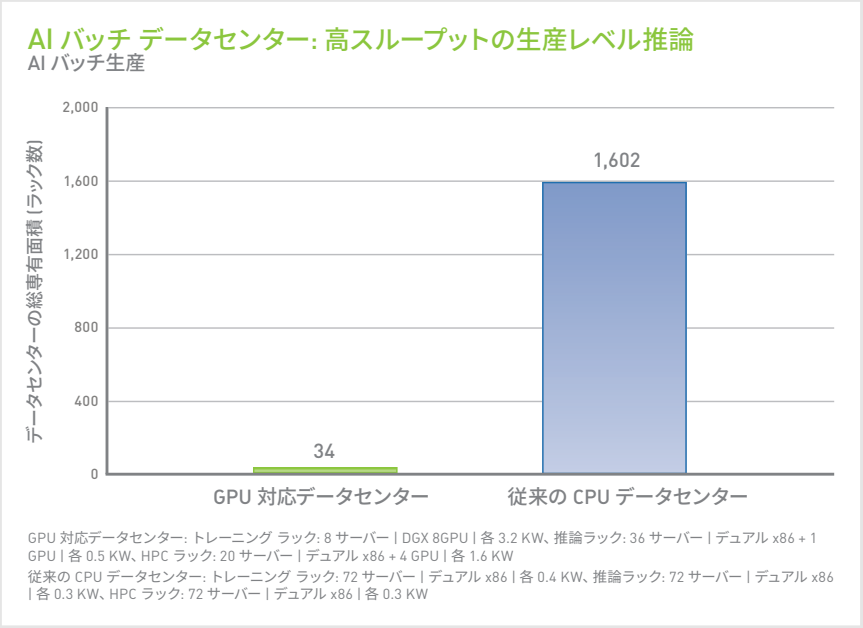


図 2

グラフ 3: 混合ワークロード用に設計された生産 AI GPU 対応データセンターの例。AI バッチまたはインタラクティブ研究と生産運用の組み合わせで、AI トレーニング、推論、およびコンピューティング リソースを混合して使用します。

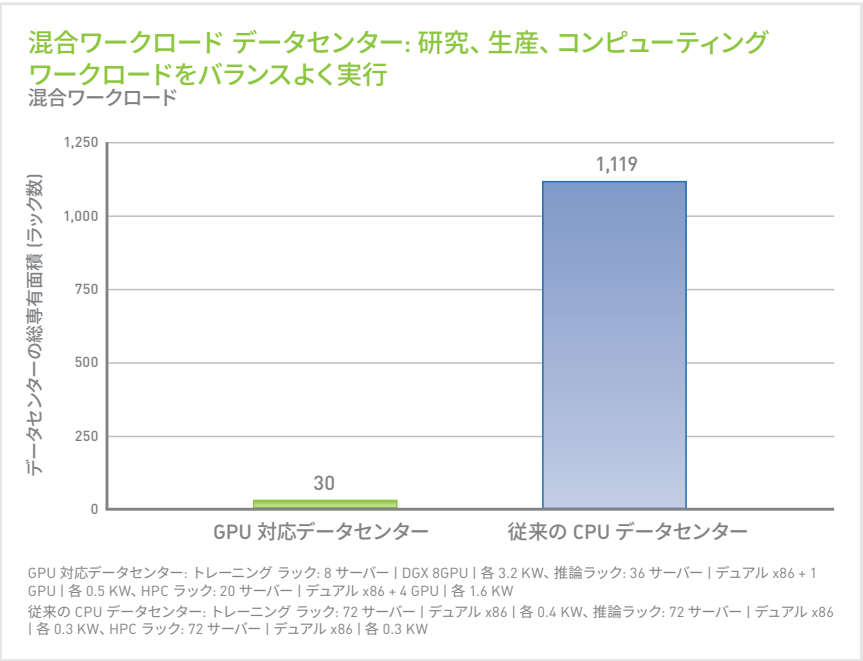


図 3

GPU ベースのサーバーでは、大規模マルチノード システムの総所有コスト (TCO) も大幅に削減できます。表 1 は、NVIDIA DGX-1 システムと 250 台の CPU ベース サーバーとの比較です。ここでは、サーバー、ネットワーク (10 Gigabit Ethernet および InfiniBand)、消費電力、コロケーション[2]、システム管理を含めた、3 年間の TCO を示しています。設置面積に対する密度の高さによって物理的なインフラストラクチャを大幅に削減できる NVIDIA DGX-1 は、従来の CPU ベース システムよりも TCO に優れています。

2. https://en.wikipedia.org/wiki/Colocation_centre (英語)

表 1: NVIDIA DGX-1 と CPU サーバーの TCO 比較

CPU ソリューションは 2RU デュアル ソケット CPU サーバー (1 台 10,000 ドル)、48 ポート 10 GBe スイッチ、36 ポート IB スイッチ、総専有面積 13 ラック、1.5 データセンター PUE で約 318 kW の総電力 (0.085 ドル/kWh) を想定。NVIDIA DGX-1 の年間システム管理コストは CPU 環境 (専属管理者 1 人の年間賃金 250,000 ドル) の 25%。

	NVIDIA® DGX-1™ システム (1 サーバー)	CPU サーバー環境 (250 サーバー)
先行支出		
サーバー (OTP)	$149,000	$2,500,000
ネットワークとケーブル (OTP)	$16,280	$187,600
定期的な運用支出		
消費電力 (3 年)	$7,153	$710,835
コロケーション (3 年)	$43,200	$1,774,800
システム管理運用コスト (3 年)	$187,500	$750,000
サポートと保守 (3 年)	$63,698	$1,125,000
3 年間の総コスト	**$466,831**	**$7,048,235**

表 1

きわめて高密度なコンピューティング能力を備えた NVIDIA GPU ベース サーバーでは、このように 3 年間の支出金額を削減できます。また、ラックごとに 15 ～ 32 kW の電力および冷却が必要です。これは、従来のデータセンターの 5 ～ 10 kW という平均値よりも高くなります。たとえば、OCP (Open Compute Project[3]) には、ワークロードに特化したサーバーやカスタム ラック設計がいくつか定義されています。

OCP V2 ラック[4] は、6.6 kW 電源を 2 つ備え、消費電力は 1 ラック 13 kW に制限されます。これにより、ラックごとに使用可能な高密度 GPU サーバーは 4 台のみとなり、高密度のメリットが失われることになります。[5] コンピューティング需要の急増により、Facebook Prineville などのハイパースケール データセンターは、10,000 (2008 年) 台のサーバーから 30,000 (2009 年)、60,000 台 (現在) に増え、307,000 平方フィート (約 28,500 平方メートル) あった面積はさらに 487,000 平方フィート (約 45,000 平方メートル) も拡大しています (合計でサッカー場 13 面分)。現在の成長は密度ではなくスペースの増加によるものです。これに比例して、ネットワーク全体の投資も 30.2 億ドル (2014 年) から 36.3 億ドル (2015 年) に増加しています。[6] 高密度サーバーでコンピューティング能力を向上させることで、このフロア スペースとネットワークの要件を大幅に削減することができます。

あらゆる規模や種類のデータセンターにも同じことが言えます。IDC のデータセンターおよびクラウド担当バイス プレジデントの Rick Villars 氏は次のように述べています。「一般的なエンタープライズ データセンターは、1 ラックあたりの電力が 8 KW 未満として構成されていますが、高密度設計を採用した最先端のクラウド サービス プロバイダーの 1 ラックあたりの電力は 12 KW 近くになります。リアルタイム分析やコグニティブ ワークロードが大幅に増える次世代のデータセンターでは、コンピューティング プールに高密度 GPU の能力を加える必要があり、7 1 ラックあたり 30 KW 程度の電力が必要になるでしょう」[7]"

3. http://www.opencompute.org/ (英語)

4. OCP V2 電源棚仕様

5. http://www.DataCenterdynamics.com/content-tracks/open-data-center/ocp-summit-facebook-refreshes-its-servers/97937.article (英語)

6. http://www.DataCenterknowledge.com/the-facebook-data-center-faq/ (英語)

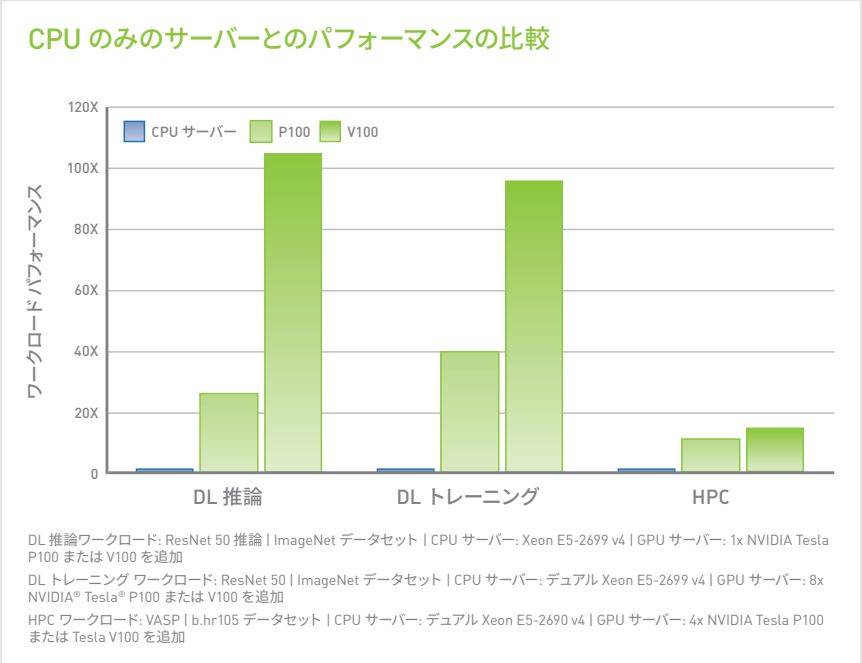
7. Rick Villars (IDC データセンターおよびクラウド担当バイス プレジデント)

GPU を利用したスケーリングの考え方

ディープ ニューラル ネットワーク (DNN) は、数千のレイヤー、数十万のニューロン、数百万の接続を扱う、今日の AI アプリケーションの中核です。AI モデルは、ギガバイトまたはテラバイト級のデータと反復計算で大規模な DNN をトレーニングし、最も正確な重み付けを導き出すことで高いパフォーマンスを実現します。AI クラスの行列乗算と畳み込みを実行ですために、GPU は強化された大規模な並列コンピューティングと最適化された高メモリ帯域幅で AI を駆動します。GPU システムのパフォーマンスは、通常の CPU のみのシステムよりも格段に優れていますが、密度や電力の要件も高くなります。

また、よりすばやく正確に大規模な AI モデルでインサイトを得るには、複数 GPU を利用したシステム パフォーマンスが必要になります。AI などの高負荷ワークロードを複数のサーバーにスケーリングするために、多数のサーバーにまたがって最小のボトルネックでアプリケーションを実行し、高いパフォーマンスを保証する必要があります。従来の軽量な CPU のみのサーバーで行うスケーリングとは対照的に、GPU システムの最大のメリットは、複数台の GPU サーバーへスケーリングする前に、コンピューティング負荷の高い複数 GPU を搭載した単体サーバーを利用できる点です。高密度 GPU サーバーは、複数のサーバーでのディープラーニング トレーニング ワークロードにも理想的なデータセンター構成要素です。

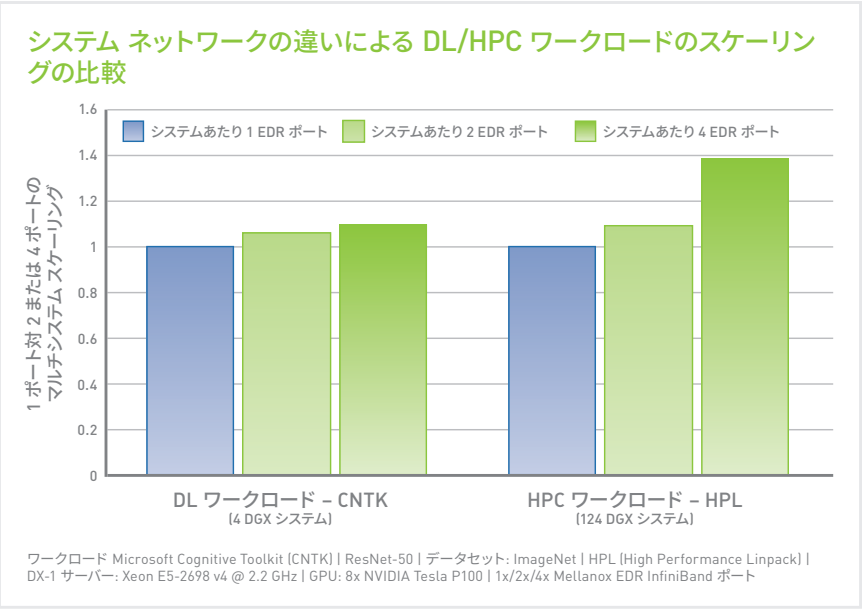
グラフ 4: GPU ベースのシステムでは、CPU のみのシステムと比較して、AI および HPC のワークロードのパフォーマンスが大幅に向上し、データセンター内の設置面積が減少するため、各ラックのパフォーマンス密度が上昇するため、システム数が減り、大規模ワークロードのスケーリング効率が大幅に向上します。



グラフ 4

また、大規模な AI 演算要素の計算でスケーラブルなパフォーマンスを得るには、システム間の強力なネットワークが必要です。以下のグラフ 5 は、システムあたりのネットワーク ポート数に対するパフォーマンスの比較、各システムで異なる数のネットワーク接続を使用する場合の違いを示しています。

グラフ 5: 高性能の GPU システムを含む設計では、ネットワーク インターコネクトが非常に重要であり、マルチノード パフォーマンスにも大きな影響を与える可能性があります。各ノード 4 つの InfiniBand リンクを使用する DGX-1 システム ベースのクラスターは、各システムで 1 つの InfiniBand リンクを使用する同等システムと比較して、DL ワークロードで 20%、HPC ワークロードでは 40% もパフォーマンスが向上します。



グラフ 5

グラフ 5 は、CNTK のディープラーニング トレーニング (マルチノード CNTK、ResNet50)[8] および HPL の計算科学 HPC ワークロードの結果を示しています。どちらのパフォーマンスも高負荷な計算と高速通信を必要とするため、1 システムあたりのネットワーク ポート数が少なくなるとマルチシステムのパフォーマンスが失われます。[9] このグラフは、システムあたり 4 ポートを使用すると、HPC ではほぼ 40%、CNTK ディープラーニング ワークロードの場合は 10% パフォーマンスが向上しています。10% という値は小さいように感じますが、これにかかるコストは、適切なインターコネクトの実装にかかるシステム コスト全体の 10% よりも格段に低くなります。コンピューティングやストレージ アクセスのボトルネックを解消して、全体的により安定した性能が提供されます。

GPU に対応した最新データセンターのベスト プラクティス

電力と冷却

大規模インフラストラクチャの問題を解決するには、計算、電力、冷却の密度を検討する必要があります。最近の冷却方式は、データセンターの高密度を活用しワットまたは単価あたりの性能が向上しています。主な方式は以下のとおりです。

- ホットまたはコールド アイル コンテインメント
- 水冷熱交換器付き後部ドア
- コンポーネント レベルでの水冷

これらの高度な冷却技術のメリットは、必要な電力とフロア スペースを最小化し、GPU サーバーのパフォーマンス効率を向上できる点です。以下の表 2 では、冷却方式を比較し、GPU サーバーと高度な冷却システムの併用によるフロア スペースへのメリットを示しています。

8. NVIDIA パフォーマンス ラボ

9. NVIDIA パフォーマンス ラボ

表 2: GPU 対応サーバー構成の例

冷却方式	冷却タイプ	ラック電力	方式のサイズ
従来の空冷	空気	8kW	52 ラック
ホット/コールド アイル コンテイ ンメント	空気	15 kW	28 ラック
熱交換器付き後部ドア	空気 + 水	35 kW	12 ラック
直接水冷	水	60 kW	7 ラック

表 2

AI やデータ集中型のワークロードが多い今日のデータセンターでは、GPU サーバーに対するさまざまなニーズが発生した場合に、ワークロードの種類に応じてさらなる最適化を行うことができます。以下の表 3 は、DNN、分析、HPC のワークロードに焦点を当て、対応する電力、ラック、冷却方式を使用した GPU サーバー構成の例を示しています。DNN トレーニング用の GPU サーバー (サーバーあたり 8 GPU の NVIDIA® DGX-1™) では超高密度の GPU サーバーとラックが適しています。冷却システムを適切に構成すれば、フロア スペースやケーブルの要件を大幅に軽減できるようになります。

表 3: GPU 対応サーバー構成の例

コンピューティング ラック		DNN トレーニング/バッチ推論	DNN リアルタイム ビデオ推論	データ分析	HPC
サーバー モデルの例		3u 8 GPU システム - NVIDIA DGX-1	1/2u 1 GPU システム	4u 8 GPU システム	1u 4 GPU システム
コンピューティング	CPU:	2 ハイエンド	2 ローからミドル	2 ハイエンド	2 ハイエンド
	GPU:	8x NVIDIA Tesla® V100	1x NVIDIA Tesla V100/低電力	8x NVIDIA Tesla V100	4x NVIDIA Tesla V100
システム メモリ		512-1,024 GB	128-256 GB	512-1,024 GB	256-512 GB
ネットワーク	内蔵:	NVIDIA NVLink™	PCIe	NVLINK	PCIe
	マルチノード:	100 GB InfiniBand	10 GB Ethernet	25 GB Ethernet	100 GB Ethernet
サーバー/ラック		4 から 8	36 から 72	4 から 10	10 から 20
サーバーあたり電力 (W)		3,200	500	2,400	1,500
ラックあたり電力 (KW)		32	18	32	15-30
低密度ラック		4 サーバー 12.8 KW 1,340 CFM	36 サーバー 18.0 KW 1,320 CFM	4 サーバー 9.6 KW 1,000 CFM	12 サーバー 18 KW 1,890 CFM
- 冷却方式		空気 – パーティション 水冷 RDHX	空気 - オープン アイル	空気 - パーティション 水冷 RDHX 直接水冷	空気 - オープン アイル
高密度ラック		8 サーバー 25.6 KW 2,630 CFM	72 サーバー 36 KW 2,650 CFM	8 サーバー 19.2 KW 2,050 CFM	24 サーバー 36 KW 3,800 CFM
- 冷却方式		空気 – パーティション 水冷 RDHX 直接水冷	空気 – パーティション 水冷 RDHX	空気 – パーティション 水冷 RDHX 直接水冷	空気 – パーティション 水冷 RDHX

表 3

表 3 では、GPU サーバーと従来の CPU データセンターの特徴的な違いが見られます。

- GPU ベースのサーバーでは、高いパフォーマンスを維持するために、各サーバーに高いエア フローが必要です。ラックを通り抜けるエア フローは、空気量と温度差がきわめて重要です。高温の空気がデータセンター内に滞留して温度が上がるのを防ぐため、装置間にすきまがないようにし、データセンター内のエア フローが確実に冷却装置に戻るように慎重に設計する必要があります。

- 高電力密度のラックでは、電力と冷却がラックだけでなくサーバー全体にバランスよく分配されるように注意する必要があります。電力消費のピーク時に負荷が高すぎて問題が発生しないように、ラック内のピーク時電力負荷の特性を確認することが重要です。複数のノードとサーバー間で電力を適切に負荷分散して、予期しないサージ電力によってノードに障害が発生しないようにする必要があります。複数の 208 V/三相/60 A または 415 V/240 V/三相/30 A 電源回路を使用する高密度電力ラック (32 kW から 50 ～ 60 kW まで) が理想です。さらに、電圧が高い方が安定性と効率に優れ、運用コストも低くなります。
- また、後部ドア冷却、コンポーネント レベルの水冷システム、液浸なども利用することができます。水冷システムは、最大で空冷システムの 3,500 倍の熱伝導能力[10] があります。また、コンポーネント レベルの水冷システムは、サーバーから 60 ～ 80% の熱を逃がすと同時にコストを 50% 抑えることができ、密度が 2 倍から 5 倍向上します。ラック レベルで水冷方式熱交換冷却システムを使用する場合でも、ラックから排出された高温の空気がサーバーの前面に循環しないようにすることが重要です。
- サーバーのエア フロー要件は、100 cfm/kW のサーバー負荷 + 5% のオーバーヘッド (空気漏れと短路循環) という「経験則」による数値を使用して計算します。この熱遮断は合計 105 cfm/kW のサーバー負荷になります (cfm: 立方フィート/分)。

GPU サーバーの HGX サーバー リファレンス アーキテクチャ

急速に進む GPU テクノロジのイノベーションにより、サーバーのアーキテクチャの重要性が高まっています。NVIDIA HGX-1 のハイパースケールな GPU アクセラレータ アーキテクチャ[11] は、世界中のクラウド サービス プロバイダー大手企業に導入されており、NVIDIA DGX-1 を含む多くのエンタープライズ クラスの GPU サーバーにその要素が組み込まれています。これらのプラットフォームは、AI やデータ分析ワークロードで業界最高のパフォーマンスを発揮するように最適化されています。HGX リファレンス アーキテクチャには、以下の項目が含まれています。

- GPU、CPU、ネットワーク、ストレージのインターコネクトに使用される PCIe および NVLink トポロジ
- CPU と GPU の比率
- システム メモリの容量
- ローカル ストレージ (SSD や NVME)

NVIDIA DGX-1 は、ディープラーニング ソフトウェアのパフォーマンスを最適化するために開発された NVIDIA 初のプラットフォームであり、最高レベルのパフォーマンスを保証します。NCCL (NVIDIA Common Collectives Library) などの NVIDIA ソフトウェア ライブラリは、HGX リファレンス アーキテクチャの PCI および NVLink トポロジ向けに最適化されています。

10. http://www.pge.com/includes/docs/pdfs/mybusiness/energysavingsrebates/incentivesbyindustry/DataCenters_BestPractices.pdf (英語)

11. https://www.nvidia.com/en-us/data-center/hgx-1/

サーバー内の GPU 密度を最大化することで、ディープラーニング トレーニング、データ分析、データベース、高性能コンピューティングなどの GPU アクセラレーション アプリケーションのパフォーマンスを最高レベルに引き上げます。ほとんどの GPU アクセラレーション アプリケーションは、CPU、メモリ、ネットワーク、ローカル ストレージを正しく構成することで、サーバーあたり 8 GPU まで適切にスケーリングされます。MXNet、TensorFlow、Caffe2、Microsoft Cognitive Toolkit などのディープラーニング フレームワークはいずれも、8 GPU まで適切にスケーリングされます。ただし、2 または 4 GPU 以上のスケーリングに最適化されていない一部の HPC アプリケーションのワークロードは、8 GPU 未満に設定する方がよい場合もあります。

NVIDIA HGX リファレンス設計の性能のバランスは、以下の条件で保証されます。

- 十分な CPU 性能。一般的に、2 基のハイエンド x86 CPU は 8 基の GPU の性能に匹敵します。
- GPU メモリの 2 倍以上のシステム メモリ。ディープラーニング トレーニング用には 4 倍が最適です。一般的に、GPU アクセラレーション データ分析およびデータベースは、サーバーに多くのシステム メモリを構成するほど有利です。
- RDMA をサポートする 100 GB ネットワーク インターフェイス カード (NIC)。分散型またはマルチノードのディープラーニング トレーニングの場合は、GPU 2 基ごとに、1 つ以上を構成して使用し、GPU と同じ PCIe スイッチ上に置く必要があります。
- ネットワーク トポロジ。システム内では NVLink を通して GPU 間の GPUDirect ピア ツー ピア転送をサポートし、システム間では InfiniBand を通して GPU 間の GPUDirect RDMA をサポートします。
- SSD および NVME ローカル ストレージ。GPU と同じまたはできるだけ近い PCIe スイッチ上に構成します。

コンピューティング ネットワーク推奨事項

複数サーバーへのスケーリングを行うには、高帯域幅、低レイテンシ、高効率の通信ネットワークが必要です。データセンターの構築時には、これらのコンピューティング ネットワーク用に 100 GB Ethernet、EDR (100 GB) または HDR (200 GB) InfiniBand[12] を使用することを検討してください。

Ethernet ネットワークで InfiniBand と同等のパフォーマンスと効率性を実現するには、以下の方法があります。

- Ethernet アダプターによる CPU 負荷を最小限に抑えるために、TCP オフロードをサポートするアダプターを使用します。
- カットスルー通信をサポートしている Ethernet スイッチ アーキテクチャを採用します。

12. http://www.mellanox.com/pdf/whitepapers/IB_Intro_WP_190.pdf (英語)

- パフォーマンスと転送効率を最大化するために、RDMA (Remote Direct Access Memory) をサポートするネットワーク アダプター使用します。
- リンクの輻輳によるボトルネックを最小限に抑えるために、スパイン リーフ トポロジ、大容量アップリンクを使用し、スイッチの数を極力減らしてレイヤー 2 ネットワークを作成します。スパイン リーフ[13] トポロジを使用してネットワークを設計すると、2 分割帯域幅の広いネットワークをコスト効率の高い方法で構築できます。これは、効率的な分散アプリケーション スケーリングの特性です。
- ルーティングによるボトルネックを最小限に抑えるために、レイヤー 3 ネットワークの数を極力減らします。
- スケーラブルなアプリケーションを実行するシステムでは、トラフィックの局所化を検討します。

マルチサーバー GPU パフォーマンスを最大限に引き出すため、InfiniBand は、高速コンピューティング用マルチサーバー アプリケーションをサポートする特別な設計になっています。InfiniBand は、複数のノードにアプリケーションをスケーリングすることを目的とした高帯域幅および低遅延の通信に関する業界標準です。これは、小規模なクラスター (20 ノード未満) からきわめて大規模なクラスター (数千ノード) までのノード間接続に使用されるテクノロジとして、HPC コミュニティで普及しています。InfiniBand ネットワークを設計する際は、以下のオプションを検討してください。

- 完全なファットツリー ネットワークを使用して、ネットワーク全体のクラスター帯域幅を最大化します。
- ノードごとに複数の InfiniBand 接続を使用して、高密度 GPU ノードのパフォーマンスを最大化します。

マルチサーバー スケーリングのパフォーマンスを実現するには、ノード内の GPU 間のトラフィックと複数サーバー間のトラフィックの帯域幅のバランスを取ることがきわめて重要です。以下の表 4 は、2 つのマルチサーバー システムを比較しています。

表 4: 異なる高速インターコネクトによるマルチノード計算コードのパフォーマンス比較

8 サーバー システムの例	サーバー	ネットワーク テクノロジ	各サーバーの受送信帯域幅	総マルチサーバー帯域幅[14] (8 サーバー)	各ソリューションの相対的アプリケーション パフォーマンス[15]
NVIDIA DGX-1 GPU サーバー 8 台、内部 GPU 間帯域幅 160 GB/秒		10 GB Ethernet (システムあたり 1 ポート)	システムあたり 2 GB/秒	合計 16 GB/秒	1X
		100 GB EDR InfiniBand (システムあたり 4 ポート)	システムあたり 47 GB/秒	合計 376 GB/秒	2X

表 4

表 4 の各システム内の GPU 間の内部帯域幅は 160 GB/秒であるため、ノード内外の通信のバランスを取ることがきわめて重要です。EDR ソリューションのオフ ノード帯域幅は 47 GB/秒ですが、これは 10 GB Ethernet ベース ソリューションの 20 倍です。さらに、これは高速コンピューティング ワークロードに最適なバランスで、実際のマルチサーバー アプリケーションではパフォーマンスが 2 倍になります。

13. http://www.cisco.com/c/en/us/products/collateral/switches/nexus-7000-series-switches/white-paper-c11-737022.html (英語)

14. 2 分割帯域幅は、2 分割したクラスター システムのネットワークで使用できる帯域幅の合計です。システム ネットワークを中央で分割し、分割したすべてのリンクの帯域幅を加算します。

15. 複数の計算コードを各ネットワークで実行した場合の平均的なパフォーマンス結果に基づく比較。

ストレージ アーキテクチャ

GPU 対応データセンターをスケールアウトする際には、GPU アプリケーションとの相性が良い共有ストレージ テクノロジを選びましょう。GPU 対応サーバーのパフォーマンスは従来の CPU サーバーよりも格段に高いため、ストレージ システムが高度なワークロードのボトルネックにならないように特に注意する必要があります。

アクセス パターンやデータ タイプが異なるため、ワークロードの特性を考慮する必要があります。並列 HPC アプリケーションを実行するには、複数プロセスが同じファイルに同時にアクセスできるストレージ テクノロジが必要です。高速な分析処理には、多数のスレッドをサポートしたり、部分的なデータにすばやくアクセスしたりできる必要があります。視覚ベースのディープラーニングでは、分類、オブジェクト検出、セグメンテーションなどに使用される画像や動画に主に読み取りでアクセスするため、高ストリーミング帯域幅、高速なランダム アクセス、高速メモリ マップ (mmap) パフォーマンスが必要になります。テキストや音声を使用するリカレント ネットワークなどのディープラーニング技術では、高速帯域幅とランダムかつ小規模なファイルの組み合わせが必要となる場合があります。

ディープラーニングでは、トレーニング パフォーマンスを最大化するために、以前読み取ったデータをキャッシュする能力が最も重要です。ディープラーニング トレーニングでは、データを何度も反復処理することによって正確性を向上させます。1 つのトレーニング エクササイズで 100 回以上反復することも珍しくありません。データがローカルにキャッシュされていれば、毎回共有ストレージにアクセスする必要はありません。ローカル メモリやローカル ディスクにデータをキャッシュすることができるファイル システム テクノロジもあります。ローカル キャッシュの容量や性能が、ディープラーニング アプリケーションのニーズに一致しているのが理想的です。

以下の表 5 は、さまざまな GPU 対応ワークロードにおけるストレージ アーキテクチャの一般的なガイドラインを示しています。最適なストレージ システム設計において、各アプリケーションの要件を理解することが最も重要です。

表 5: ストレージ アーキテクチャ

ユース ケース	読み取りキャッシュの必要性	推奨されるネットワーク タイプ	ネットワーク ファイル システム オプション
データ分析	N/A	10 GBe	マルチスレッドの読み取りおよび小規模ファイルのパフォーマンスに優れたオブジェクト ストレージや NFS などのシステム
HPC	N/A	10/40/100 GBe、InfiniBand	クライアント数が多く、高速単一ノード性能、マルチスレッドの書き込みをサポートする NFS または HPC 対象のファイル システム
ディープラーニング、256x256 画像	はい	10 GBe	小規模ファイルを十分サポートする NFS またはストレージ
ディープラーニング、1080p 画像	はい	10/40 GBe、InfiniBand	ハイエンド NFS、HPC ファイル システム、高速ストリーミング性能を持つストレージ
ディープラーニング、4K 画像	はい	40 GBe、InfiniBand	HPC ファイル システム、ハイエンド NFS、ノードあたり 3 GB/秒以上の高速ストリーミング性能を持つストレージ
ディープラーニング、未圧縮画像	はい	InfiniBand、40/100 GBe	HPC ファイル システム、ハイエンド NFS、ノードあたり 3 GB/秒以上の高速ストリーミング性能を持つストレージ
ディープラーニング、キャッシュされないデータセット	いいえ	InfiniBand、10/40/100 GBe	同上。全体的ストレージ性能をスケーリングしてすべてのアプリケーション要件に同時に対応

表 5

これまでの説明はパフォーマンスに焦点を当てましたが、信頼性、回復性、管理性も同様に重要です。さまざまな選択肢の中からパフォーマンス ニーズを満たすソリューションを見つけるためには、ストレージ システムの運用に必要なすべての要素と組織のニーズを慎重に検討して、すべてにおいて最大限に価値を高められるものを選択してください。

システム ランタイム監視と管理

システムの監視および管理ツールが GPU に対応していることも重要です。温度、クロック レート、GPU メモリ使用状況などの主要な GPU パラメーターを監視する必要があります。既存の管理ツールに GPU 監視機能がない場合、または GPU 固有の監視機能を追加したい場合は、NVIDIA DCGM (Data Center GPU Manager)[16] を使用してください。

DCGM には、アクセラレーテッド データセンターの管理に使用できるエンタープライズ クラスのツールが一式揃っています。これを使用すると、システム ポリシーの適用、GPU ヘルスの監視、システム イベントの診断、データセンターのスループット最大化などを行うことができます。**Bright Cluster Manager**、**Altair PBSWorks**、**IBM Spectrum LSF**、**Adaptive Computing**、**SchedMD**、**Univa** など、DCGM を統合したツールもあります。

DCGM は、GPU の運用を監視して、全体のパフォーマンス、パフォーマンスの変動性、ノードの正常性などへの影響を最小化します。GPU の温度を監視することで、急激な温度変化による電力制限を防止します。DCGM をスケジューリング ソフトウェアに統合すると、ジョブ インスタンスごとの GPU 使用率やスループットを正確に測定できるようになります。また、DCGM で GPU の正常性や状態を定期的に診断することで、保守作業が必要なコンポーネントを事前に把握して、稼働時間を最大限まで高めることができます。

高密度 GPU ノード群を監視する重要なシステム メトリックスには、ファン速度、シャーシとコンポーネントの温度、システム エラー ログなどがあり、これらは PCIe バス、電源の状態、各電源の消費電力などに関連しています。IPMI (Intelligent Platform Management Interface) は、これらのサーバー コンポーネントの管理および監視機能を提供する標準的な方法として長く使用されています。サーバー ヘルスに関して豊富な情報を提供します。IPMI センサーは、サーバーの正常性に関するインサイトを提供し、障害の可能性を事前に知らせます。

まとめ

今日のエンタープライズおよびハイパースケールのデータセンターは、大量のデータで高付加な計算を行うディープ ニューラル ネットワーク (DNN) や人工知能 (AI) を使用したワークロードを念頭に構築されることが増えています。要求されるコンピューティング レベルがきわめて高く、GPU がもたらすメリットは多大です。大規模な並列処理能力を備え、高いメモリ帯域幅に最適化されている GPU を使用することで、AI クラスの行列積と分析によってすばやくインサイトを獲得できます。

16. http://www.nvidia.com/object/data-center-gpu-manager.html (英語)

GPU システムは、CPU のみのシステムよりも格段に優れたシステム パフォーマンスを発揮します。大規模データセンターに GPU システムを導入することで、コンピューティング ラックの削減、パフォーマンス強化、ワットあたりのパフォーマンスの向上、迅速な問題解決などが可能になります。GPU 対応データセンターでこれらを実現するには、以下のように、設計や運用に高度なアプローチを採用する必要があります。

- 高い電力密度をサポートするデータセンターを**設計**します。最大限に効率化するために、ラックあたり 30 KW から 50 KW の電力供給と温度管理されたエア フロー供給を検討します。さらに、冷却効率と運用コストの改善においては、ラック レベルまたはコンポーネント レベルで水冷システムの導入を検討しましょう。同時に、計算密度、電力密度、冷却密度を見直します。たとえば、コンポーネント レベルを冷却することによってデータセンター内の密度が増し、ワットおよび単価あたりのパフォーマンスが向上します。
- データおよび AI 中心のワークロードに使用する場合、大規模な計算と高い I/O スループットをサポートする**システム アーキテクチャ**を構築します。GPU によるパフォーマンス向上に合わせて、すべてのシステム サブコンポーネントを再評価し、ネットワークやストレージなどのボトルネックを解消する必要があります。
- パフォーマンスの高い AI ディープラーニング、高速分析、HPC ワークロードなどのボトルネックを解消するためには、高帯域幅、低遅延、高効率の**ネットワークおよびストレージ アーキテクチャ**を採用します。マルチシステム GPU ワークロードを適切にスケーリングするには、堅牢で競合性の少ないネットワークでデータを大量に転送できる必要があります。
- 重要なコンポーネントの**管理およびシステム監視**は、高密度の GPU システムおよび複数のノードでの効率的なスケーリングに欠かせません。マルチシステム ワークロードは、ジョブ内の最も遅いシステムに合わせて制限されるため、システム全体でパフォーマンス レベルが一致していることが重要です。これが一致していない場合、高速なシステムは遅いシステムの処理が完了するのを待たなければなりません。

このホワイト ペーパーで説明した GPU 対応データセンターの設計原則は、ボトルネックを解消し、パフォーマンスと効率を最大化し、真の NVIDIA GPU システムの能力を実現する鍵となります。

免責事項

以下をお読みください

説明、オプション、NVIDIA デザイン仕様書、リファレンス ボード、ファイル、図、診断、リスト、およびその他のドキュメント (以下、併せておよびそれぞれ「資料」という) を含む、本ホワイト ペーパーに示す情報はすべて、「現状有姿」とします。NVIDIA は資料について、明示または黙示、あるいは法定または非法定にかかわらず保証しません。さらに、特定の目的に対する黙示的保証、非抵触行為、商品性、および適正すべてに対する責任を明示的に否認します。

NVIDIA は、この仕様に対する訂正、修正、拡充、改善、その他の変更を随時行える権利と、任意の製品またはサービスを通知なしに終了する権利を留保します。お客様は、注文を行う前に最新の関連仕様を入手し、それらの情報が最新かつ完全であることを確認する必要があります。NVIDIA とお客様のそれぞれの承認を得た担当者によって署名された個別の販売契約に別段の定めがない限り、NVIDIA 製品は、注文確認時点で提供される NVIDIA の標準的な販売条件に従って販売されます。NVIDIA は、この仕様で参照される NVIDIA 製品の購入に関連した一切の顧客向け一般条件を適用することに明示的に反対します。NVIDIA 製品は、医療、軍事、航空、宇宙、生命維持の各装置で使用したり、NVIDIA 製品の故障または誤動作の結果、負傷、死亡、物的損害、環境劣化などが起こることを合理的に予想できるような用途で使用したりするよう設計または許可されておらず、また、そのような用途への適性も保証されていません。NVIDIA は、そのような装置や用途に NVIDIA 製品を含めたり使用したりすることに対して一切の法的責任を負いません。そのため、そのような使用はお客様自身の責任において行っていただきます。NVIDIA は、これらの仕様に基づく製品が追加的なテストや修正を行わずに特定の用途に適合することを表明するものでも、保証するものでもありません。各製品の全パラメーターのテストが NVIDIA によって実行されるとは限りません。お客様によって計画された用途への製品の適合性を確認し、用途または製品の不履行を避けるために必要なテストを実施することは、お客様側の責任です。お客様の製品設計に含まれる欠点は、NVIDIA 製品の品質および信頼性に影響する可能性があり、その結果、この仕様には含まれていない追加的あるいは異なる条件や要件が生じる可能性があります。NVIDIA は、次に基づく、またはそれに起因する一切の不履行、損害、コスト、あるいは問題に対しても責任を負いません。(i) この仕様に違反する方法で NVIDIA 製品を使用すること。(ii) お客様の製品設計。この仕様の下では、明示か黙示かを問わず、NVIDIA の特許権、著作権、その他の知的財産権が適用されるいかなるライセンスも供与されません。サードパーティ製品またはサービスに関して NVIDIA によって公開される情報は、それらの製品またはサービスを使用するための NVIDIA からのライセンスを構成するものでも、それらの製品またはサービスを保証もしくは是認するものでもありません。これらの情報を使用するには、サードパーティの特許またはその他の知的財産権の下でサードパーティから提供されるライセンスが必要になるか、NVIDIA の特許またはその他の知的財産権の