# NVIDIA DGX Station A100

## AI時代のワークグループアプライアンス

データサイエンスチームは、AIイノベーションの最先端に立ち、企業や私たちの世界を変革できるプロジェクトを構築しています。しかしながら、最も複雑なモデルをトレーニングするのに役立つ予備のコンピューティング環境を探し続けることも少なくありません。そのようなデータサイエンスチームは、どこにでも接続可能で、ハードウェアとソフトウェア全体に対して完全に最適化され、世界中にいる大勢の同時接続ユーザーに革新的な性能を提供できる専用のAIプラットフォームを必要としています

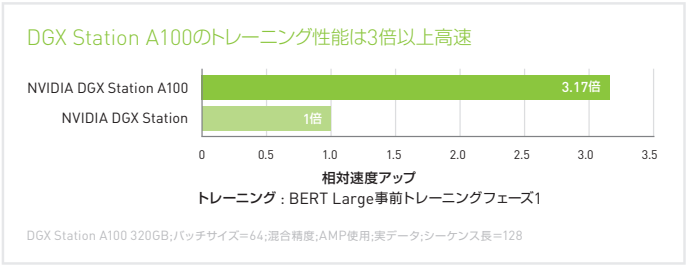
## NVIDIA DGX Station™A100の概要

- お客様のチームが無制限に使用できる2.5petaFLOPSの性能を提供可能なAIワークグループサーバー：トレーニング、推論、およびデータ分析用
- サーバーグレードのプラグアンドゴー方式。データセンターの電力と冷却が不要。
- 世界最高峰のAIプラットフォーム。複雑なインストールやITのサポートが不要。
- 完全に相互接続された4基のNVIDIA A100 TensorコアGPUと、最大320ギガバイト（GB）のGPUメモリを搭載した世界で唯一のワークステーション型システム。
- NVIDIAのノウハウと経験を活かしてAIトランスフォーメーションを速やかに実現

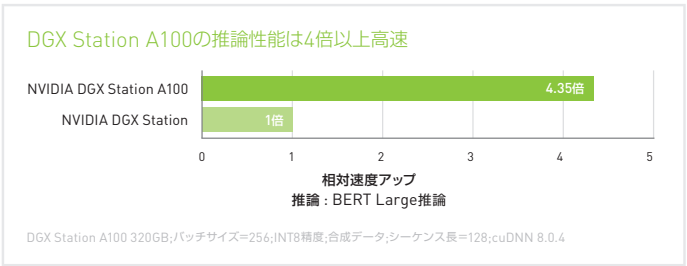
NVIDIA DGX Station A100は、AIスーパーコンピューティングをデータサイエンスチームにもたらし、データセンターやITインフラストラクチャの増設なしにデータセンターテクノロジを提供します。複数ユーザーが同時に接続できるように設計されており、DGX Station A100はオフィスに適したフォームファクタでサーバーグレードコンポーネントを活用します。完全に相互接続されたマルチインスタンスGPU（MIG）対応の4基のNVIDIA A100 TensorコアGPUと全体で最大320GBのGPUメモリを備えた唯一のシステムであり、標準的な電源コンセントにプラグ接続できるため、どこにでも配置できる強力なAIアプライアンスです。

## データサイエンスチーム向けのAIスーパーコンピューティング

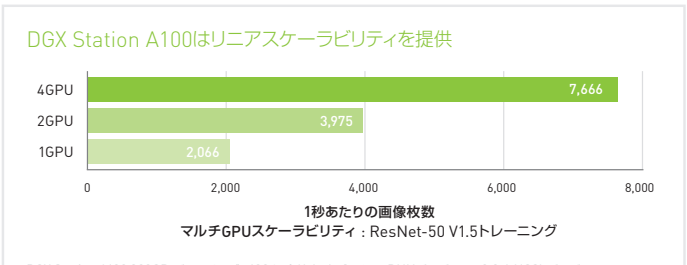
DGX Station A100は、複数の同時接続ユーザーに集中型AIリソースを手軽に提供できるAI時代のワークグループアプライアンスです。トレーニング、推論、分析のワークロードをMIGで並列実行可能。最大28台の独立したGPUデバイスが個々のユーザーとジョブに割り当てられるので、システム全体の性能に影響を与えません。DGX Station A100は、すべての




DGX Station A100 320GB;バッチサイズ=64;混合精度;AMP使用;実データ;シーケンス長=128




DGX Station A100 320GB;バッチサイズ=256;INT8精度;合成データ;シーケンス長=128;cuDNN 8.0.4




DGX Station A100 320GB;バッチサイズ=192;混合精度;実データ;cuDNNバージョン=8.0.4;NCCLバージョン=2.7.8; NGC MXNet20.10コンテナ

DGXシステムと同じく完全に最適化されたNVIDIA DGX™ソフトウェアスタックを備えており、個々のシステムからNVIDIA DGX POD™やNVIDIA DGX SuperPOD ™に至るまで、最高の性能とDGXベースインフラストラクチャとの完全相互運用性を提供し、規模の大小を問わず、あらゆる組織のチームにとって理想的なプラットフォームとなります。

## データセンターなしでデータセンターの性能を実現

NVIDIA DGX Station A100は、ワークステーションのフォームファクタでデータセンター級のAIサーバーを実現し、専用の電源と冷却なしに標準的なオフィス環境で使用するのに適しています。4つの超強力なNVIDIA A100 TensorコアGPU、最上位のサーバーグレードCPU、超高速NVMeストレージ、および最先端のPCIe Gen4バスを備える設計になっています。DGX Station A100にもNVIDIA DGX A100と同じベースボードマネジメントコントローラ（BMC）が搭載されており、システム管理者はリモート接続によって要求タスクを実行できます。DGX Station A100は、オフィス環境向けの最も強力なAIシステムであり、データセンターなしにデータセンターテクノロジを提供します。

## どこにでも配置できるAIアプライアンス

NVIDIA DGX Station A100は、企業のオフィス、実験室、研究施設、さらには自宅から作業する昨今のアジャイルデータサイエンスチーム向けに設計されています。大規模なAIインフラストラクチャを設置するには多大なIT投資と工業用の強力な電源供給と冷却機能を備えた大型データセンターが必要ですが、DGX Station A100なら、お客様のチームの作業スペースがどこにあっても標準的な壁のコンセントに接続するだけです。さらに、その革新的な冷却ベースのデザインにより、手で触れても熱くありません。1人で簡単にセットアップできるので、たった2本のケーブルで動作する世界最高峰のAIプラットフォームを数分で立ち上げて稼働させることができます。

## モデルの大規模化、応答の高速化

NVIDIA DGX Station A100はワークステーションではありません。これは、デスクの下に設置できるAIワークグループサーバーです。64コアのデータセンターグレードCPUに加えて、NVIDIA DGX A100サーバーと同じNVIDIA A100 TensorコアGPUを備え、それぞれ40GBまたは80GBのGPUメモリが高速なSXM4を介して接続されています。NVIDIA DGX Station A100は、NVIDIA®NVLink®を活用して4つのGPUを完全相互接続し、かつMIG対応でシステム性能に影響を与えずに並列ジョブと複数のユーザー向けに最大28台の独立したGPUデバイスを提供する唯一のオフィスに適したシステムです。

## 比類なきAIノウハウへの統合アクセス

NVIDIA DGX Station A100は、NVIDIAの数千人のAI専門家に支えられた完全なハードウェア&ソフトウェアプラットフォームであり、世界最大のDGX実験場であるNVIDIA DGX SATURNVから得られた知識に基づいて構築されています。DGX Station A100を所有すれば、NVIDIA DGXpertsに直接アクセスできます。NVIDIA DGXpertsは、AIに精通した専門家の国際チームであり、NVIDIAの10年を超えるAIリーダーシップのノウハウと経験に基づき、AI変革の急速な発展に貢献する規範的なガイダンスと設計の専門知識を提供します。これにより、ミッションクリティカルなアプリケーションが確実に素早く立ち上がってスムーズに稼働し続けることになり、インサイトを得るまでの時間が大幅に改善されます。

システム仕様

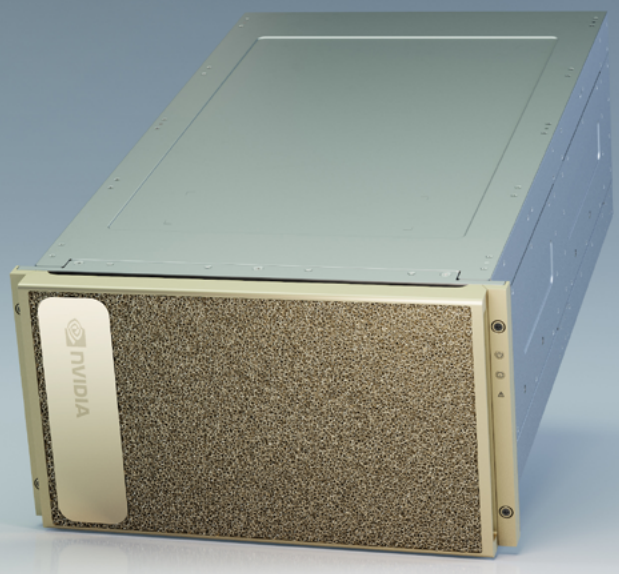
| | NVIDIA DGX Station A100 320GB | NVIDIA DGX Station A100 160GB |
|---|---|---|
| GPU | 4x NVIDIA A100 80GB GPU | 4x NVIDIA A100 40GB GPUs |
| GPUメモリ | 総計320GB | 総計160GB |
| 性能 | 2.5petaFLOPS AI<br>5petaOPS INT8 | |
| システム消費電力 | 1.5kW（100〜120Vac時） | |
| CPU | Single AMD 7742、64コア、2.25GHz（ベース）〜3.4GHz（最大ブースト） | |
| システムメモリ | 512GB DDR4 | |
| ネットワーク | デュアルポート10Gbase-T イーサネットLAN<br>シングルポート1Gbase-T イーサネットBMCマネジメントポート | |
| ストレージ | OS:1x 1.92TB NVMEドライブ<br>内部ストレージ:7.68TB U.2 NVMEドライブ | |
| DGXディスプレイアダプタ | 4GB GPUメモリ、4x Mini DisplayPort | |
| システム音響特性 | 37dB未満 | |
| ソフトウェア | Ubuntu Linux OS | |
| システム重量 | 43.1kg（91.0lbs） | |
| システム梱包重量 | 57.93kg（127.7lbs） | |
| システムサイズ | 高さ:639mm（25.1in）<br>幅:256mm（10.1in）<br>長さ:518mm（20.4in） | |
| 運用温度範囲 | 5〜35 °C（41〜95 °F） | |

お問い合わせ先

日本GPUコンピューティングパートナーシップ
https://www.gdep.jp

# NVIDIA DGX A100

## AI インフラストラクチャ向けの ユニバーサル システム



## エンタープライズAIのスケーリングへの挑戦

あらゆるビジネスで、人工知能(AI)を活用した変革が求められています。それは、困難な時代に生き残るためだけではなく、飛躍を遂げるためでもあります。ただし、そのためには、従来のアプローチを改善するAIインフラストラクチャ用のプラットフォームが必要です。これまでは、分析、トレーニング、推論のワークロードごとにサイロ化された低速のコンピューティングアーキテクチャが採用されていましたが、このアプローチでは、複雑さとコストが増大し、スケーリングの速度が制限され、現代のAIには対応できていませんでした。企業、開発者、データサイエンティスト、研究者に本当に必要なのは、すべてのAIワークロードを統合し、インフラストラクチャを簡素化し、ROIを向上させる新たなプラットフォームです。

## あらゆるAIワークロードに対応する ユニバーサルシステム

NVIDIA DGX™ A100は、分析からトレーニング、推論に至るまで、あらゆるAIワークロードに対応するユニバーサルシステムです。6Uのフォームファクターで5petaFLOPSのAIパフォーマンスを発揮し、従来のコンピューティングインフラストラクチャに代わる1つの統合システムとして、計算処理密度の新たな水準を確立します。また、NVIDIA A100 TensorコアGPUに搭載されたマルチインスタンス-GPU(MIG)機能を利用することにより、コンピューティングパワーをきめ細かく配分するかつてない能力を実現し、管理者は特定のワークロードに適したサイズのリソースを割り当てられるようになります。総計640ギガバイト(GB)までのGPUメモリが利用できるので、大規模なトレーニングジョブのパフォーマンスが最大3倍に向上し、MIGインスタンスのサイズが2倍になります。DGX A100は、単純で小さなジョブだけでなく、大規模かつ非常に複雑なジョブにも対応します。NGCの最適化されたソフトウェアでDGXソフトウェアスタックが実行され、高密度な計算能力と完全なワークロードの柔軟性を組み合わせることにより、シングルノードでの展開にも、NVIDIA DeepOpsで展開された大規模なSlurmクラスターやKubernetesクラスターにも最適な選択肢となっています。

## NVIDIA DGXpertsへのダイレクトアクセス

NVIDIA DGX A100は、単なるサーバーではありません。DGXの世界最大の実験場であるNVIDIA DGX SATURNVで得られた知識に基づいて構築された、ハードウェアとソフトウェアの完成されたプラットフォームです。そして、NVIDIA

### システムの仕様

| | NVIDIA DGX A100 640GB | NVIDIA DGX A100 320GB |
|---|---|---|
| GPU | NVIDIA A100 80 GB GPU x 8 | NVIDIA A100 40 GB GPU x 8 |
| GPUメモリ | 総計 640 GB | 総計 320 GB |
| パフォーマンス | AI で 5 petaFLOPS<br>INT8 で 10 petaOPS | |
| NVIDIA NVSwitch | 6 | |
| 消費電力 | 6.5 kW(最大) | |
| CPU | Dual AMD Rome7742、総計 128 コア、2.25 GHz (ベース)、3.4 GHz (最大ブースト) | |
| システムメモリ | 2 TB | 1 TB |
| ネットワーク | シングルポート Mellanox ConnectX-6 VPI<br>200 Gb/秒 HDR InfiniBand x 8<br>デュアルポート Mellanox ConnectX-6 VPI<br>10/25/50/100/200 Gb/秒 Ethernet x 2 | シングルポート Mellanox ConnectX-6 VPI x 8<br>200 Gb/秒 HDR InfiniBand<br>デュアルポート Mellanox ConnectX-6 VPI x 1<br>10/25/50/100/200 Gb/秒 Ethernet |
| ストレージ | OS:1.92 TB M.2 NVME ドライブ x 2<br>内部ストレージ:30 TB(3.84 TB x 8) U.2 NVMe ドライブ | OS:1.92 TB M.2 NVME ドライブ x 2<br>内部ストレージ:15 TB(3.84 TB x 4) U.2 NVME ドライブ |
| ソフトウェア | Ubuntu Linux OS<br>その他:<br>Red Hat Enterprise Linux<br>CentOS | |
| 重量 | 123.16 kg(最大) | |
| 梱包重量 | 163.16 kg(最大) | |
| サイズ | 全高:264.0 mm<br>全幅:482.3 mm<br>奥行:897.1 mm | |
| 運用温度範囲 | 5ºC〜30ºC | |

の何千人ものDGXpertsによるサポートを提供します。DGXpertはAIに精通した専門家で、役立つアドバイスや設計に関する専門知識を提供し、AI変革の加速に向けて支援します。過去10年にわたって蓄積してきた豊富なノウハウと経験を活かし、お客様がDGXへの投資から最大限の価値を引き出せるようお手伝いします。DGXpertのサポートによって、重要なアプリケーションを迅速に実行し、スムーズな運用を維持し、インサイトを得るまでの時間を飛躍的に短縮することができます。

## 最速での解決

8基のNVIDIA A100 TensorコアGPUを搭載するNVIDIA DGX A100は、比類のないアクセラレーションを提供し、NVIDIA CUDA-X™ソフトウェアとエンドツーエンドのNVIDIAデータセンターソリューションスタックに完全に最適化されています。NVIDIA A100 GPUは、FP32と同じように動作しながらも1秒あたりの浮動小数点演算回数（FLOPS）が前世代の20倍のAIを実現するTensor Float 32（TF32）という新しい精度に対応しています。最大の特長は、コードを変更することなくこの高速化を実現できる点です。またFP16を活用したNVIDIAの自動混合精度機能を使用すれば、A100ではコードを1行追加するだけで、さらに2倍の性能が得られます。

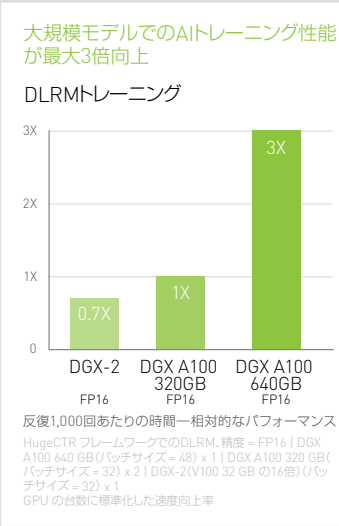
A100 80GB GPUは、高帯域幅メモリが40GB（HBM）から80GB（HBM2e）に倍増し、GPUメモリ帯域幅がA100 40GB GPUを30%上回る、世界初の毎秒2テラバイト超を実現しています。DGX A100は第3世代のNVIDIA® NVLink®を初めて搭載し、GPU間の直接帯域幅を毎秒600ギガバイト（GB/秒）に倍増させています。これは、PCIe Gen 4のほぼ10倍に相当します。他にも、前世代の2倍の速度を持つ新しいNVIDIA NVSwitch™も搭載しています。このかつてないパワーによって、最短でソリューションを実現でき、これまで不可能だったり、現実的ではなかったりした課題に取り組めるようになります。

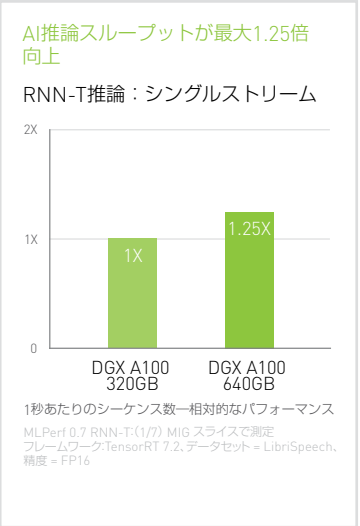
## 世界で最も安全なエンタープライズ向けAIシステム

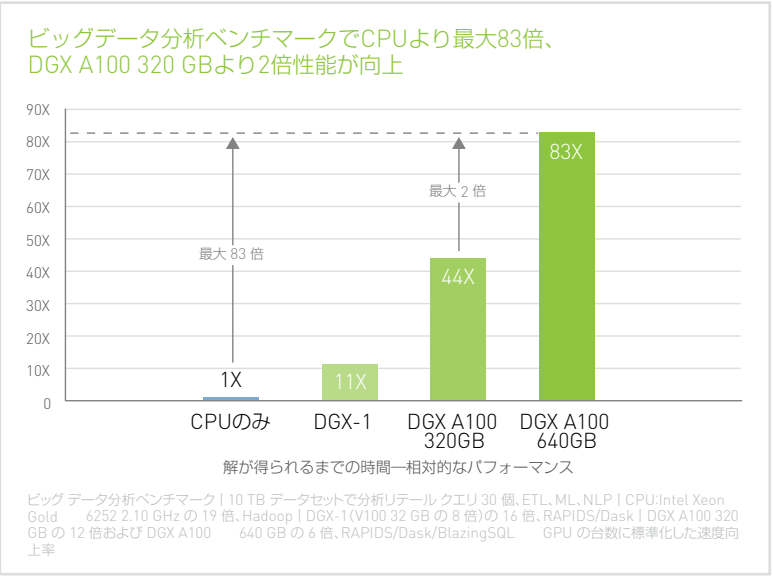
NVIDIA DGX A100は、あらゆる主要なハードウェアおよびソフトウェアコンポーネントを保護する多層的なアプローチによって、AIを活用する企業において最も堅牢なセキュリティ体制を実現します。ベースボード管理コントローラー（BMC）、CPUボード、GPUボード、自動暗号化ドライブ、セキュアブートなど、幅広いセキュリティ機能が組み込まれているため、IT部門は脅威の評価や軽減に時間を費やすことなく、AIの運用に集中できます。

## NVIDIA Mellanoxによるデータセンターの比類なきスケーラビリティ

DGXシステムの中で最速のI/OアーキテクチャをNVIDIA DGX A100は、NVIDIA DGX Super POD™のような大規模なAIクラスターのための基本構成要素となり、企業は拡張性の高いAIインフラストラクチャの計画を策定できます。DGX A100は、クラスタリング用に8つのシングルポートNVIDIA Mellanox® ConnectX®-6 VPI HDR InfiniBandアダプターと、ストレージとネットワーキング用に最大2つのデュアルポートConnectX-6 VPI Ethernetアダプターを備えており、いずれも毎秒200 Gbの性能を発揮します。大規模なGPUアクセラレーテッドコンピューティングと、最先端のネットワーキングハードウェアおよびソフトウェアの最適化を組み合わせることで、数百、数千ノードにまでスケールアップが可能になり、対話型AIや大規模な画像分類などの難易度の高い課題に対応できます。




## 信頼できるデータセンターのリーダー企業と共に構築された実証済みのインフラストラクチャソリューション

ストレージとネットワーキングの技術を誇るリーディングプロバイダーとの連携により、インフラストラクチャソリューションのポートフォリオに、NVIDIA DGX POD™の最高クラスのリファレンスアーキテクチャが加わりました。これらのソリューションは、NVIDIAパートナーネットワーク（NPN）を通じて、すぐに導入可能な完全統合型サービスとして提供されるため、データセンターへの AI 導入を簡素化かつ迅速化できます。

お問い合わせ先

日本GPUコンピューティングパートナーシップ
https://www.gdep.jp