EXA5ファイルシステムを搭載したDDN A^{3}I AI400ストレージアプライアンス

NVIDIA DGX SuperPODリファレンスアーキテクチャ

Document History

RA-09734-001

Version	Date	Authors	Description of Change
001	2019-11-13	Robert Sohigian and Craig Tierney	Initial release

概要

NVIDIA DGX SuperPOD™は、世界で最も困難なAIの問題を解決することを目的として設計された、今までにない人工知能 (AI) スーパーコンピューティング インフラストラクチャです。革新的な性能を提供し、完全に統合されたシステムとして数週間で展開できます。

DGX SuperPODが提供する画期的な性能により、大規模なディープラーニングモデルの迅速なトレーニングが可能になります。画像分類、物体検出、自然言語の非常に正確なモデルを作成するには、大量のトレーニングデータが必要になります。これらのデータには、SuperPOD全体のどこからでも素早くアクセスできる必要があります。DGX SuperPODの計算能力を最大限引き出すためには、タスクに合わせてDGX SuperPODとストレージシステムの組み合わせを決めることが不可欠です。

このホワイトペーパーでは、DGX SuperPODへの接続時にディープラーニング (DL) ワークロードをサポートするためのDDN® A³I AI400アプライアンスの適合性を評価しました。AI400アプライアンスはコンパクトで省電力のストレージソリューションです。ストレージにNVMeドライブを使用し、ネットワークトランスポートとしてInfiniBandを使用することで、驚くべき基本性能を発揮します。AI400アプライアンスでは、データ管理機能が追加、強化されたエンタープライズ版のLustre並列ファイルシステムを提供するEXAScalerファイルシステムが活用されています。

Lustreなどの並列ファイルシステムにより、データアクセスが簡素化されます。また、効率的なトレーニングのために大量のデータを高速で処理する必要があり、ローカルキャッシュが適切でないようなユースケースにも対応します。シングルスレッドとマルチスレッドでの高い読み取り性能は、以下の用途で活用されます。

- データセットをDGX-2システムメモリやDGX-2 NVMe RAIDにローカルにキャッシュできない場合のトレーニング
- ローカルディスクへのデータの高速ステージング
- 大規模な個別データオブジェクト (非圧縮またロスレス圧縮の1080p画像など) を使用したトレーニング
- TFRecordなどの最適化されたデータ形式を使用したトレーニング
- データセットの長期保存 (LTS) 用のワークスペース

目次

ストレージの概要

トレーニングの性能は、ストレージからのデータの読み取りおよび再読み取りの速度によって制限されることがあります。性能を決める鍵となるのは、データを複数回読み取る能力です。データがキャッシュされる場所がGPUに近いほど、読み取りを速くすることができます。ストレージの設計では、永続的な保存の場合も一時的な保存の場合も、性能、容量、コストの各ニーズのバランスが保たれるように、さまざまなストレージテクノロジーの階層を考慮する必要があります。

表1に、ストレージのキャッシュ階層を示します。データサイズと性能のニーズに応じて、階層の各層を活用してアプリケーションの性能を最大限引き出すことができます。

表1.DGX SuperPODのストレージおよびキャッシュ階層

ストレージ階層のレベル	テクノロジー	総容量	読み取り性能
RAM	DDR4	1.5 TB/ノード	100 GB/s以上
内蔵ストレージ	NVMe RAID	30 TB/ノード	25 GB/s以上
高速ストレージ	汎用	具体的なニーズにより異なる	必要な性能: • システム読み取り総量: 32 GB/s以上 • システム書き込み総量: 16 GB/s以上 • シングルノードの読み取り: 5 GB/s以上が望ましい • シングルノードでGPUあたり1 GB/sの読み取り (16 GB/s)

データをローカルRAMにキャッシュすると最高の読み取り性能が得られます。このキャッシュは、データがファイルシステムから読み取られるとシステムからは見えなくなります。ただし、RAMのサイズはDGX-2システムで1.5 TBに制限されており、この容量をオペレーティングシステム、アプリケーション、およびその他のシステムプロセスと共有する必要があります。DGX-2システムのローカルストレージは、30 TBの非常に高速なNVMeを提供します (必要に応じて60 TBまでアップグレードできます)。ローカルストレージは高速ですが、ローカルディスクのみで動的に変更可能な環境を管理することは実用的ではありません。

高速ストレージでは、組織のデータをすべてのノードで共有して表示できます。小さいランダムなI/Oパターンに合わせて最適化する必要があるものの、高いピークノード性能と、高いファイルシステム集約性能をもたらし、組織で発生するさまざまなワークロードに対応できます。

今日の30 TBというデータセットでも大きいと考えられていますが、自動車やその他のコンピュータービジョンタスクでは、トレーニングに1080pの画像を使用し、場合によっては非圧縮というユースケースも見られます。これらの形式のデータセットは、サイズが30 TBを容易に超える可能性があります。そうしたケースでは、読み取り性能としてGPUあたり1 GB/sが必要と見られています。

前述の指標は、高速ストレージシステムから直接、ローカルでトレーニングする場合における、さまざまなワークロード、データセット、およびニーズを前提としたものです。性能や容量の最終的な要件を決める前に、ワークロードの特性を理解することが最善となります。

ストレージ階層はさらに拡張することができます。LTSは多くの場合、データや履歴結果を保存するために重要です。LTSは数十または数百ペタバイト(PB)にもなる可能性があります。この階層からのデータアクセスの頻度は低く、格納や呼び出しの性能はあまり重要ではありません。そのためのソリューションは、前述の高速ストレージとは異なる方法でコストを最適化できます。例えば、低速の回転ディスクをベースとしたり、S3互換のオブジェクトストレージテクノロジーを使用することができ、場合によってはクラウドも使用できます。

DLトレーニングのファイルシステムにおいては、読み取り、特に再読み取りの性能が重要な指標となります。DLトレーニングの実行中は、モデルを反復処理して、最も正確なモデルが見つかるまで、データが何度も繰り返し読み取られます。トレーニングのデータセットが十分に小さければ、ローカルメモリまたはローカルNVMeディスクにキャッシュして、リモートファイルシステムへのアクセスを制限できます。しかし、データセットが大きくなると、作業用のデータセット全体を格納できるほどの十分大きな容量をローカルシステムでプロビジョニングすることが、常に実用的になるとは限りません。そのような場合は、エポックごとにネットワークファイルシステムからデータを再読み取りするようなモデルを、トレーニングすることが必要になります。DGX SuperPODなどの大規模な構成でこうしたレベルのI/Oを処理するためには、多数のI/Oパターン (大きなブロック(1 MB超)、小さなブロック(1 MB未満、場合によっては32 kB未満)、メモリマップファイルなど) からなるデータの大量のスループットが必要になります。DGX SuperPODのニーズを満たすストレージソリューションの場合は、このようなタイプのI/Oパターンを処理でき、すべてのノードに対して同時に、毎秒数十ギガバイトという読み取り性能にまでスケーリングできる能力が必要です。

DDN AI400アプライアンスについて

DDN AI400アプライアンス (図1) は、前述の要件を満たし、DGX SuperPODやデータセンターのシステム管理の性能を最大限引き出すうえで重要な機能をいくつも備えています。

- DDN AI400アプライアンスは、性能、負荷分散、およびレジリエンシーを実現するために、複数のEDR InfiniBandまたは100 GbEネットワーク接続を使用して、DGX SuperPODクライアントと通信します。DDNの並列プロトコルにより、GPUあたり1 GiB/sの目標を超える20 GiB/s以上の速度でストレージにアクセスできます。画像サイズが1080p、4K、またはそれ以上に大きくなる画像ベースのネットワークをトレーニングするには、こうした性能が必要です。
- DDN AI400アプライアンスはスケーラブルに構成可能なビルディングブロックであり、単一のファイルシステムに簡単に集約できるため、容量、性能、機能をシームレスにスケーリングできます。
- DDN AI400アプライアンスは30 TiBから240 TiBに至るさまざまな容量で構成できます。
- DDN AI400アプライアンスは、すべてNVMeを使用したアーキテクチャに基づいています。ランダム読み取り性能も優れており、多くの場合、シーケンシャル読み取りパターンと同じくらい高速です。

図1.DDN AI400アプライアンス

InfiniBandのネイティブサポートは、性能を最大限引き出し、CPUのオーバーヘッドを最小限に抑える重要な機能です。DGX SuperPODのコンピューティングファブリックはInfiniBandです。つまり、ストレージファブリックをInfiniBandとして設計することで、管理を必要とする高速ファブリックのタイプが1つだけになり、運用が簡素化されます。

注:このペーパーで公開したテストが完了したことから、DDNは次世代のDDN AI400Xアプライアンスをリリースしました。AI400Xアプライアンスは、さらに優れたIOPS性能とスループットを提供するように刷新されています。また、MellanoxのHDR100 InfiniBand接続が可能になり、次世代のHDRファブリックをサポートします。ここで紹介しているベンチマークは現行世代の製品のものですが、AI400XアプライアンスはDLストレージのニーズに応える、さらに優れた性能を提供できます。

テストの方法と結果

すべての、シングルノード、および、マルチノード帯域幅の性能測定には、ファイルシステムテストツールFIOを使用しました。マルチノードメタデータの性能測定には、MDTestを使用しました。また、実際のアプリケーションの性能を評価するために、MLPerfのいくつかのベンチマークを使用しました。

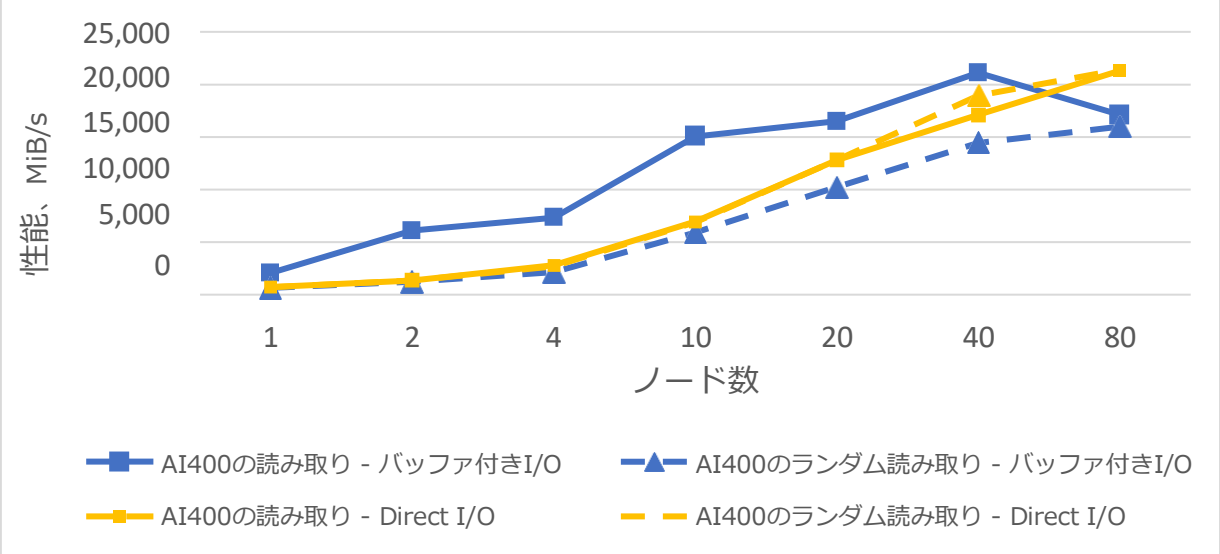
DDN AI400アプライアンスは、Lustre 2.12のエンタープライズリリースに基づくEXA5ファイルシステムを使用して、構成しました。ストレージは、DGX SuperPOD上の独立したInfiniBandファブリックに接続しました。DGX SuperPODの各システムを、2つのMellanox InfiniBandアダプターを使用してストレージファブリックに接続することで、システムあたり20 GiB/sを超える帯域幅が可能となりました。

以下のセクションでは、DLワークロードに対するDDN AI400アプライアンスの全体的な能力を測定するために使用したいくつかのテストについて説明します。スレッドごとの読み取り性能、ノードごとの読み取り性能、およびDLトレーニング性能の、三つのテストを行いました。読み取り性能のベンチマークは、DLアプリケーションに必要な要件と、最も強く関連付けられます。DLトレーニングのベンチマークは、NVMeベースのDDN AI400アプライアンスがもたらす性能上のさまざまな利点が、実アプリケーション性能とどのように結びつくかを示す例となります。

スレッドごとの読み取り性能

DLトレーニングでは、シーケンシャルとランダムの両方の読み取り性能が重要になります。図2に、スレッド数が増加したときのDDN AI400アプライアンスのシーケンシャル読み取りとランダム読み取りの性能を示します。バッファ付きI/Oの場合、シングルスレッドのシーケンシャル読み取り性能は1.5 GiB/sを超え、ランダム読み取りでは640 MiB/sを超えます。80スレッドでのDirect I/Oでは、読み取り性能は20 GiB/s超までスケールします。 これは、目標性能である、ノードあたり16 GiB/s (またはGPUあたり1 GiB/s)を上回っています。シーケンシャルアクセス時のバッファ付きI/Oの性能は、スレッド数が最も多い場合以外はDirect I/Oよりも優れています。

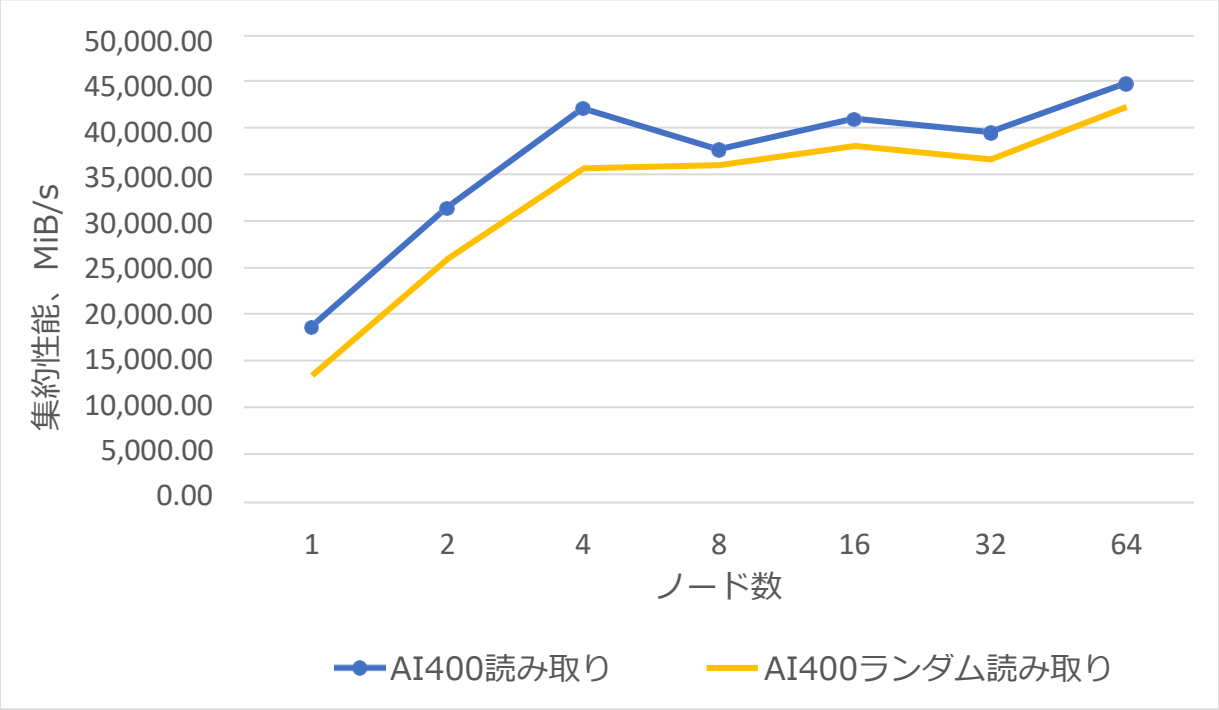
図2.DDN AI400アプライアンスのシーケンシャル読み取りとランダム読み取りの性能



ノードごとの読み取り性能

図3に、システムレベルの検証による、単一のDDN AI400アプライアンスの、ピーク集約性能を示します。図に示されるように、大きなブロックに対するシーケンシャル読み取りとランダム読み取りの性能は同程度となります。シングルノードのパフォーマンスは16 GiB/sを超えています。最大の読み取り性能は44 GiB/sを超え、シーケンシャル読み取りとランダム読み取りの両社とも、同等の性能になります。使用するDDN AI400アプライアンスを増やすことで、集約性能は、ほぼ直線的にスケーリングします。

図3.DGX SuperPOD全体での読み取りおよびランダム読み取りの性能



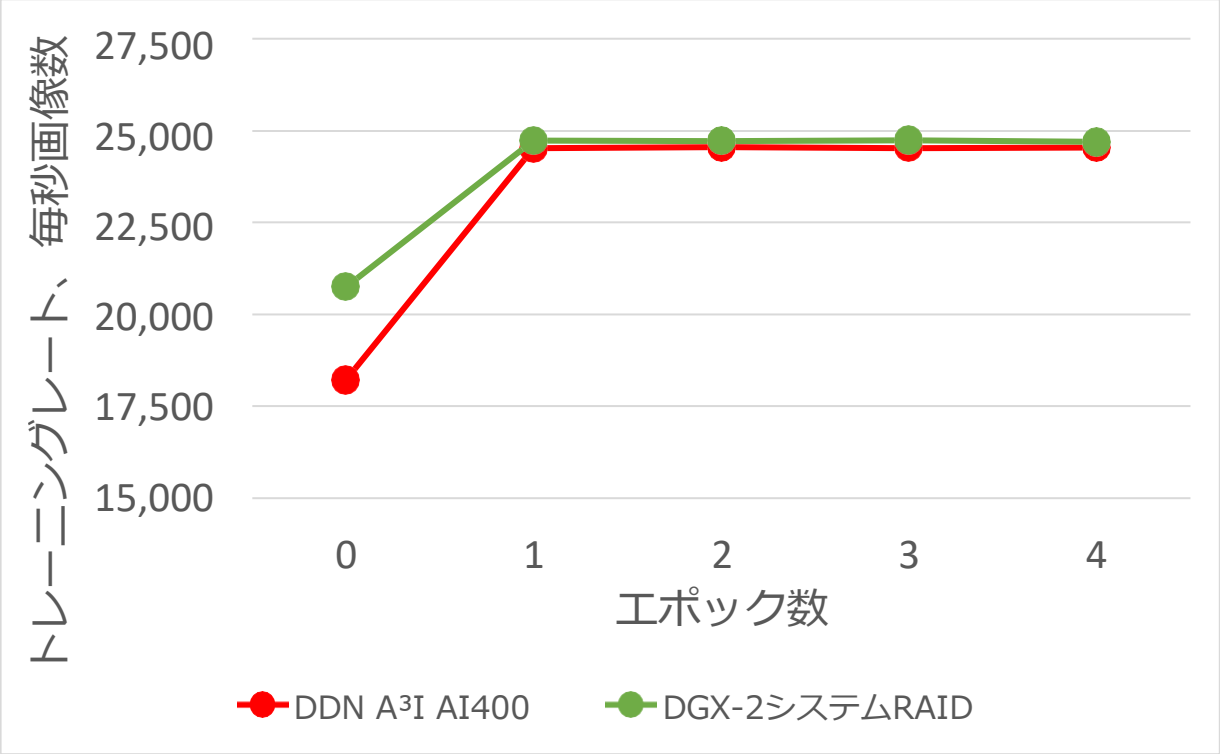
MLPerfおよびResNet-50の性能

FIOやMDTESTなどのマイクロベンチマークはファイルシステムの性能を特徴付けるうえで重要なツールですが、最善の指標となるのは実アプリケーション性能です。MLPerfのベンチマークは、今日のDLワークロードを適切に表現するさまざまなI/O要件に対応した、さまざまなDLモデルを提供します。

これらのワークロードの中で最もI/O処理に負荷がかかるのはResNet-50です。DGX-2システムでは、トレーニングレートは毎秒24,700画像を超えます。ImageNetデータベースにある画像の平均サイズは122 KiBです。堅牢なモデルを実現するためには、エポックごとにデータをランダムに処理することが重要です。例えば、16個のGPUにわたって比較的小さなファイルをランダムに読み取る場合、I/Oの要件は3 GiB/sを超えます。また、データはメモリマップファイルとして読み取られるため、さらに状況は複雑になります。メモリマップファイルとしてのファイル読み取りは、ローカルファイルシステムの場合にはよい最適化になりますが、ネットワークファイルシステムでは、より扱いが難しくなります。その理由として、ページ境界 (この場合は4 KiB) をまたぐ際の読み取り性能、および、複数ノード上のプロセスがページを読み取る際に、他のプロセスがそのページに書き込もうとしないことを保証するために、オーバーヘッドが必要になることが挙げられます。

図4に、MLPerf v0.6収録されているResNet-50を用いて、トレーニング性能を秒あたりの画像数で計測した結果を示します。このベンチマークではデータ形式をRecordIOファイルとしています。RecordIOファイルはMXNetフレームワークによって使用されます。すべてのファイルは、単一の大きなファイルにまとめられています。RecordIOファイルはMMAPを使用して読み取られます。NVIDIAのMLPerfリファレンスコードでは、NVIDIA Data Loading Library (DALI) フレームワークをデータの読み取りに使用しています。

図4.MLPerf v0.6およびResNet-50のトレーニング性能



ここでは、データをDGX-2システムのローカルRAIDディスクから読み取った結果と、DDN AI400アプライアンスから読み取った結果が、示されています。まず注目すべき点として、DGX-2システムRAIDを用い、まだ、データがキャッシュされていない状態でのトレーニング性能です。ローカルRAIDから読み取る場合であっても、データがキャッシュされていないエポック (エポック0) では、データがキャッシュされている後続のエポックの場合よりも遅くなります。DGX-2システムのRAIDディスクは25 GiB/s以上の読み取りを維持できます。優れた性能の達成に必要なのは読み取り性能だけではありません。

DDN AI400アプライアンスは、最初のエポックで、DGX-2システムRAIDと比して87%の性能を達成できます。ファイルが小さく、MMAPを使用しているため、この性能は優秀であると考えられます。また、データをキャッシュした場合のトレーニング性能の違いは1%以内に収まっています。DDN A³Iソリューションは、非常に困難なMLPerfベンチマークにおいても、I/O要件に関して、優れた性能を発揮しています。

まとめ

EXA5を搭載したDDN A³I AI400アプライアンスは、DGX SuperPOD向けのスケーラブルな共有型高性能ストレージの構築における、優れたビルディングブロックとなります。DDN AI400アプライアンスはスケーラブルに構成可能なビルディングブロックであり、単一のファイルシステムに簡単に集約でき、容量、性能、機能をシームレスにスケーリングできます。DDN AI400アプライアンスのアーキテクチャは、すべてNVMeにより構成されており、優れたランダム読み取り性能を備え、多くの場合、シーケンシャルパターンの読み取りと同じくらい高速です。これらの性能はすべて、2Uのビルディングブロックで提供され、DGX SuperPODのあらゆる性能要件を満たします。加えて、ノードあたり16 GiB/s (またはGPUあたり1 GiB/s) という、より大きな望ましい目標を達成します。

DDN AI400アプライアンスは、DGX SuperPOD向けのストレージとして、現在および将来のストレージニーズに対応できる優れた選択肢です。

Notice

The information provided in this specification is believed to be accurate and reliable as of the date provided. However, NVIDIA Corporation (“NVIDIA”) does not give any representations or warranties, expressed or implied, as to the accuracy or completeness of such information. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This publication supersedes and replaces all other specifications for the product that may have been previously supplied.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and other changes to this specification, at any time and/or to discontinue any product or service without notice. Customer should obtain the latest relevant specification before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer. NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this specification.

NVIDIA products are not designed, authorized or warranted to be suitable for use in medical, military, aircraft, space or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer’s own risk.

NVIDIA makes no representation or warranty that products based on these specifications will be suitable for any specified use without further testing or modification. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer’s sole responsibility to ensure the product is suitable and fit for the application planned by customer and to do the necessary testing for the application to avoid a default of the application or the product. Weaknesses in customer’s product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this specification. NVIDIA does not accept any liability related to any default, damage, costs or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this specification, or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this specification. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA. Reproduction of information in this specification is permissible only if reproduction is approved by NVIDIA in writing, is reproduced without alteration, and is accompanied by all associated conditions, limitations, and notices.

ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, “MATERIALS”) ARE BEING PROVIDED “AS IS.” NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE.

Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA’s aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the NVIDIA terms and conditions of sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, NVIDIA DGX SuperPOD, NVIDIA DGX POD, NVSwitch, NVLink, Tesla, and GPU Direct are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051

http://www.nvidia.com