NVIDIA-Certified Systems Configuration Guide for PCIe Servers

Design Guide

DG-10105-001 _v04 | March 2022

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

This PCIe server system configuration guide provides the server topology and system configuration recommendations for server designs that integrate NVIDIA® PCIe form factor graphics processing units (GPUs) from the NVIDIA® Turing™ and Ampere GPU architectures. NVIDIA® HGX™ system configurations are available to system designers through the NVIDIA Partner Network - https://www.nvidia.com/en-us/about-nvidia/partners/.

The optimal PCIe server configuration depends on the target workloads (or applications) for that server. In some cases, the server design can be optimized for a specific use case or target workload, but in general a GPU server can be configured to execute the following general applications or target workloads:

- Inference and Intelligent Video Analytics (IVA)
- Deep learning (DL) training / AI and high-performance computing (HPC)
- Aerial Edge AI and 5G vRAN
- Cloud gaming
- Rendering and virtual workstation

Note: This system configuration guide provides system design recommendations for inference and DL training. Other target workloads are discussed in separate documents.

1.1. References

The following NVIDIA documents are relevant to this system design guide:

- *NVIDIA Form Factor 5.0 Specification for PCIe Server Products* (SP-10066-001)
- *NVIDIA A100 PCIe Product Specification* (SP-10026-001)
- *NVIDIA A40 Product Specification* (SP-10159-001)
- *NVIDIA A30 Product Specification* (SP-10367-001)
- NVIDIA-Certified Systems documents - various

Chapter 2. Inference Systems Configurations

Inference application performance is greatly accelerated with the use of NVIDIA GPUs and include workloads such as:

- DeepStream – GPU accelerated Intelligent Video Analytics (IVA)
- NVIDIA® TensorRT™, Triton – inference software with GPU acceleration
- Natural language recognition (NLR)

A GPU server designed for executing inference workloads can be deployed at the edge or in the data center. Each server location has its own set of environmental and compliance requirements. For example, an edge server may require NEBS compliance, with more stringent thermal and mechanical requirements.

Table 1 provides the system configuration requirements for an inference server using NVIDIA GPUs.

Table 1. Inference Server System Configuration

Parameter	Inference Server Configuration
GPU	A100 A40 A30
GPU Configuration	1x / 2x / 4x / 8x GPUs per server
CPU	AMD EPYC (Rome or Milan) Intel Xeon (Skylake, Cascade Lake, Ice Lake)
CPU Sockets	1P / 2P
CPU Speed	2.1 GHz minimum base clock
CPU Cores	6x physical CPU cores per GPU
System Memory	Minimum 1.5x of total GPU memory / 2.0x is recommended Evenly spread across all CPU sockets and memory channels

Parameter	Inference Server Configuration
PCI Express	One PCIe Gen4 x16 or Gen3 x16 root port per GPU when PCIe switches are not used One PCIe Gen4 x16 or Gen3 x16 root port for every two GPUs when PCIe switches are used
PCIe Topology	Balanced PCIe topology with GPUs spread evenly across CPU sockets and PCIe root ports. NIC and NVMe drives should be under the same PCIe switch or PCIe root complex as the GPUs. Note that a PCIe switch may not be needed for low-cost inference servers. See topology diagrams for details.
PCIe Switches	Broadcom PEX88000 Gen4 and Microsemi PM40100 Gen4 PCIe Switches
Network Adapter (NIC)	Mellanox ConnectX-6, ConnectX-6 DX, or BlueField-2 network adapters are recommended. See Section Network for details.
NIC Speed	25 Gbps per GPU
Storage	One NVMe per server (either socket)
Remote Systems Management	Redfish 1.0 (or greater) compatible
Security	TPM 2.0 module (secure boot)

Chapter 3. Deep Learning Training System Configurations

DL training application performance is greatly accelerated by use of NVIDIA GPUs and include workloads such as:

- TensorFlow / Resnet
- PyTorch
- RAPIDS

A GPU server designed for executing training workloads is typically deployed in the data center. Each data center or cloud service provider (CSP) may have its own set of environmental and compliance requirements, but those requirements are typically not as stringent as NEBS or edge server requirements.

Table 2 provides the system configuration requirements for a DL training server using NVIDIA GPUs.

Table 2. Training Server - System Configuration

Parameter	Deep Learning Server Configuration
NVIDIA GPU	A100
GPU Configuration	2x / 4x / 8x GPUs per server GPUs are balanced across CPU sockets and root ports See topology diagrams for details
CPU	AMD EPYC (Rome or Milan) Intel Xeon (Skylake, Cascade Lake, Ice Lake)
CPU Sockets	1P / 2P
CPU Speed	2.1 GHz minimum base clock
CPU Cores	6x physical CPU cores per GPU (minimum)
System Memory	Minimum 1.5x of total GPU memory / 2.0x is recommended Evenly spread across all CPU sockets and memory channels

Parameter	Deep Learning Server Configuration
PCI Express	One PCIe Gen4 x16 or Gen3 x16 root port per GPU when PCIe switches are not used One PCIe Gen4 x16 or Gen3 x16 root port for every two GPUs when PCIe switches are used
PCIe Topology	Balanced PCIe topology with GPUs spread evenly across CPU sockets and PCIe root ports. NIC and NVMe drives should be under the same PCIe switch or PCIe root complex as the GPUs. See topology diagrams for details.
PCIe Switches	Broadcom PEX88000 Gen4 and Microsemi PM40100 Gen4 PCIe Switches
Network Adapter (NIC)	Mellanox ConnectX-6, ConnectX-6 DX, or BlueField-2 network adapters are recommended - see Section Network for details.
NIC Speed	100 Gbps per A100 GPU
Storage	One NVMe drive per CPU socket
Remote Systems Management	Redfish 1.0 (or greater) compatible
Security	TPM 2.0 module (secure boot)

Chapter 4. Topology Diagrams

This chapter shows the system configurations that correspond to those outlined in Table 1 and Table 2 for inference and DL training servers, starting from the simplest configuration to the most complex.

Note that some server configurations may not require a PCIe switch. For example, a server with one or two GPUs per socket may not require a PCIe switch, depending on the number of PCIe lanes available from the CPU.

Figure 1. 1P Server with One GPU

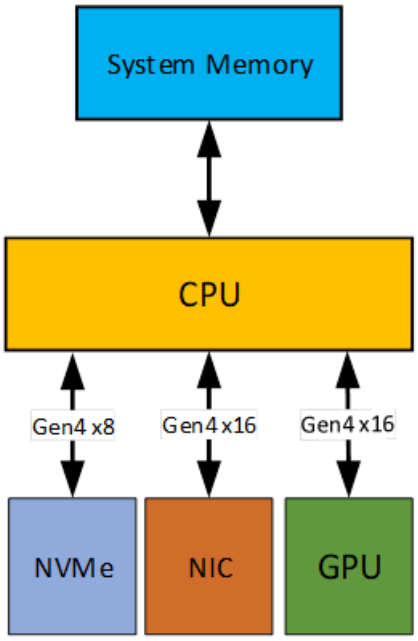


Figure 2. 1P Server with Two GPUs

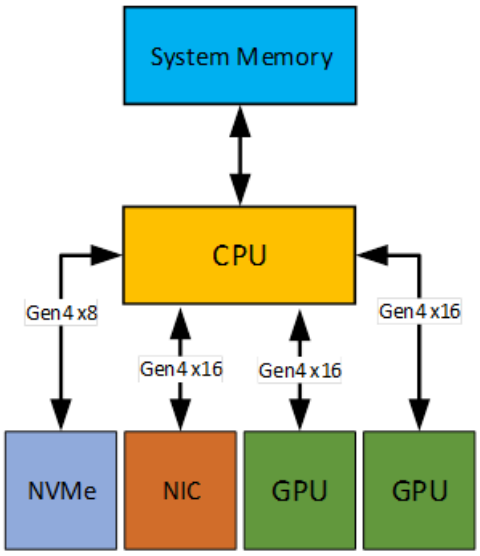


Figure 3. 2P Server with One GPU

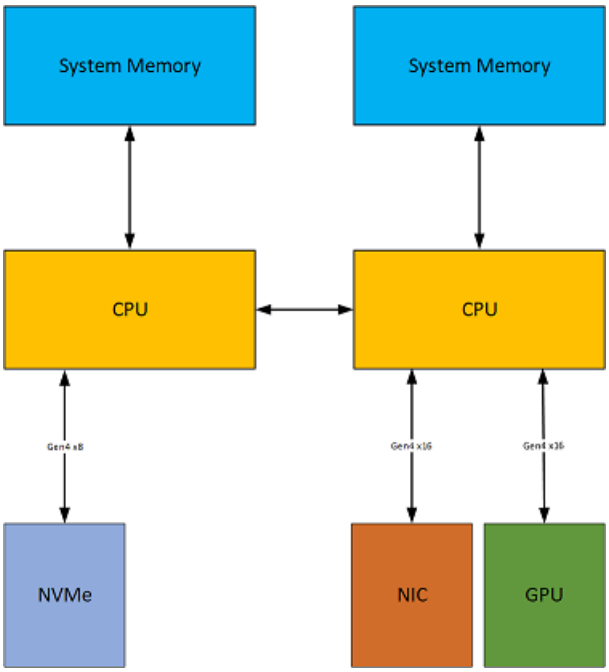


Figure 4. 2P Server with Two GPUs

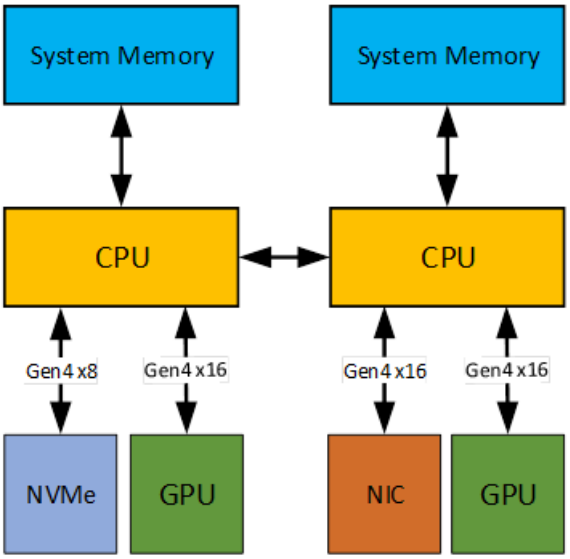


Figure 5. 1P Server with Four GPUs

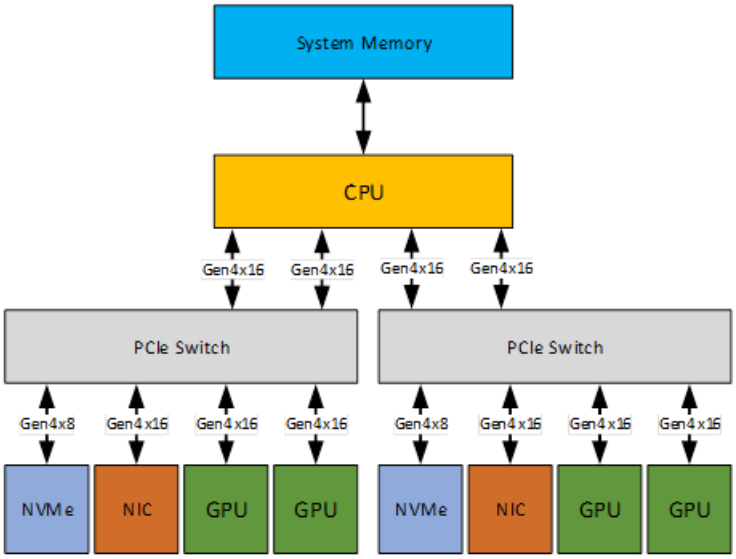


Figure 6. 2P Server with Four GPUs

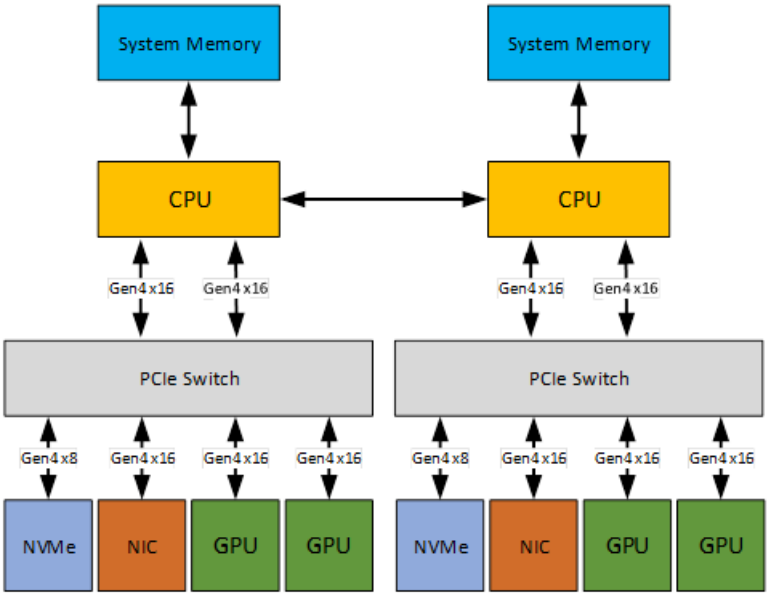
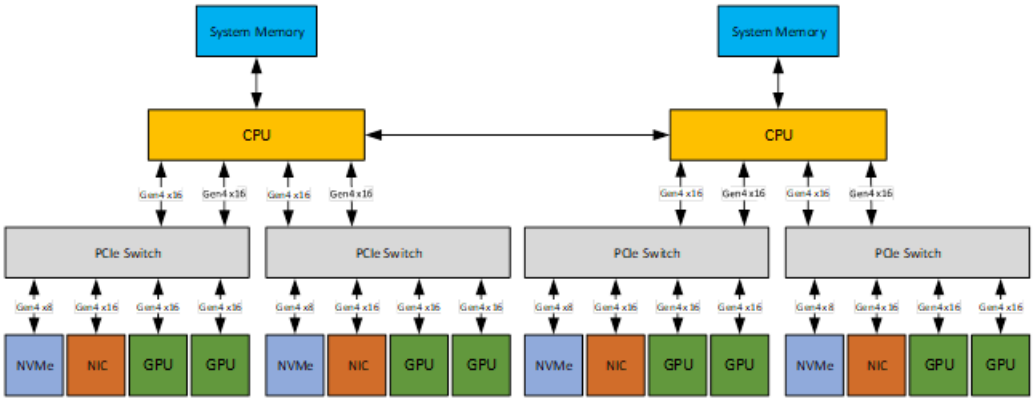


Figure 7. 2P Server with Eight GPUs



Chapter 5. Design Discussion

5.1. GPU

GPUs are best deployed in servers in powers of two (1, 2, 4, or 8 GPUs), with the GPUs evenly distributed across CPU sockets and root ports. That is, a balanced GPU configuration results in optimal GPU application performance and will avoid performance bottlenecks in critical interfaces such as PCIe and system memory.

5.2. CPU

Single-socket (1P) or dual-socket (2P) configurations are allowed. The primary design constraint for the CPU is to choose a configuration with sufficient CPU cores per GPU. At least six physical CPU cores per GPU is recommended.

Depending on the number of GPUs used in the server, and the number of cores available in either an AMD or Intel enterprise class CPU, that will dictate whether a 1P or 2P configuration can be used. 1P server designs are typically used in edge server applications.

NVIDIA has started its evaluation of 4P (four CPU socket) server configurations, but more data is needed before a 4P system configuration recommendation can be made. If partners design a 4P server, the usual recommendation for a balanced GPU configuration apply. That is, distribute the GPUs evenly across all CPU sockets and root ports. Contact NVIDIA if you are interested in designing a 4P server with NVIDIA GPUs.

5.2.1. AMD PCIe Port Mapping

Servers using the AMD Rome and Milan CPUs may suffer from IO performance issues if the correct PCIe ports are not used in the server configuration. Complete details are provided in the *AMD Rome System Configuration Design Guide* (DG-10282-001) by NVIDIA.

The general design guidelines for optimal IO performance and device to host bandwidth when using AMD CPUs are:

- 2 GPUs in 1P – use P0/P3 or P1/P2
- 4 GPUs in 1P – use ports P0/P1/P2/P3
- 4 GPUs in 2P – use P0/P3 or P1/P2 on both sockets

- 8 GPUs in 2P – use ports P0/P1/P2/P3

5.3. System Memory

Total system memory should be at least 1.5 times greater than the total amount of GPU frame buffer memory. At least 2.0 times greater is preferred to best GPU application performance.

For example, in the server with four NVIDIA A100 GPUs – each with 40 GB of frame buffer memory – the total amount of GPU memory is 160 GB. Thus, the total system memory should be at least 240 GB, and 320 GB is preferred.

The system memory should be evenly distributed across all CPU sockets and memory channels for optimal performance.

5.4. PCIe Interface

This section details the PCIe interface including speed and land width, topology, and port mapping.

5.4.1. PCIe Speed and Lane Width

PCIe Gen4 x16 to each GPU is recommended for maximum performance.

- Since Intel Sky Lake and Cascade Lake systems do not support Gen4, a Gen3 x16 connection to each GPU is acceptable.
- PCIe Gen4 x8 links to the GPU are also supported. However, GPU performance may be impacted for some applications.

5.4.2. PCIe Topology

For high volume servers with one or two GPUs installed, it is likely that a PCIe switch is not needed to connect the PCIe devices (GPUs, NICs and NVMe drives) to the host PCIe root ports. This is especially true with the AMD Rome and Milan CPUs and the upcoming Intel Ice Lake CPUs, which all provide sufficient PCIe Gen4 lanes for device connectivity.

- In general, it is strongly recommended to distribute the GPUs across the CPU sockets and root ports for a balanced configuration. One Gen4 x16 link per GPU is strongly recommended for best performance.
- As the number of GPUs and other PCIe devices increases for larger, higher performance servers, PCIe switches may be needed to provide sufficient connectivity. In that case, the number of PCIe switch layers should be limited to one or two, to minimize PCIe latency.
- If a PCIe switch is used, a maximum of two GPUs should share an upstream PCIe Gen4 x16 link for optimal performance.

5.4.3. AMD PCIe Port Mapping

Refer to Section AMD PCIe Port Mapping for specific recommendation for using the PCIe ports on the AMD Rome and Milan CPUs.

5.5. Network

For servers designed with GPUs, NVIDIA recommends the use of Mellanox 200Gb/s InfiniBand and Ethernet BlueField-2 data processing units (DPUs) and ConnectX Smart Network adapters. The NVIDIA-Certified systems testing process requires a standardized testing environment that uses one of the following NVIDIA Mellanox products:

- ConnectX-6
- ConnectX-6 Dx
- BlueField-2

Customers may purchase and deploy NVIDIA-Certified systems with their choice of networking adaptors. NVIDIA software support services are available on any NVIDIA-Certified system regardless of the networking used.

The ConnectX-6 VPI HDR / 200 Gb/s network adapter accelerates HPC and AI workloads for optimal performance and efficiency. ConnectX-6 Dx Ethernet secure network cards accelerate a wide number of use cases and workloads with offload capability for networking, virtualization, security, storage, telco, streaming, and other workloads.

BlueField-2 is the world's most advanced data processing unit (DPU), delivering software-defined, hardware-accelerated networking, storage, and security to every host with zero CPU overhead. BlueField-2 DPUs integrate a ConnectX-6 Dx with a set of powerful Arm cores, security engines, and high bandwidth memory interface into a single SOC.

The following table details what network protocols are supported and what security protocols are offloaded by NVIDIA Mellanox networking cards.

Table 3. Supported Network Protocols and Offloaded Security Protocols

SmartNIC/DPU	Network Protocol	Security Offload Engine		
		IPsec	TLS	AES-XTS
ConnectX-6	Ethernet/ InfiniBand Up to 2x200 Gb/s	No	No	Yes
ConnectX-6 Dx	Ethernet Up to 1x100 Gb/s	Yes	Yes	Yes
BlueField-2	Ethernet/ InfiniBand	Yes	Yes	Yes

SmartNIC/DPU	Network Protocol	Security Offload Engine		
		IPsec	TLS	AES-XTS
	Up to 1x100 Gb/s			

5.5.1. NIC Speeds

The following PCIe recommendations apply to the different NIC speeds used in the server design for best performance.

- Each 200 Gbps NIC needs Gen4 x16
- Each 100 Gbps NIC needs Gen4 x8 or Gen3 x16

5.6. Storage

NVMe drives are recommended for optimal application performance. However, for NVIDIA-Certified 2.0, there is no minimum number of NVMe drives required, and NVMe is not needed to be NVIDIA-Certified.

In server configurations that use PCIe switches, the NVMe drives should be located under the same PCIe switch as the GPU(s) for best NVIDIA® GPUDirect® Storage (GDS) performance.

GDS is supported for NVMe drives located in the same server as the GPUs, or in external storage servers. In all cases the GPU and NIC must sit under the same PCIe switch, or under the same PCIe root complex if PCIe switches are not present in the server design. If NVMe drives are present in the server, they must be on the same PCIe switch or root complex to be used with GDS.

For more details on GPUDirect Storage, refer to the *GPUDirect Storage Design Guide* (DG-09719-001).

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation ("NVIDIA") makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. 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 document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information 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 ("Terms of Sale"). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

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 this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order 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 document. NVIDIA accepts no 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 document 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 document. 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 document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND 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. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. 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 Terms of Sale for the product.

Trademarks

NVIDIA and the NVIDIA logo are trademarks and/or registered trademarks of NVIDIA Corporation in the Unites States 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
www.nvidia.com