NVIDIA A100 40GB PCIe GPU Accelerator

Product Brief

PB-10137-001_v03 | September 2020

Document History

PB-10137-001_v03

Version	Date	Authors	Description of Change
01	September 14, 2020	AS, SM	Initial Release
02	September 15, 2020	AS, SM	Removed "confidential" markings
03	September 22, 2020	AS, SM	Updated naming convention for vCompute Server to NVIDIA Virtual Compute Server (Table 3)

Table of Contents

List of Figures

List of Tables

Overview

The NVIDIA® A100 GPU is a dual-slot 10.5 inch PCI Express Gen4 card based on the NVIDIA Ampere GA100 graphics processing unit (GPU). It uses a passive heat sink for cooling, which requires system air flow to properly operate the card within its thermal limits. The A100 PCIe supports double precision (FP64), single precision (FP32) and half precision (FP16) compute tasks, unified virtual memory, and page migration engine.

For performance optimization, NVIDIA GPU Boost™ feature is supported. NVIDIA GPU Boost automatically and dynamically adjusts the GPU clock during runtime to optimize performance within the power cap and thermal limits.

A100 PCIe boards are shipped with ECC enabled by default to protect the GPU's memory interface and the on-board memories. ECC protects the memory interface by detecting any single, double, and all odd-bit errors. The GPU will retry any memory transaction that has an ECC error until the data transfer is error-free. ECC protects the DRAM content by fixing any single-bit errors and detecting double-bit errors. The A100 with 40 GB of HBM2 memory has native support for ECC and has no ECC overhead, both in memory capacity and bandwidth.

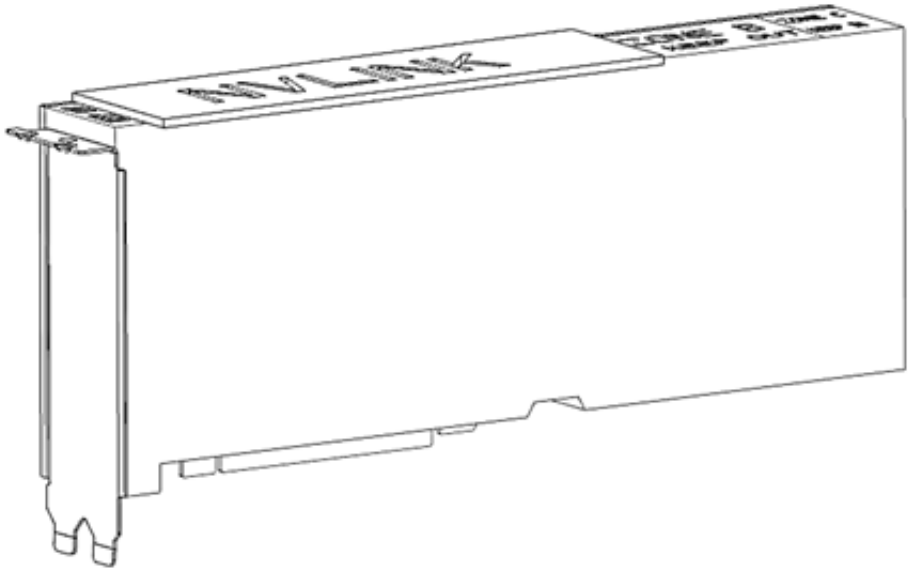
The NVIDIA A100 GPU operates unconstrained up to its thermal design power (TDP) level of 250 W to accelerate applications that require the fastest computational speed and highest data throughput.

For more information on Tensor Cores, download the white paper at https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf

The thermal requirements for A100 are similar to those of the NVIDIA V100S product. See the thermal section for further details. Refer to the following website for the latest list of qualified A100 servers:

https://www.nvidia.com/en-us/data-center/tesla/tesla-qualified-servers-catalog/

Figure 1. NVIDIA A100 with NVLink Bridge



Specifications

Product Specifications

Table 1 through Table 3 provide the product, memory, and software specifications for the NVIDIA A100 GPU card.

Table 1. Product Specifications

Specification	NVIDIA A100
Product SKU	P1001 SKU 200 NVPN: 699-21001-0200-xxx
Total board power	250 W
Thermal solution	Passive
Mechanical Form Factor	Full-height, full-length (FHFL) 10.5", dual-slot
GPU SKU	GA100-883AA-A1
PCI Device IDs	Device ID: 0x20F1 Vendor ID: 0x10DE Sub-Vendor ID: 0x10DE Sub-System ID: 0x145F
GPU clocks	Base: 765 MHz Boost: 1410 MHz
VBIOS	EEPROM size: 8 Mbit UEFI: Supported
PCI Express interface	PCI Express 4.0 ×16 Lane and polarity reversal supported
Power connectors and headers	One CPU 8-pin auxiliary power connector
Weight	Board: 1240 Grams (excluding bracket and extenders) Bracket with screws: 20 Grams Long offset extender: 64 Grams Straight extender: 39 Grams

Table 2. Memory Specifications

Specification	Description
Memory clock	1215 MHz
Memory type	HBM2
Memory size	40 GB
Memory bus width	5120 bits
Peak memory bandwidth	Up to 1555 GB/s

Table 3. Software Specifications

Specification	Description[1]
SR-IOV support	Supported -- 16 VF (virtual functions)
BAR address (physical function)	BAR0: 16 MiB[1] BAR1: 64 GiB[1] BAR3: 32 MiB[1]
BAR address (virtual function)	BAR0: 512 MiB, (256 KiB per VF)[1] BAR1: 64 GiB, 64-bit (4 GiB per VF)[1] BAR3: 512 MiB, 64-bit (32 MiB per VF)[1]
Message signaled interrupts	MSI-X: Supported MSI: Not supported
ARI Forwarding	Supported
Driver Support	R450.x
NVIDIA® CUDA® Support	CUDA 11.x (or later)
Virtual GPU Software Support	Supports vGPU 11.x (or later): NVIDIA Virtual Compute Server Edition
NVIDIA® NGC-Ready™ Test Suite	NGC-Ready 2.x (or later)
PCI class code	0x03 – Display Controller
PCI sub-class code	0x02 – 3D Controller
ECC support	Enabled (by default). Can be disabled via software
SMBus (8-bit address)	0x9E (write), 0x9F (read)
SMBus direct access	Supported
SMBPBI SMBus Post-Box Interface)	Supported

Note:

[1]The KiB, MiB and GiB notation emphasizes the "power of two" nature of the values. Thus,

- 256 KiB = 256 x 1024
- 16 MiB = 16 x 1024^2
- 64 GiB = 64 x 1024^3

The operator is given the option to configure this power setting to be persistent across driver reloads or to revert to default power settings upon driver unload.

Multi-Instance GPU Support

The A100 PCIe card supports Multi-Instance GPU (MIG) capability by providing up to 7 GPU instances per NVIDIA A100 GPU. MIG technology can partition the A100 GPU into individual instances, each fully isolated with its own high-bandwidth memory, cache, and compute cores, enabling optimized computational resource provisioning and quality of service (QoS).

For detailed information on MIG provisioning and use, consult the *Multi-Instance GPU User Guide*: https://docs.nvidia.com/datacenter/tesla/mig-user-guide/index.html

Environmental and Reliability Specifications

Table 5 provides the environment conditions specifications for the A100 PCIe card.

Table 4. Board Environmental and Reliability Specifications

Specification	Description
Ambient operating temperature	0 °C to 50 °C
Storage temperature	-40 °C to 75 °C
Operating humidity	5% to 95% relative humidity
Storage humidity	5% to 95% relative humidity
Mean time between failures (MTBF)	Uncontrolled environment:[1] 945,568 hours at 35 °C Controlled environment:[2] 1,303,691 hours at 35 °C

Notes:

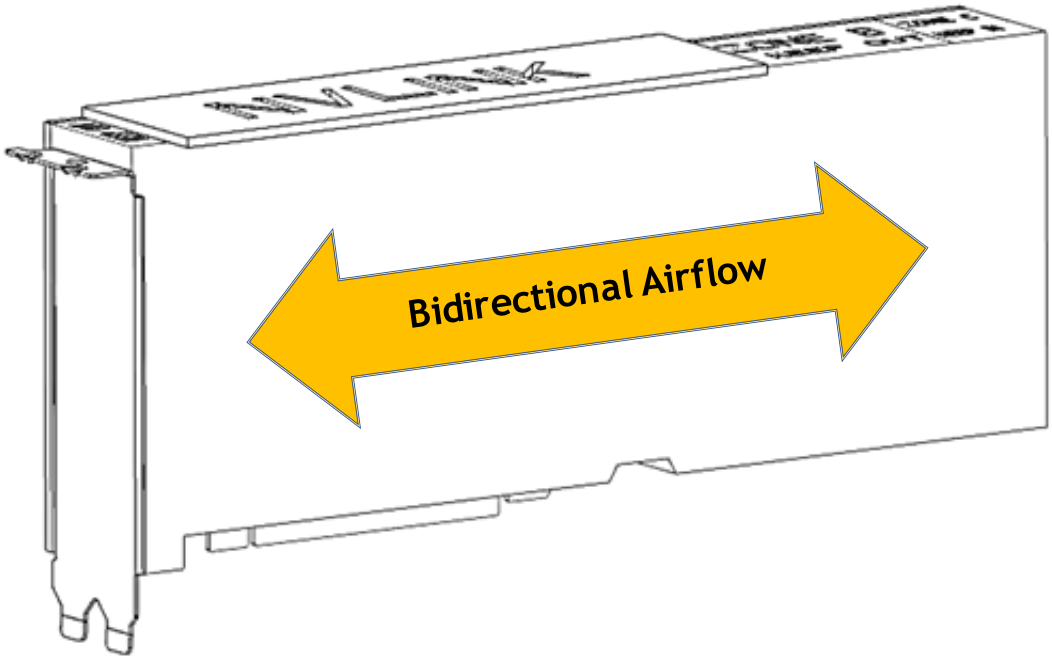
[1]Some environmental stress with limited maintenance (GF35).

[2]No environmental stress with optimum operation and maintenance (GB35).

Airflow Direction Support

The NVIDIA A100 PCIe card employs a bidirectional heat sink, which accepts airflow either left-to-right or right-to-left directions.

Figure 2. A100 Airflow Directions



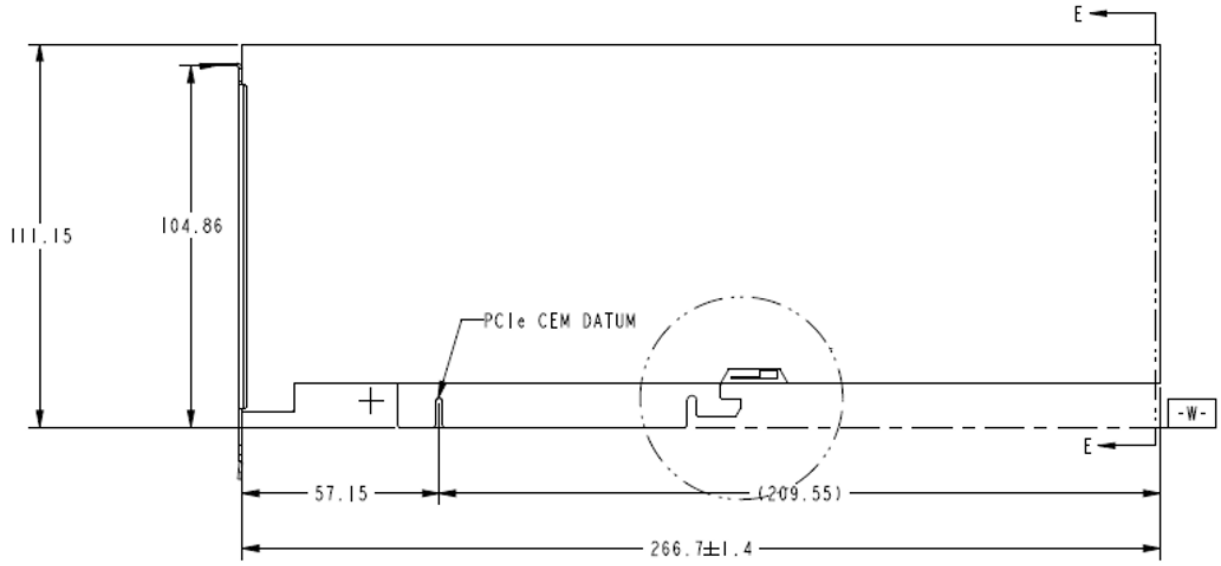
Product Features

Form Factor

The NVIDIA A100 PCIe card conforms to NVIDIA Form Factor 5.0 specification for a full-height, full-length (FHFL) dual-slot PCIe card. For details refer to the *NVIDIA Form Factor 5.0 Specification* (NVOnline reference number 1052306).

In this product brief, nominal dimensions are shown. For tolerances, see the 2D mechanical drawings identified in the "Mechanical Collateral" section of the product specification.

Figure 3. NVIDIA A100 PCIe Card Dimensions

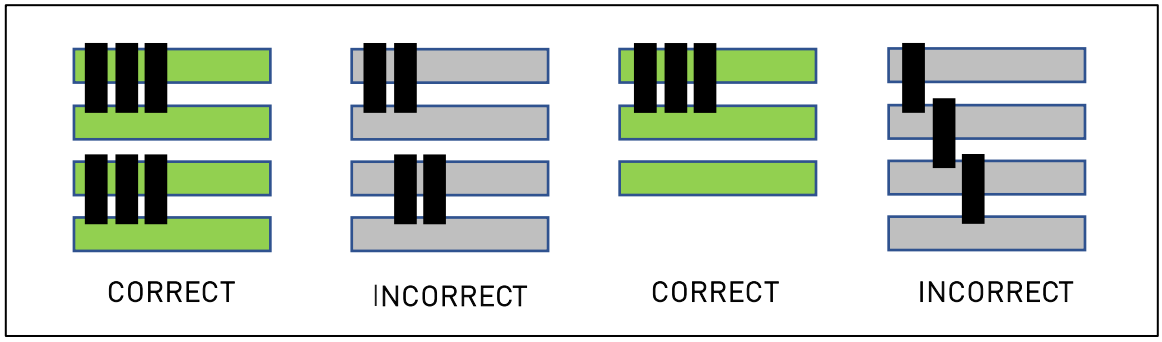


NVLink Bridge Support

NVIDIA® NVLink® is a high-speed point-to-point peer transfer connection, where one GPU can transfer data to and receive data from one other GPU. The NVIDIA A100 card supports NVLink bridge connection with a single adjacent A100 card.

Each of the three attached bridges spans two PCIe slots. To function correctly as well as to provide peak bridge bandwidth, bridge connection with an adjacent A100 card must incorporate all three NVLink bridges. Wherever an adjacent pair of A100 cards exists in the server, for best bridging performance and balanced bridge topology, the A100 pair should be bridged. Figure 4 illustrates correct and incorrect A100 NVLink connection topologies.

Figure 4. NVLink Topology – Top Views



For systems that feature multiple CPUs, both A100 cards of a bridged card pair should be within the same CPU domain—that is, under the same CPU's topology. Ensuring this benefits workload application performance. The only exception is for dual CPU systems wherein each CPU has a single A100 PCIe card under it; in that case, the two A100 PCIe cards in the system may be bridged together.

A100 NVLink speed and bandwidth are given in the following table.

Table 5. A100 NVLink Speed and Bandwidth

Parameter	Value
Total NVLink bridges supported by NVIDIA A100	3
Total NVLink Rx and Tx lanes supported	96
Data rate per NVIDIA A100 NVLink lane (each direction)	50 Gbps
Total maximum NVLink bandwidth	600 Gbytes per second

NVLink Connector Placement

Figure 5 shows the connector keep-out area for the NVLink bridge support of the A100.

Figure 5. NVLink Connector Placement – Top View

Sufficient clearance must be provided both above the north edge of the card and behind the backside of the card's PCB to accommodate NVIDIA A100 NVLink bridges. The clearance above the card's north edge should meet or exceed 2.5 mm. The backside clearance (from the rear card's rear PCB surface) should meet or exceed 2.67 mm. Consult *NVIDIA Form Factor 5.0 for Server Cards* for more specifics.

NVLink bridge interfaces of the A100 PCIe card include removable caps to protect the interfaces in non-bridged system configurations.

Power Connector Placement

The board provides a CPU 8-pin power connector on the east edge of the board.

Figure 6. CPU 8-Pin Connector

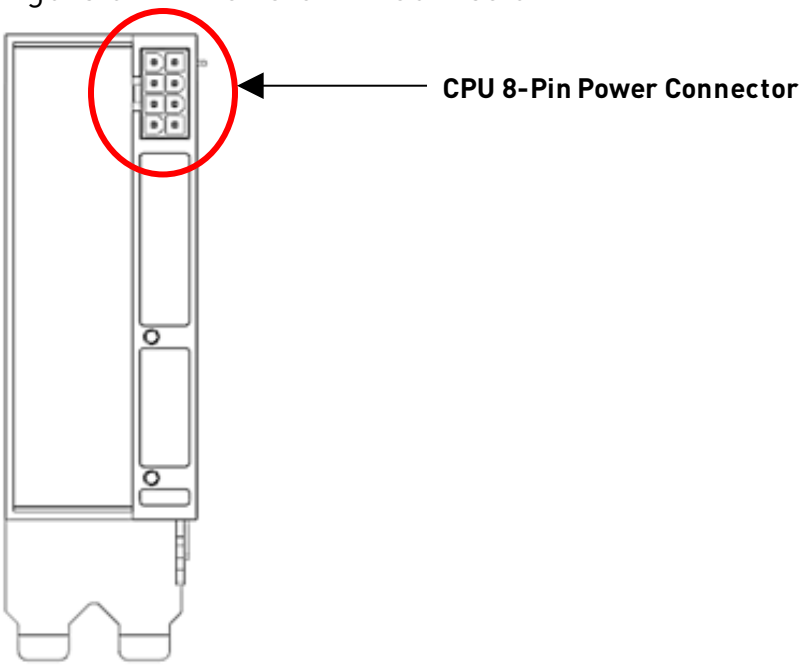


Table 7 lists supported auxiliary power connections for the NVIDIA A100 GPU card.

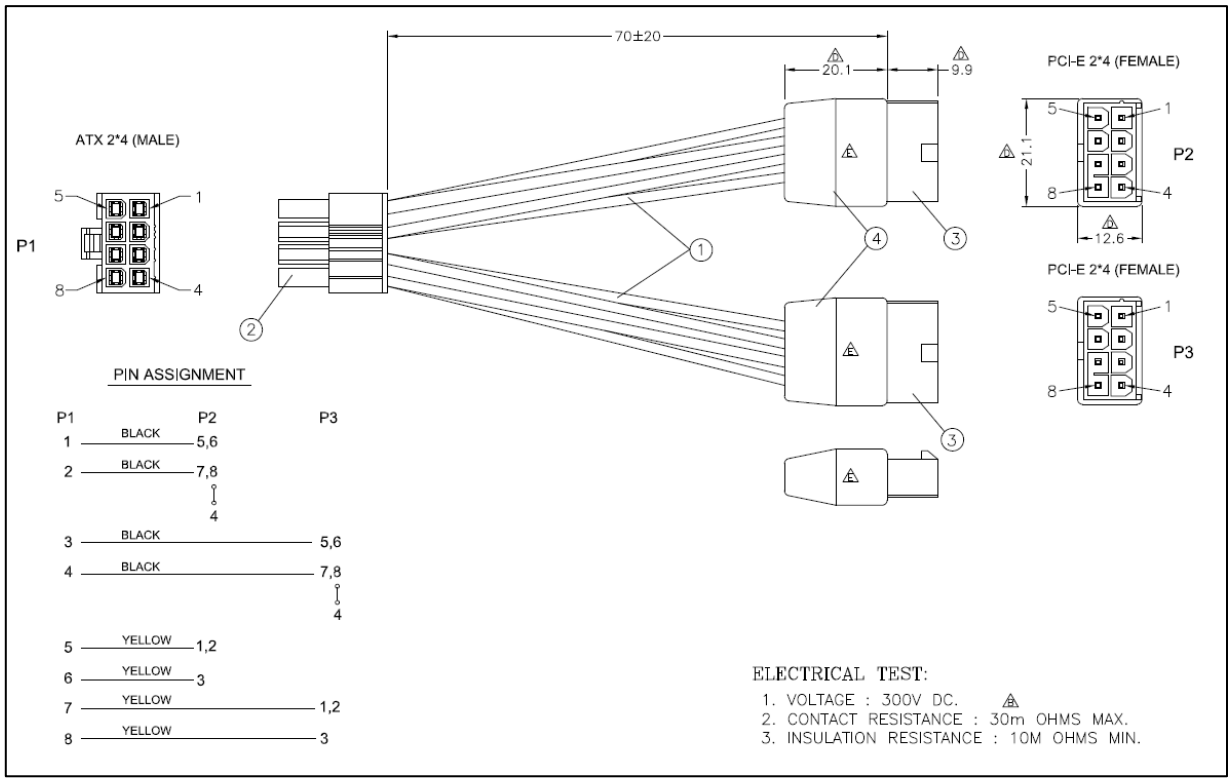
Table 6. Supported Auxiliary Power Connections

Board Connector	PSU Cable
CPU 8-pin	1× CPU 8-pin cable
CPU 8-pin	CPU 8-pin to PCIe 8-pin cable adapter

CPU 8-Pin to PCIe 8-Pin Power Adapter

Figure 7 lists the pin assignments of the power adapter.

Figure 7. CPU 8-Pin to PCIe 8-Pin Power Adapter



Extenders

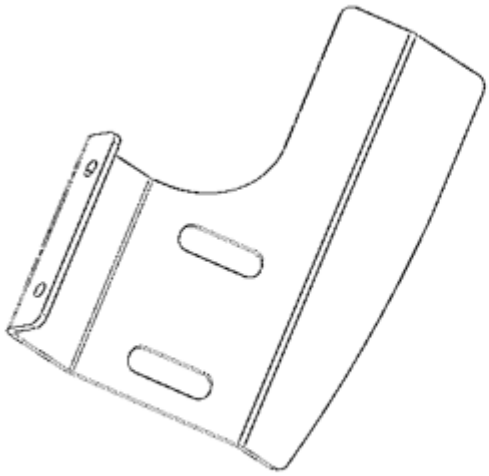
The A100 PCIe card provides two extender options, shown in Figure 8.

- NVPN: 682-00003-5555-006 – Long offset extender
  - Card + extender = 339 mm
- NVPN: 682-00003-5555-007 – Straight extender
  - Card + extender = 312 mm

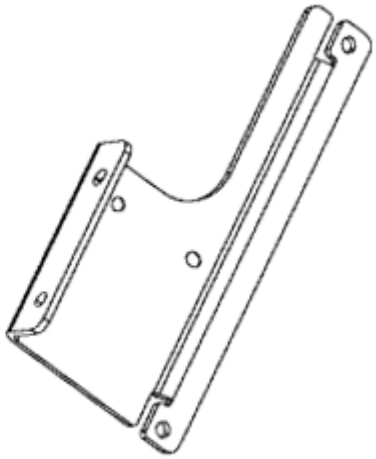
Using the standard NVIDIA extender ensures greatest forward compatibility with future NVIDIA product offerings.

If the standard extender will not work, OEMs may design a custom attach method using the extender mounting holes on the east edge of the PCIe card.

Figure 8. Extenders



LONG OFFSET EXTENDER



STRAIGHT EXTENDER

Languages Supported

Table 8 shows the languages supported for NVIDIA A100 40GB PCIe GPU Accelerator.

Table 7. Languages Supported

Languages	Windows[1]	Linux
English (US)	Yes	Yes
English (UK)	Yes	Yes
Arabic	Yes	
Chinese, Simplified	Yes	
Chinese, Traditional	Yes	
Czech	Yes	
Danish	Yes	
Dutch	Yes	
Finnish	Yes	
French (European)	Yes	
German	Yes	
Greek	Yes	
Hebrew	Yes	
Hungarian	Yes	
Italian	Yes	
Japanese	Yes	
Korean	Yes	
Norwegian	Yes	
Polish	Yes	
Portuguese (Brazil)	Yes	
Portuguese (European/Iberian)	Yes	
Russian	Yes	
Slovak	Yes	
Slovenian	Yes	

Languages	Windows[1]	Linux
Spanish (European)	Yes	
Spanish (Latin America)	Yes	
Swedish	Yes	
Thai	Yes	
Turkish	Yes	

Note:

[1]Microsoft Windows 7, Windows 8, Windows 8.1, Windows 10, Windows Server 2008 R2, Windows Server 2012 R2, and Windows 2016 are supported.

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, the NVIDIA logo, CUDA, NGC-Ready, NVIDIA GPU Boost, NVLink, and Tesla 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