NVIDIA A16 PCIe GPU Accelerator

Product Brief

PB-10518-001_v02 | March 2022

Document History

PB-10518-001_v02

Version	Date	Authors	Description of Change
01	June 29, 2021	VK, SM	Initial Release
02	March 17, 2022	VK, SM	• Removed Secure Boot from Table 1 • Updated Secure Boot and CEC description in Table 3 • Updated Table 4 "Board Environmental and Reliability Specifications" • Updated "Root of Trust" section • Added Table 5 "Root of Trust Feature Set"

Table of Contents

List of Figures

List of Tables

Overview

NVIDIA A16 is a PCI Express Gen4 graphics processing unit (GPU) card that is ideal for providing high-user density for Virtual Desktop Infrastructure (VDI) environments. It is a full height, full length (FHFL) design with four GPUs on a single board. The A16 is a dual-slot card featuring 64 GB of GDDR6 memory and a 250 W maximum power limit. The A16 also supports x16 PCIe Gen4 connectivity. It is a passively cooled card with a superior thermal design that requires system airflow to operate and handles challenging ambient environments with ease (NEBS-3 capable).

Powered by NVIDIA Ampere architecture, the NVIDIA A16 provides the highest encoder throughput and frame buffer for the best user experience in a VDI environment using NVIDIA Virtual PC (vPC) software. Video transcoding and Android™ cloud gaming are among the other workloads that can take advantage of the multiple encoders and decoders on the A16 GPU. The quad GPU design enables the highest frame buffer, encoder, and decoder density in a dual-slot form factor for VDI use cases.

Refer to the following website for the latest list of servers qualified with NVIDIA A16: https://www.nvidia.com/en-us/data-center/tesla/tesla-qualified-servers-catalog/

Figure 1. Single Board with Four GPUs

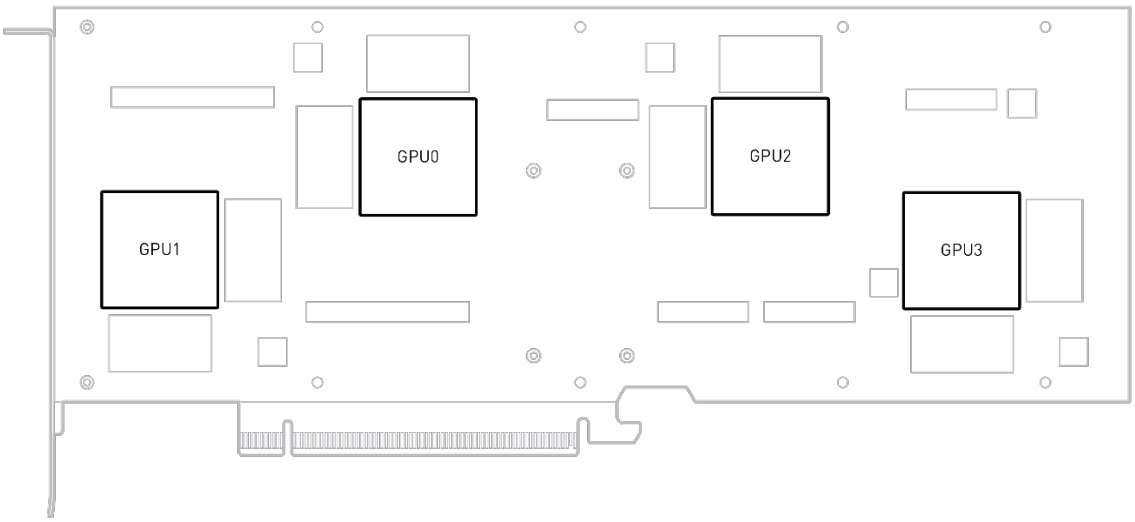
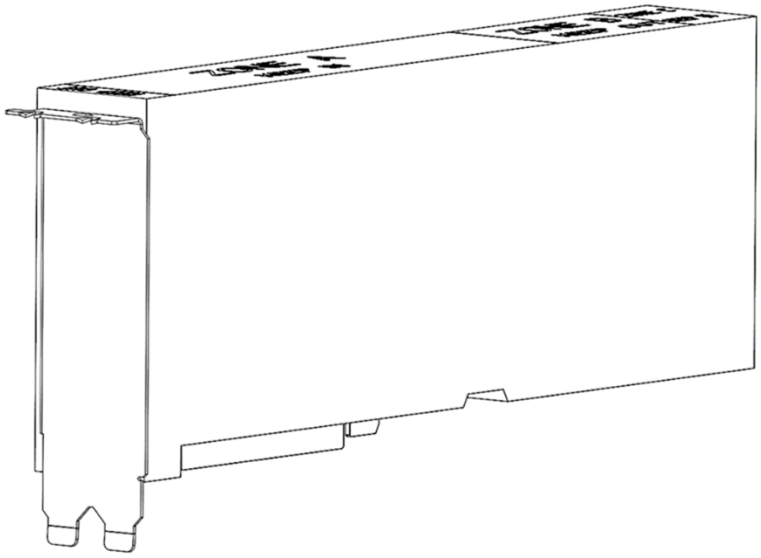


Figure 2. NVIDIA A16 PCIe Card



Specifications

Product Specifications

Table 1 through Table 3 provide the product, memory, and software specifications for the NVIDIA A16 PCIe card.

Table 1. Product Specifications

Specification	NVIDIA A16
Product SKU	PG171 SKU 200 NVPN: 699-2G171-0200-xxx
Total board power	250 W default 250 W maximum 235 W MaxQ (maximum performance per watt) 195 W minimum
Thermal solution	Passive
Mechanical Form Factor	Full-height, full-length (FHFL) 10.5", dual-slot
GPU SKU	GA107-890 (quantity four)
Board PCI Device IDs	Device ID: 0x25B6 Vendor ID: 0x10DE Sub-Vendor ID: 0x10DE Sub-System ID: 0x14A9
Switch PCI Device IDs	Device ID: 0x4123 Vendor ID: 0x15B3 Sub-Vendor ID: 0x15B3 Sub-System ID: 0x1103
GPU clocks	Base: 1312 MHz Boost: 1755 MHz
VBIOS	EEPROM size: 8 Mbit UEFI: Supported
PCI Express interface	Physical x16 PCIe lanes PCIe Gen4 x16 or Gen3 x16 supported

Specification	NVIDIA A16
	Lane and polarity reversal supported
Zero Power	Not supported
Weight	Board: 1088 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	6.25 GHz
Memory type	GDDR6
Memory size	64 GB (16 GB per GPU)
Memory bus width	128 bits
Peak memory bandwidth	Up to 4x 200 GB/sec

Table 3. Software Specifications

Specification	Description[1]
SR-IOV support	Supported: 16 VF (virtual functions) per GPU
BAR address (physical function)	BAR0: 16 MiB[1] BAR1: 16 GiB[1] BAR3: 32 MiB[1]
BAR address (virtual function)	BAR0: 4 MiB, (16 VF x 256 KiB per VF)[1] BAR1: 32 GiB, 64-bit (16 VF x 2 GiB per VF)[1] BAR3: 512 MiB, 64-bit (16 VF x 32 MiB per VF)[1]
Message signaled interrupts	MSI-X: Supported MSI: Not supported
ARI Forwarding	Supported
Driver support	R470 or later
Secure Boot	Supported (see "Root of Trust" section)
CEC Firmware	v4.01 or later (for CEC-enabled cards)
NVIDIA® CUDA® support	CUDA 11.4 or later
Virtual GPU software support	Supports vGPU 13.0 or later Supports NVIDIA vPC, vApps, NVIDIA RTX™ vWS, vCS
NVIDIA® NGC-Ready™ test suite	NGC-Next Certification 2.x or later
PCI class code	0x03 – Display Controller

Specification	Description[1]
PCI sub-class code	0x02 – 3D Controller
ECC support	Enabled (by default). Can be disabled via software
SMBus (8-bit address)	GPU1: 0x9E (write), 0x9F (read) GPU2: 0x9C (write), 0x9D (read) GPU3: 0x9A (write), 0x9B (read) GPU4: 0x98 (write), 0x99 (read)
Reserved I2C addresses[2]	CEC1: 0xAA (write) CEC2: 0xAC (write)
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

[2]See Section "Root of Trust" of this product brief

Environmental and Reliability Specifications

Table 4 provides the environment conditions specifications for the A16 PCIe card.

Table 4. Board Environmental and Reliability Specifications

Specification	Description
Ambient operating temperature	0 °C to 50 °C
Ambient operating temperature (short term)[1]	-5 °C to 55 °C
Storage temperature	-40 °C to 75 °C
Operating humidity (short term)[1]	5% to 93% relative humidity
Operating humidity	5% to 85% relative humidity
Storage humidity	5% to 95% relative humidity
Mean time between failures (MTBF)	Uncontrolled environment:[2] 677,793 hours at 35 °C Controlled environment:[3] 874,410 hours at 35 °C

Notes: Specifications in this table are applicable up to 6000 feet.

[1]A period not more than 96 hours consecutive, not to exceed 15 days per year.

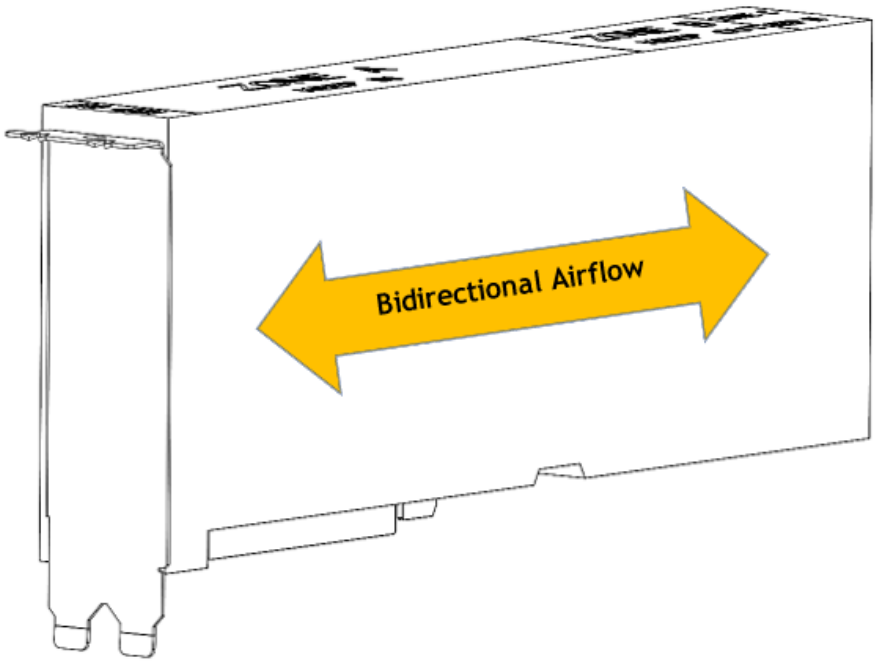
[2]Some environmental stress with limited maintenance (GF35).

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

Airflow Direction Support

The NVIDIA A16 PCIe card employs a bidirectional heat sink, which accepts airflow either left-to-right or right-to-left directions.

Figure 3. A16 Airflow Direction



Product Features

PCI Express Interface Specifications

The following subsections describe the PCIe interface specifications for the A16 PCIe card.

PCIe Support

The A16 card supports PCIe Gen4. Gen4 x16 interface should be used when connecting to the A16 PCIe card. For optimal GPU performance, a Gen4 ×16 connection is recommended. However, Gen3 ×16 link connection is supported as well.

Polarity Inversion and Lane Reversal Support

Lane Polarity Inversion, as defined in the PCIe specification, is supported on the A16 PCIe card.

Lane Reversal, as defined in the PCIe specification, is supported on the A16 PCIe card. When reversing the order of the PCIe lanes, the order of both the Rx lanes and the Tx lanes must be reversed.

Root of Trust

The NVIDIA A16 GPU has a primary root of trust within the GPU chip that provides the following:

- Secure boot
- Secure firmware upgrade
- Firmware rollback protection
- Ability to disable In-band firmware update (established after each GPU reset)
- Secure application processor recovery

In addition, NVIDIA offers some A16 PCIe boards with an onboard CEC1712 chip, acting as a secondary root of trust, extending the security capabilities allowing for firmware attestation, key revocation, and out-of-band firmware updates. The CEC1712 device authenticates the contents of the GPU firmware ROM before permitting the GPU to boot from its ROM. For

CEC1712-enabled cards, the root of trust feature occupies up to two I2C addresses (in addition to the SMBus addresses). I2C addresses 0xAA and 0xAC should therefore be avoided for system use.

Identification of the two variants of A16 PCIe boards (with or without CEC1712) can be done using the 900-level part number on the back of the GPU or running the nvidia-smi -q command.

- 900-2G171-XXXX-1XX A16 GPUs without CEC1712 (secondary root of trust)
- 900-2G171-XXXX-0XX A16 GPUs with CEC1712 (secondary root of trust)

The following table shows the features that are available using the primary and secondary root of trust.

Table 5. Root of Trust Feature Set

Features	Primary Root of Trust within GPU Chip	Secondary Root of Trust Using External CEC Chip on Board
Secure Boot	Yes	Yes
Secure Firmware Upgrade	Yes	Yes
Firmware Rollback Protection	Yes	Yes
In-Band Firmware Update Disable	Yes[1]	Yes
Key Revocation	No	Yes
Firmware Attestation	No	Yes

Notes:

[1] "In-Band Firmware Update Disable" feature must be established after every GPU reset.

nvidia-smi

nvidia-smi is an in-band monitoring tool provided with the NVIDIA driver and can be used to set the maximum power consumption with driver running in persistence mode. An example command to enable Max-Q with a power limit of 195 W is shown:

```
nvidia-smi -pm 1
nvidia-smi -pl 195
```

To restore the A16 back to its default TDP power consumption, either the driver module can be unloaded and reloaded, or the following command can be issued:

```
nvidia-smi -pl 250
```

SMBPBI

An out-of-band channel exists through the SMBus Post-Box Interface (SMBPBI) protocol to set the power limit of the GPU. This also requires that the NVIDIA driver be loaded for full functionality. Max-Q mode can be enabled through the following asynchronous command:

Table 6. SMBPBI Commands

Specification	Value
Opcode	10h – Submit/poll asynchronous request
Arg1	0x01 – Set total GPU power limit
Arg2	0x00

Power Connector Placement

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

Figure 4. CPU 8-Pin Power Connector

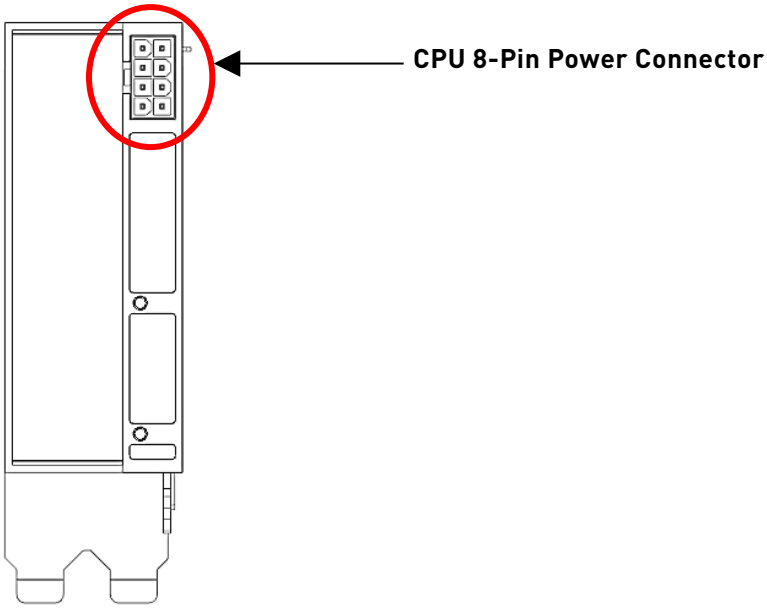


Table 7 lists supported auxiliary power connections for the NVIDIA A16 PCIe card.

Table 7. Supported Auxiliary

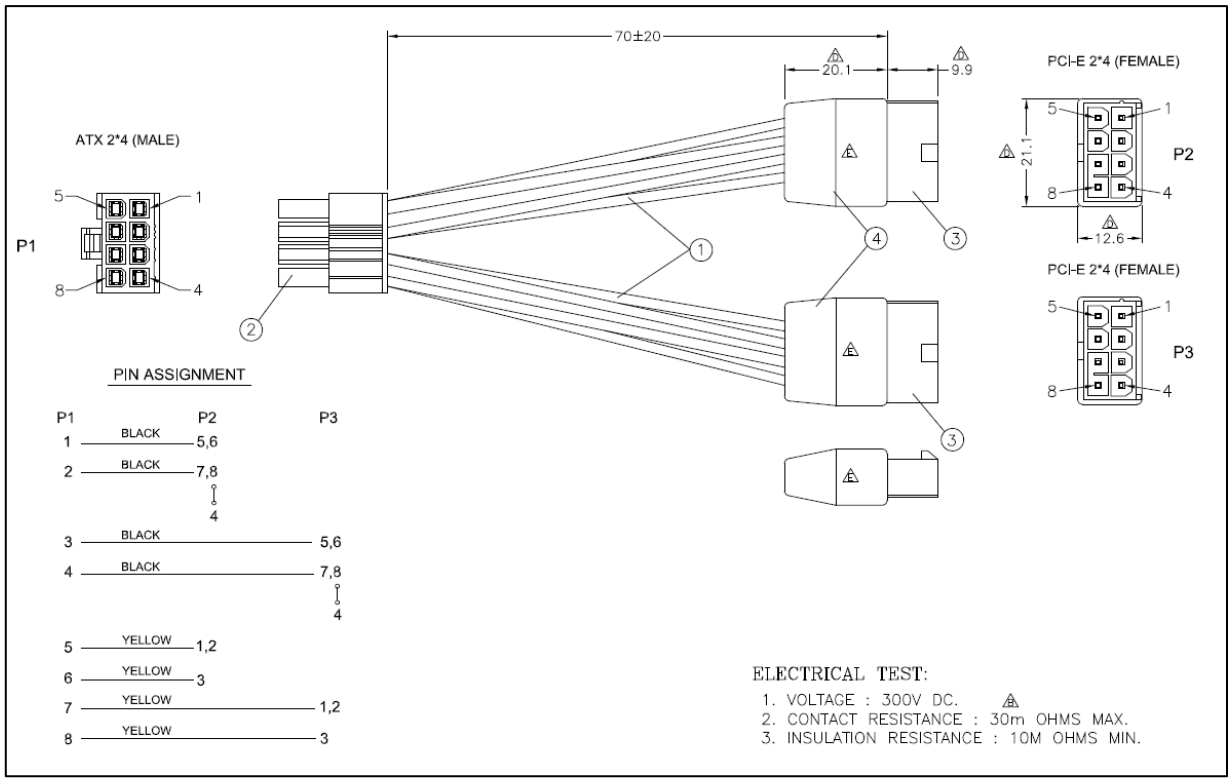
Board Connector	PSU Cable
CPU 8-pin	1x CPU 8-pin cable

CPU 8-Pin to PCIe 8-Pin Power Adapter

Figure 5 lists the pin assignments of the power adapter. Consult NVIDIA Applications Engineering for qualified suppliers of the power adapter.

The CPU 8-pin smart power adapter NVPN is 030-1233-000.

Figure 5. CPU 8-Pin to PCIe 8-Pin Power Adapter



Support Information

Certifications

- Windows Hardware Quality Lab (WHQL):
  - Certified Windows 10
  - Certified Windows Server 2008 R2, Windows Server 2012 R2
- Ergonomic requirements for office work W/VDTs (ISO 9241)
- EU Reduction of Hazardous Substances (EU RoHS)
- Joint Industry guide (J-STD) / Registration, Evaluation, Authorization, and Restriction of Chemical Substance (EU) – (JIG / REACH)
- Halogen Free (HF)
- EU Waste Electrical and Electronic Equipment (WEEE)

Agencies

- Australian Communications and Media Authority and New Zealand Radio Spectrum Management (RCM)
- Bureau of Standards, Metrology, and Inspection (BSMI)
- Conformité Européenne (CE)
- Federal Communications Commission (FCC)
- Industry Canada - Interference-Causing Equipment Standard (ICES)
- Korean Communications Commission (KCC)
- Underwriters Laboratories (cUL, UL)
- Voluntary Control Council for Interference (VCCI)

Languages

Table 8. 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	
Spanish (European)	Yes	
Spanish (Latin America)	Yes	
Swedish	Yes	
Thai	Yes	
Turkish	Yes	

Note:

[1]Microsoft 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, and NVIDIA RTX 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.

Google

Android, Android TV, Google Play, and the Google Play logo are trademarks of Google, Inc.

Copyright

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.nvidia.com