Tesla V100S PCIe GPU Accelerator

Product Brief

PB-09804-001_v01 | January 2020

Document History

PB-09804-001_v01

Version	Date	Authors	Description of Change
01	January 13, 2020	WT, SM	Initial Release

Table of Contents

List of Figures

List of Tables

Overview

The NVIDIA® Tesla® V100S GPU Accelerator for PCIe is a dual-slot 10.5 inch PCI Express Gen3 card with a single NVIDIA Volta™ GV100 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 Tesla V100S 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. By automatically adjusting the GPU clock dynamically, maximum performance is achieved within the power cap limit.

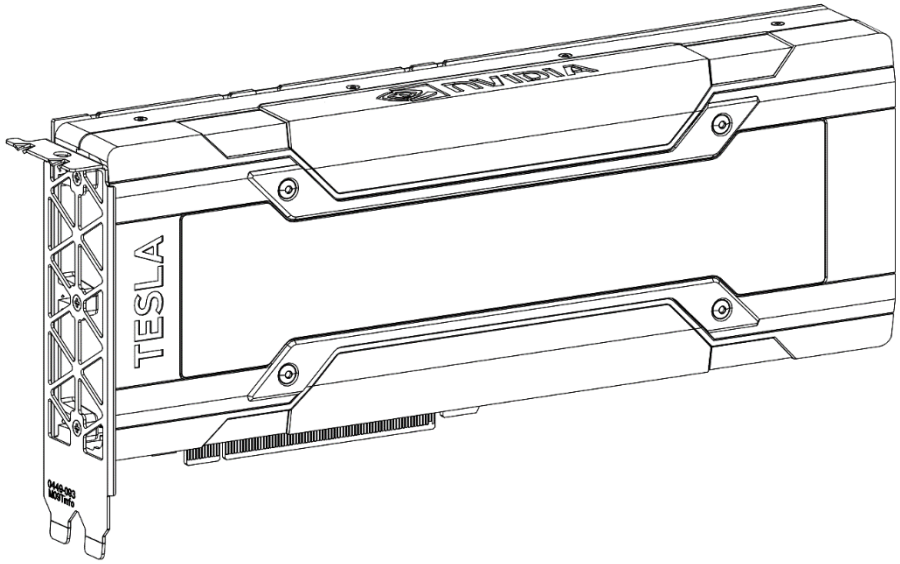
Tesla V100S PCIe boards are shipped with ECC enabled by default to protect the GPU's memory interface and the on-board memories. ECC protects the memory interface by detecting any single, double, and all odd-bit errors. The GPU will retry any memory transaction that has an ECC error until the data transfer is error-free. ECC protects the DRAM content by fixing any single-bit errors and detecting double-bit errors. The Tesla V100S PCIe with 32 GB of HBM2 memory has native support for ECC and has no ECC overhead, both in memory capacity and bandwidth.

Tesla V100S PCIe supports Maximum Performance (Max-P) and Maximum Efficiency (Max-Q) modes. In Max-P mode, the Tesla V100S PCIe Accelerator will operate 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.

Max-Q mode allows data center managers to tune power usage of their Tesla V100S PCIe Accelerators to operate with optimal performance per watt. A power cap limit can be set via software across all GPUs in a rack, reducing power consumption dramatically, while still obtaining excellent rack performance. NVIDIA has provided electrical and thermal specification for Max-Q at 180 W, but customers can provide different power levels for Max-Q depending on the optimal performance point for their target applications.

For more information on Tensor Cores, download the white paper at http://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf

Figure 1. Tesla V100S PCIe Board with Optional I/O Bracket



Specifications

Product Specifications

Table 1 provides the product specifications for the Tesla V100S PCIe board.

Table 1. Product Specifications

Specification	Tesla V100S PCIe 32GB
Product SKUs	NVPN: 699-2G500-0212-XXX
Total board power	Max-P: 250 W (default) Max-Q[1]: 180 W
GPU SKUs	GV100-907A-A1
PCI Device IDs	Device ID: 0x1DF6 Vendor ID: 0x10DE Sub-Vendor ID: 0x10DE Sub-System ID: 0x13D6
GPU clocks	Base: 1267 MHz Maximum boost: 1597 MHz
VBIOS	EEPROM size: 8 Mbit UEFI: Supported
PCI Express interface	PCI Express 3.0 ×16, Lane and polarity reversal supported
Power connectors and headers	One CPU 8-pin auxiliary power connector
Weight	Board: 1196 Grams Bracket with screws: 21 Grams Long offset extender: 52 Grams Straight extender: 42 Grams

Note:

[1]The allowable power range for Max-Q is 100 W to 250 W. Electrical and thermal reference data is provided at 180 W for Max-Q. Other Max-Q power levels must be qualified by the NVIDIA partner.

Table 2 provides the memory specifications for the Tesla V100S PCIe board.

Table 2. Memory Specifications

Specification	Tesla V100S PCIe 32GB
Maximum memory clock	1107 MHz
Memory size	32 GB HBM2
Memory bus width	4096-bit
Peak memory bandwidth	Up to 1134 GB/s

Table 3 provides the software specifications.

Table 3. Software Specifications

Specification	Description
Compatibility mode supported	Compute only
Base address	BAR0: 16 MB BAR1: 32 GB BAR3: 32 MB
PCI class code	0x03 - Display Controller
PCI sub-class code	0x02 – 3D Controller
ECC support	Supported (Enabled by default)
SMBus (8-bit address)	0x9E (write), 0x9F (read)
SMBus direct access	Supported
SMBus Post Box Interface (SMBPBI)	Supported
Max customer boost clock	Supported
Zero Power	Not supported

Max-Q Mode

Max-Q mode, optimized for GPU performance per watt, can be enabled through setting the power limit to the specified Max-Q board power rating. The Max-Q point may vary with a workload from 100 W to 250 W. The characterized Max-Q setting for DGEMM is 180 W.

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 is shown (power limit 180 W):

```
nvidia-smi -pm 1
nvidia-smi -pl 180
```

To restore the SXM2 module back to its default TDP power consumption, you can either unload the driver module and reload, or use the following command:

```
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 SXM2 module, but this also requires that the NVIDIA driver be loaded for full functionality. Max-Q mode can be enabled through the following asynchronous command:

Table 4. SMBPBI Commands

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

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.

PCI Express Interface Specifications

The following sub-section describe the PCIe interface specifications for the Tesla V100S PCIe board.

Polarity and Lane Reversal Support

Polarity and lane reversal features are supported on the Tesla V100S PCIe GPU Accelerator.

Environmental and Reliability Specifications

Table 5 provides the environmental conditions specifications for the Tesla V100S PCIe board

Table 5. Board Environmental and Reliability Specifications

Specification	Description
Ambient operating temperature	0 °C to 45 °C
Storage temperature	-40 °C to 75 °C
Operating humidity	5% to 90% relative humidity
Storage humidity	5% to 95% relative humidity
Mean time between failures (MTBF)	Uncontrolled environment[1]: 1,111,592 hours at 35 °C Controlled environment[2]: 2,051,014 hours at 35 °C

Notes:

[1]Some environmental stress with limited maintenance.

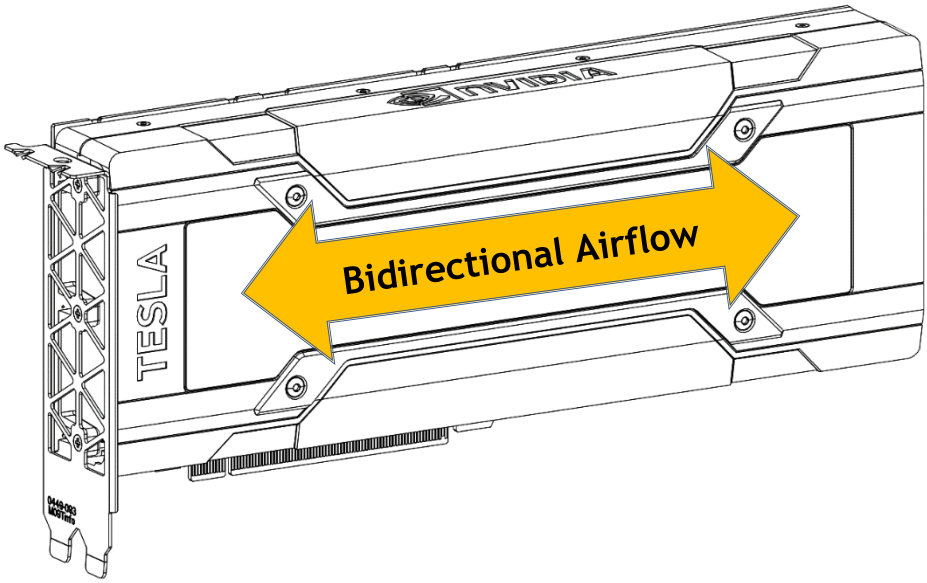
[2]No environmental stress with optimum operation and maintenance.

System Airflow Requirements

Airflow Direction Support

The Tesla V100S PCIe board employs a bidirectional heat sink, which accepts airflow either left-to-right or right-to-left directions.

Figure 2. Tesla V100S PCIe Airflow Directions with Optional I/O Bracket



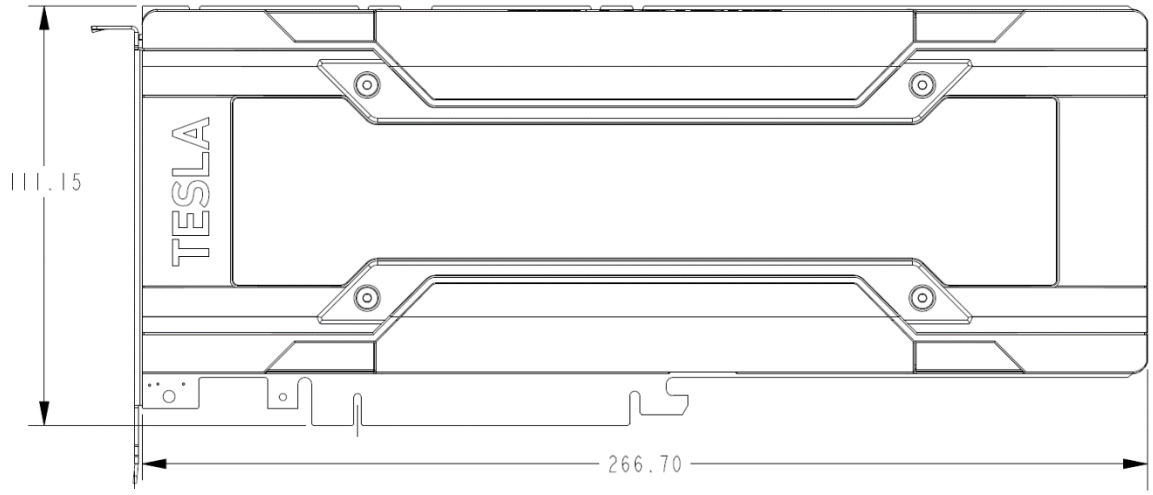
Product Features

Form Factor

The Tesla V100S PCIe board conforms to NVIDIA Form Factor 3.0 specification. For details about NVIDIA Form Factor 3.0 consult the *System Design Guide for NVIDIA Enterprise Products Design Guide* (DG-07562-001) and the NVIDIA Form Factor 3.0 specification.

In this product specification, nominal dimensions are shown; for tolerances, see the attached 2D mechanical drawings.

Figure 3. Tesla V100S PCIe Board Dimensions with Optional I/O Bracket



Power Connector Placement

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

Figure 4. CPU 8-Pin Power Connector with Optional I/O Bracket

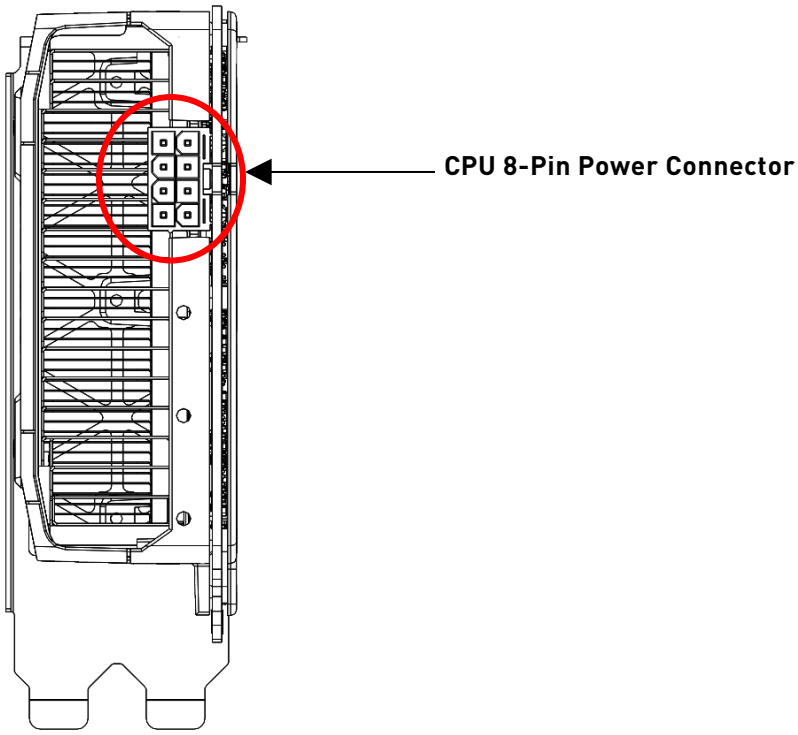


Table 6 lists supported auxiliary power connections for the Tesla V100S PCIe board.

Table 6. Supported Auxiliary Power Connections

Board Connector	PSU Cable
CPU 8-pin	1x CPU 8-pin cable
1x CPU 8-pin cable	2x PCIe 8-pin cable 2x PCIe 6-pin cable[1] 1x PCIe 8-pin cable and 1x PCIe 6-pin cable[1]

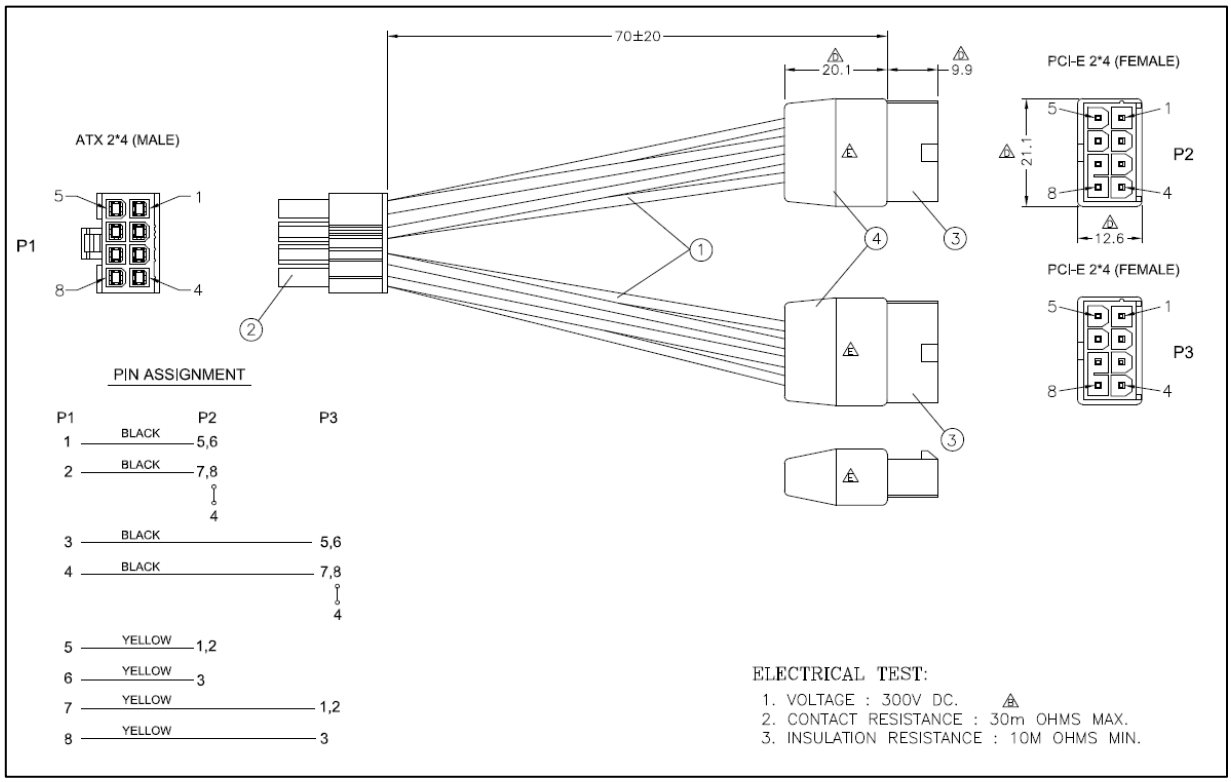
Note:

[1]The PCIe 6-pin cable must be capable of carrying up to 120 W.

CPU 8-Pin to PCIe 8-Pin Dongle

Figure 5 lists the pin assignments of the dongle. The part number for the dongle is NVPN: 030-0571-000.

Figure 5. CPU 8-Pin to PCIe 8-Pin Dongle



Extenders

The Tesla V100S PCIe board provides two extender options as shown in the following figures.

- NVPN: 682-00003-5555-002 –Long offset extender (Figure 6)
  - Card + extender = 339 mm
- NVPN: 682-00003-5555-000 – Straight extender (Figure 7)
  - Card + extender = 312 mm

Figure 6. Long Offset Extender

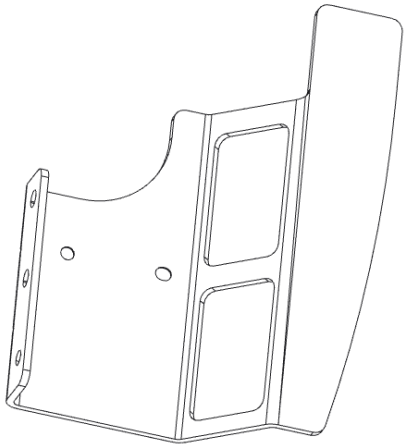
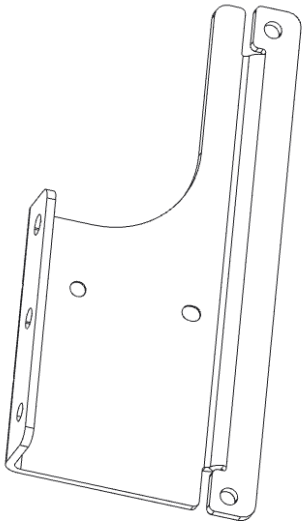


Figure 7. Straight Extender



- 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 heat sink baseplate. The extender mounting holes will vary among NVIDIA products, so designing for flexibility is recommended.

Support Information

Languages

Table 7 lists the languages supported for the Tesla V100S PCIe GPU Accelerator.

Table 7. Languages Supported

Languages	Windows[1]	Linux
English (US)	Yes	Yes
English (UK)	Yes	Yes
ArabicTabhle	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	

Languages	Windows[1]	Linux
Slovak	Yes	
Slovenian	Yes	
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, NVIDIA GPU Boost, NVIDIA Volta, 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