

NVIDIA L40 GPU Accelerator

Product Brief

Document History

PB-11131-001_v03

Version	Date	Authors	Description of Change
01	October 10, 2022	MD, SM	Initial release – Preliminary
02	October 13, 2022	MD, SM	Minor document modification
03	January 20, 2023	MD, SM	• Updated board power and GPU clocks in Table 1
			Updated MTBF in Table 4

Table of Contents

Overview	1
Specifications	2
Product Specifications	2
Environmental and Reliability Specifications	5
Airflow Direction Support	6
Product Features	7
PCI Express Interface Specifications	7
PCIe Support	7
Single Root I/O Virtualization Support	7
Interrupt Messaging	7
Polarity Inversion and Lane Reversal Support	8
Root of Trust	8
Display	8
Display On and Off	8
Display Off Mode	8
Display On 8GB BAR1 Mode	9
Display On 256MB BAR1 Mode	
Switching Operating Modes	9
Frame Lock	9
Display Bracket	9
Programmable Power	
nvidia-smi	
SMBPBI	
IPMI FRU EEPROM	
Form Factor	
Power Connector	
Power Connector Placement	
CPU 8-Pin to PCIe 16-Pin Power Adapter	
Extenders	14
Support Information	15
Certification	15
Agencies	15
Languages	

List of Figures

Figure 1.	NVIDIA L40 PCIe Card	1
-	L40 Airflow Direction	
Figure 3.	NVIDIA L40 PCIe Card Dimensions	11
Figure 4.	PCIe 16-Pin Power Connector	12
Figure 5.	PCIe 16-Pin Power Connector Pin Assignments	12
Figure 6.	CPU 8-Pin to PCle 16-Pin Power Adapter	13
Figure 7.	CPU 8-Pin to PCle 16-Pin Power Adapter Pin Assignments	14

List of Tables

Table 1.	Product Specifications	2
Table 2.	Memory Specifications	3
Table 3.	Software Specifications	3
Table 4.	Board Environmental and Reliability Specifications	5
Table 5.	Display Modes	8
Table 6.	SMBPBI Commands	10
Table 7.	PCIe CEM 5.0 16-Pin PCIe PSU Power Level vs. Sense Logic	13
Table 8.	Supported Auxiliary Power Connections	13
Table 9.	Power Adapter Sideband Signals	14
Table 10.	Languages Supported	16

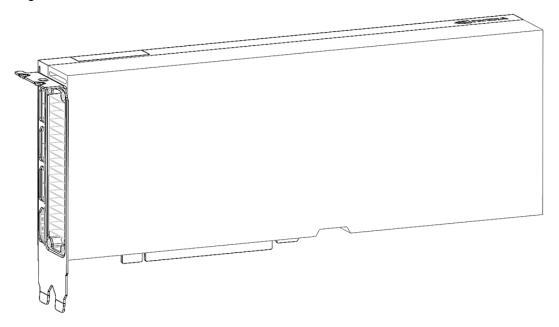
Overview

The NVIDIA L40 GPU Accelerator is a full height, full-length (FHFL), dual-slot 10.5 inch PCI Express Gen4 graphics solution based on the latest NVIDIA Ada Lovelace Architecture. The card is passively cooled and capable of 300 W maximum board power.

The NVIDIA L40 supports the latest hardware-accelerated ray tracing, revolutionary AI features, advanced shading, and powerful simulation capabilities for a wide range of graphics and compute use cases in data center and edge server deployments. This includes NVIDIA Omniverse[™], cloud gaming, batch rendering, virtual workstations, and deep learning training as well as inference workloads.

As part of the NVIDIA OVX™ server platform, L40 delivers the highest level of graphics, ray tracing, and simulation performance for NVIDIA Omniverse. With 48 GB of GDDR6 memory, even the most intense graphics applications run with the highest level of performance.





Specifications

Product Specifications

Table 1 through Table 3 provide the product, memory, and software specifications for the NVIDIA L40 PCIe card.

Product Specifications Table 1.

Specification	NVIDIA L40
Product SKU	PG133 SKU 250
	NVPN: 699-2G133-0250-xxx
Total board power	300 W default
	300 W maximum
	100 W minimum
Thermal solution	Passive
Mechanical Form Factor	Full-height, full-length (FHFL) 10.5", dual-slot
GPU SKU	AD102-895/895A
PCI Device IDs	Device ID: 0x26B5
	Vendor ID: 0x10DE
	Sub-Vendor ID: 0x10DE
	Sub-System ID: 0x169D
GPU clocks	Base: 735 MHz
	Boost: 2,490 MHz
Performance States	P0
VBIOS	EEPROM size: 8 Mbit
	UEFI: Supported
PCI Express interface	PCI Express Gen4 x16
	Lane and polarity reversal supported
Multi-Instance GPU (MIG)	Not supported
NVIDIA® NVLink®	Not supported
Zero Power	Not supported

Specification	NVIDIA L40
Connectors	One PCIe 16-pin auxiliary power connector
	Four DisplayPort connectors
Weight	Board: 1,051 grams (excluding bracket, and extenders)
	Bracket with screws: 20 grams
	Enhanced straight extender: 35 grams
	Long offset extender: 48 grams
	Straight extender: 32 grams

Memory Specifications Table 2.

Specification	Description
Memory clock	9,001 MHz
Memory type	GDDR6
Memory size	48 GB
Memory bus width	384 bits
Peak memory bandwidth	864 GB/s

Software Specifications Table 3.

Specification	Description ¹
SR-IOV support	Supported 32 VF (virtual functions)
BAR address (physical functions)	BAR0: 16 MiB
	BAR1: 64 GiB (Display Off mode; default)
	BAR1: 8 GiB (Display On, 8 GB BAR1 mode)
	BAR1: 256 MiB (Display On, 256 MB BAR1 mode)
	BAR3: 32 MiB
BAR address (virtual functions)	Display Off Mode (default):
	• BAR0: 8 MiB (32 VF × 256 KiB)
	• BAR1: 64 GiB, 64-bit (32 VF × 2 GiB)
	• BAR3: 1 GiB, 64-bit (32 VF × 32 MiB)
	Display On Modes:
	VF BAR sizes are not applicable to Display On modes
Message signaled interrupts	MSI-X: Supported
	MSI: Not supported
ARI Forwarding	Supported
Driver support	Linux: R525 or later
	Windows: R525 or later
Secure Boot	Supported (See "Root of Trust" section)

Specification	Description ¹
CEC firmware	Not Supported
NVIDIA® CUDA® support	CUDA 12.0 or later
Virtual GPU software support	Supports vGPU 15.1 or later: NVIDIA Virtual Compute Server Edition
Operating modes	Display Off mode (default)
	Display On, 8 GiB BAR1 mode
	Display On, 256 MiB BAR1 mode
PCI class code	0x03 – Display controller
PCI subclass code	0x02 – 3D controller (Display Off)
	0x00 – VGA controller (Display On)
Primary boot device capability	Supported in Display On modes
ECC support	Enabled (default); can be disabled by software
SMBus (8-bit address)	0x9E (write), 0x9F (read)
IPMI FRU EEPROM I2C address	0x50 (7-bit), 0xA0 (8-bit)
Reserved I2C addresses	0xAA, 0xAC
SMBus direct access	Supported
SMBPBI	Supported

Note:

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

- 256 KiB = 256 × 1024
- 16 MiB = 16 × 1024²
- 64 GiB = 64×1024^3

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.

Environmental and Reliability Specifications

Table 4 provides the environment conditions specifications for the NVIDIA L40 PCIe card.

Board Environmental and Reliability Specifications Table 4.

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:} 1,363,144 hours at 35°C
	Controlled environment:3: 1,859,084 hours at 35°C

Notes: Specifications in this table are applicable up to 6,000 feet.

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

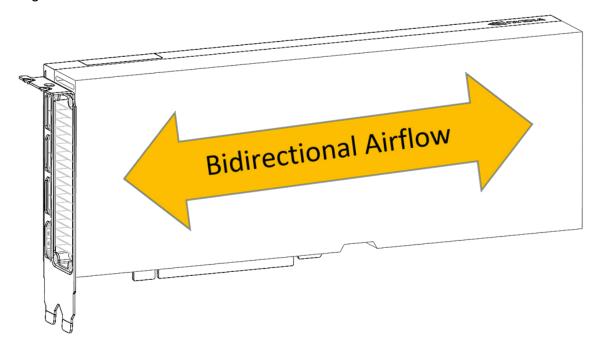
²Some environmental stress with limited maintenance (GF35).

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

Airflow Direction Support

The NVIDIA L40 PCIe card employs a bidirectional heat sink, which accepts airflow either leftto-right or right-to-left directions.

Figure 2. L40 Airflow Direction



Product Features

PCI Express Interface Specifications

The following subsections describe the PCIe interface specifications for the NVIDIA L40 PCIe card.

PCIe Support

The NVIDIA L40 GPU card supports PCIe Gen4. Either a Gen4 x16, Gen4 x8, or Gen3 x16 interface should be used when connecting to the NVIDIA L40 PCIe card.

Single Root I/O Virtualization Support

Single Root I/O Virtualization (SR-IOV) is a PCIe specification that allows a physical PCIe device to appear as multiple physical PCIe devices. Per PCIe specification, each device can have up to a maximum of 256 virtual functions (VFs). The actual number can depend on the device. SR-IOV is enabled in an NVIDIA L40 PCIe card with 32 VFs supported.

For each device, SR-IOV identifies two function classes:

- Physical functions (PFs) constitute full-featured functionality. They are fully configurable, and their configuration can control the entire device. Naturally, a PF also has full ability to move data in and out of the device.
- ▶ Virtual functions (VFs), which lack configuration resources. VFs exist on an underlying PF, which may support many such VFs. VFs only have the ability to move data in and out of the device. They cannot be configured and cannot be treated like a full PCIe device. The OS or hypervisor instance must be aware that they are not full PCIe devices.

The NVIDIA L40 requires that SBIOS and software support in the operating system (OS) instance or hypervisor is configured to enable support for SR-IOV. The OS instance or hypervisor must be able to detect and initialize PFs and VFs.

Interrupt Messaging

The NVIDIA L40 PCIe card only supports the MSI-X interrupt messaging protocol. The MSI interrupt protocol is not supported.

Polarity Inversion and Lane Reversal Support

Lane Polarity Inversion and Lane Reversal, as defined in the PCIe specification, are supported on the NVIDIA L40 PCIe card.

Root of Trust

The NVIDIA L40 provides a primary root of trust within the GPU that provides the following:

- Secure boot
- ► Secure firmware upgrade
- ► Firmware rollback protection
- ► Support for in-band firmware update disable (established after each GPU reset)
- ► Secure application processor recovery

Display

This section details the operating modes for NVIDIA L40.

Display On and Off

The L40 PCIe card supports three operating modes as described by Table 5.

Table 5. Display Modes

Display Mode	BAR Address (Physical Functions)
Display Off (default)	BAR1: 64 GiB (Display Off mode; default)
Display On 8GB BAR1 Mode	BAR1: 8 GiB (Display On, 8 GB BAR1 mode)
Display On 256MB BAR1 Mode	BAR1: 256 MiB (Display On, 256 MB BAR1 mode)

Display Off Mode

The default Display Off mode supports SR-IOV and is required to run NVIDIA Virtual GPU software. NVIDIA L40 supports all four editions of NVIDIA Virtual GPU software:

- ► NVIDIA RTX™ Virtual Workstation (RTX vWS)
- ► NVIDIA Virtual PC (vPC)
- NVIDIA Virtual Applications (vApps)
- ► NVIDIA AI Enterprise (NVAIE)

Display On 8GB BAR1 Mode

The Display On, 8GB BAR1 mode is the recommended configuration for scalable visualization system deployments. In this mode, the NVIDIA L40 card requires a BAR1 size of 8 GB and can drive up to four VESA® DisplayPort™ monitors through the integral DisplayPort connectors on the card's bracket.

Synchronizing content across multiple monitors driven from different L40 cards is accomplished by use of the NVIDIA® Quadro® Sync II card. The following document provides usage guidance for this technology:

Quadro Sync II User's Guide (DU-08348-001)

Display On 256MB BAR1 Mode

The Display On, 256MB BAR1 mode is the recommended configuration for professional desktop systems. In this mode, the NVIDIA L40 card can drive up to four DisplayPort monitors through the integral DisplayPort connectors on the card's bracket.

Synchronizing content across multiple monitors driven from different L40 cards is accomplished by use of the Quadro Sync II card. The following document provides usage quidance for this technology:

Quadro Sync II User's Guide (DU-08348-001)

Switching Operating Modes

For running NVIDIA L40 in the Display On modes, register on the NVIDIA Developer Zone and download the DisplayModeSelector tool.

System requirements should be checked before switching modes. After switching modes, the system must be rebooted for the configured modes to take effect.

Frame Lock

The NVIDIA L40 supports frame lock by use of the NVIDIA Quadro Sync II board. See the Quadro Sync II Board Specification (BD-08152-001) and Quadro Sync II User's Guide (DU-08348-001) for details. The L40 frame lock and stereo connectors are on the north edge of the NVIDIA I 40 PCIe card.

Display Bracket

The L40 card provides a display bracket that may be removed for system configurations that do not use the NVIDIA bracket.

Programmable Power

The Programmable Power feature provides the ability to configure the power cap of the card for system power and thermal budget or performance-per-watt reasons.

The power cap can be modified using either of these two NVIDIA tools:

- ▶ In-band: nvidia-smi (power cap adjustment must be reestablished after each new driver load
- Out-of-band: SMBPBI (power cap adjustment remains in force across driver loads and system boots)

Power limit specifications for the NVIDIA L40 are presented in Table 1.

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 reduce the power cap to 150 W is shown:

```
nvidia-smi -pm 1
nvidia-smi -pl 150
```

To restore the NVIDIA L40 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 300

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. The power cap can be adjusted through the following asynchronous command:

SMBPBI Commands Table 6.

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

Using SMBPBI, the configured power limit setting can be made persistent across driver reloads. Refer to the SMBus Post-Box Interface (SMBPBI) Design I(DG-06034-002) for full implementation details.

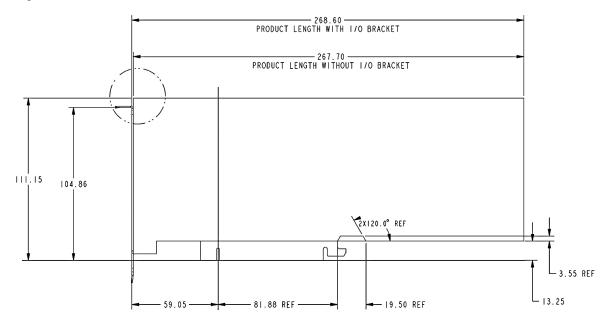
IPMI FRU FFPROM

The NVIDIA L40 supports the Intelligent Platform Management Interface (IPMI) FRU Specification v1.0 r1.3. See Table 3 for the FRU EEPROM I2C address.

Form Factor

In this product specification, nominal dimensions are shown.

Figure 3. NVIDIA L40 PCIe Card Dimensions



Power Connector

This section details the power connector for the NVIDIA L40 PCIe card.

Power Connector Placement

The board provides a PCIe 16-pin power connector on the east edge of the board.

Figure 4. PCIe 16-Pin Power Connector

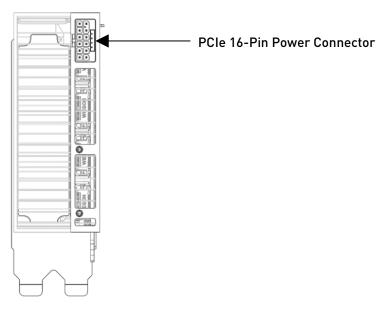
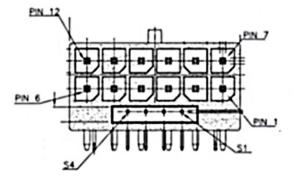


Figure 5 shows the pin assignments for the PCIe 16-pin power connector, per PCIe CEM 5.0 specification.

Figure 5. PCIe 16-Pin Power Connector Pin Assignments



12V PWR Cable Plug Pin Assignment

Pin	Signal	
1 to 6	+12 V	
7 to 12	Ground	
S1	CARD_PWR_STABLE	
S2	CARD_CBL_PRES#	
S3	SENSE0	
S4	SENSE1	

Table 7 lists the power level options identifiable by the PCIe 16-pin power connecter per CEM5 PSU, and the corresponding Sense0 and Sense1 logic. The NVIDIA card senses the Sense0 and Sense1 levels and recognizes the power available to the NVIDIA card from the power connector. If the power level identified by Sense0 and Sense1 is equal to or greater than what the NVIDIA card needs from the 16-pin connector, the NVIDIA card operates per normal. If the power level identified by Sense0 and Sense1 is less than the default power cap of the NVIDIA card, the card will not boot.

The NVIDIA L40 requires up to 300 W from the 16-pin power connector. Thus, the top three rows of Table 7 are acceptable. The configuration [Sense0=1, Sense1=1] will prevent the NVIDIA L40 from booting.

PCIe CEM 5.0 16-Pin PCIe PSU Power Level vs. Sense Logic Table 7.

Power Level	Sideband 3 (Sense0)	Sideband 4 (Sense1)
451 - 600 W	0	0
301 - 450 W	1 (float)	0
151 - 300 W	0	1 (float)
Up to 150 W	1 (float)	1 (float)

Table 8 lists supported auxiliary power connections for the NVIDIA L40 GPU card.

Table 8. Supported Auxiliary Power Connections

Board Connector		PSU Cable	
	PCIe 16-pin	PCle 16-pin	
	PCIe 16-pin	CPU 8-pin to PCIe 16-pin	

CPU 8-Pin to PCIe 16-Pin Power Adapter

A CPU 8-pin to PCle 16-pin power adapter for systems that do not have native PCle 16-pin power connectors may be used. Figure 6 illustrates an example power adapter.

Figure 6. CPU 8-Pin to PCle 16-Pin Power Adapter

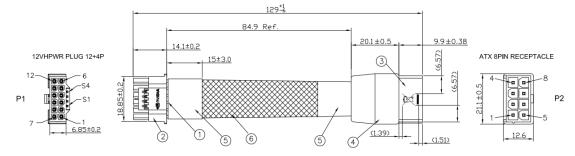
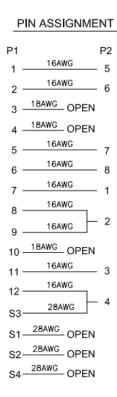


Figure 7 shows CPU 8-Pin to PCle 16-Pin power adapter pin assignments.

Figure 7. CPU 8-Pin to PCle 16-Pin Power Adapter Pin Assignments



The power adapter must support the four Sideband signals, hardware-strapped per Table 9. Strapping corresponding to the "151 – 300 W" power level PCIe CEM 5.0 specification (shown in Row 3 of Table 7), or strapping rows 1 and 2 of Table 7 are acceptable for the NVIDIA L40 PCIe card. When using a CPU 8-pin to PCIe 16-pin power adapter, the NVIDIA L40 recognizes that there is sufficient power available on its PCIe 16-pin auxiliary power connector.

Power Adapter Sideband Signals Table 9.

Sideband 1	Sideband 2	Sideband 3	Sideband 4
Card Power Stable	Card Cable Present	Sense0	Sense1
1 (float)	1 (float)	0	

Extenders

The NVIDIA L40 PCIe extender options are TBD.

Support Information

Certification

- ▶ Windows Hardware Quality Lab (WHQL):
 - Certified Windows 7, Windows 8.1, Windows 10, Windows 11
 - Certified Windows Server 2012 R2, Windows Server 2019, Windows Server 2022
- ► 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 10. 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:

¹Microsoft 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, NVIDIA Omniverse, NVIDIA OVX, NVIDIA RTX, NVLink, and Quadro 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.

VESA DisplayPort

DisplayPort and DisplayPort Compliance Logo, DisplayPort Compliance Logo for Dual-mode Sources, and DisplayPort Compliance Logo for Active Cables are trademarks owned by the Video Electronics Standards Association in the United States and other countries.

Copyright

© 2022, 2023 NVIDIA Corporation. All rights reserved.

