NVIDIA A40 GPU Accelerator

Product Brief

PB-09976-001_v04 | January 2021

Document History

PB-09976-001_v04

Version	Date	Authors	Description of Change
01	May 29, 2020	MD, VK, MV, SM	Preliminary Information
02	October 16, 2020	VK, SM	• Updated software features • Updated Max-Q section • Updated product name • General edits throughout
03	December 18, 2020	VK, SM	• Removed "Preliminary" from product brief • Updated software specifications in Table 3 • Added "Display" section • Updated diagram in Figure 3
04	January 5, 2021	MD, SM	Removed confidential markings from product brief

Table of Contents

List of Figures

List of Tables

Overview

The NVIDIA A40 is a full height, full-length (FHFL), dual-slot 10.5-inch PCI Express Gen4 graphics solution based on the state-of-the-art NVIDIA Ampere architecture. The card is passively cooled and capable of 300 W maximum board power.

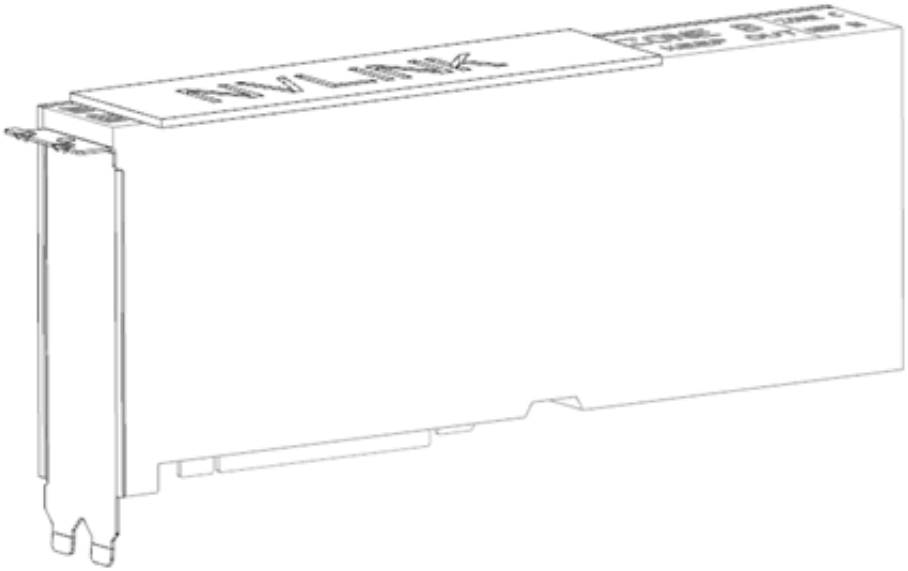
The NVIDIA A40 supports the latest hardware-accelerated ray tracing, revolutionary AI features, advanced shading, and powerful simulation capabilities for a wide range of graphics and compute use cases in data center and edge server deployments, including batch rendering, virtual workstations, and deep learning training as well as inference workloads.

With 48 GB of GDDR6 memory (expandable to 96 GB with NVIDIA® NVLink®), even the most intense graphics or deep learning applications run with the highest level of performance, including those with very large data sets.

The NVIDIA A40 card enables use cases requiring either virtual or physical displays. By default, the GPU is configured to support virtual graphics and compute workloads (display off mode). A utility can be requested to enable support for specific use cases that require the physical display ports (display on modes), such as location-based entertainment and virtual studio production.

The card is designed to meet the requirements of NEBS Level 3 compliant servers and supports security features like secure boot with hardware root of trust.

Figure 1. NVIDIA A40 PCIe Card



Specifications

Product Specifications

Table 1 through Table 3 provides the product, memory, and software specifications for the NVIDIA A40 PCIe card.

Table 1. Product Specifications

Specification	NVIDIA A40
Product SKU	PG133 SKU 200 NVPN: 699-2G133-0200-xxx
Total board power	300 W (default)
Thermal solution	Passive
Mechanical form factor	Full-height, full-length (FHFL) 10.5-inch, dual-slot
GPU SKU	GA102-895
PCI Device IDs	Device ID: 0x2235 Vendor ID: 0x10DE Sub-Vendor ID: 0x10DE Sub-System ID: 0x145A
GPU clocks	Base: 1305 MHz Boost: 1740 MHz
VBIOS	EEPROM size: 8 Mbit UEFI: Supported
PCI Express interface	PCI Express 4.0 ×16 Lane and polarity reversal supported
Secure Boot	Supported
Zero Power	Not supported
NEBS readiness	Supported
Connectors and headers	One CPU 8-pin auxiliary power connector Three DisplayPort connectors
Weight	Board: 990 Grams (excluding bracket and extenders)

Specification	NVIDIA A40
	Bracket with screws: 20 Grams Long offset extender: 64 Grams Straight extender: 39 Grams

Note:

[1]The allowable power range for Max-Q is 100 W to 300 W. Max-Q power and thermal levels must be qualified by the NVIDIA partner.

Table 2. Memory Specifications

Specification	Description
Memory clock	7250 MHz
Memory type	GDDR6
Memory size	48 GB
Memory bus width	384 bits
Peak memory bandwidth	Up to 696 GB/s

Table 3. Software Specifications

Specification	Description
SR-IOV support	Supported -- 32 VF (virtual functions)
BAR address (physical function)	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 function)	Display Off Mode (default): • BAR0: 8 MiB (32 VF x 256 KiB) • BAR1: 64 GiB, 64-bit (32 VF x 2 GiB) • BAR3: 1 GiB, 64-bit (32 VF x 32 MiB) Display On Modes: • VF BAR sizes are not applicable to Display On modes
Message signaled interrupts	MSI-X: Supported MSI: Not supported
Multi-Instance GPU (MIG)	Not supported
ARI Forwarding	Supported
Driver Support	R460.16 or later
CEC Firmware	v5.01 or later

Specification	Description
NVIDIA® CUDA® Support	CUDA 11.2 or later
Virtual GPU Software Support	Supports vGPU 12.0 or later
NVIDIA® NGC-Ready™ Test Suite	NGC-Next Certification 2.x or later
Operating modes	Display Off mode (default) Display On, 8 GB BAR1 mode Display On, 256 MB BAR1 mode
PCI class code	0x03 - Display Controller
PCI sub-class code	0x02 – 3D Controller
Primary Boot Device Capability	Not supported in either operating mode
ECC Support	Enabled (by default); can be disabled via software
SMBus (8-bit address)	0x9E (write), 0x9F (read)
SMBus direct access	Supported
SMBus Post Box Interface (SMBPBI)	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.

Environmental and Reliability Specifications

Table 4 provides the environment conditions specifications for the NVIDIA A40 card.

Table 4. Board Environmental and Reliability Specifications

Specification	Description
Ambient operating temperature	0 °C to 55 °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]: TBD hours at 35 °C Controlled environment[2]: TBD hours at 35 °C

Notes:

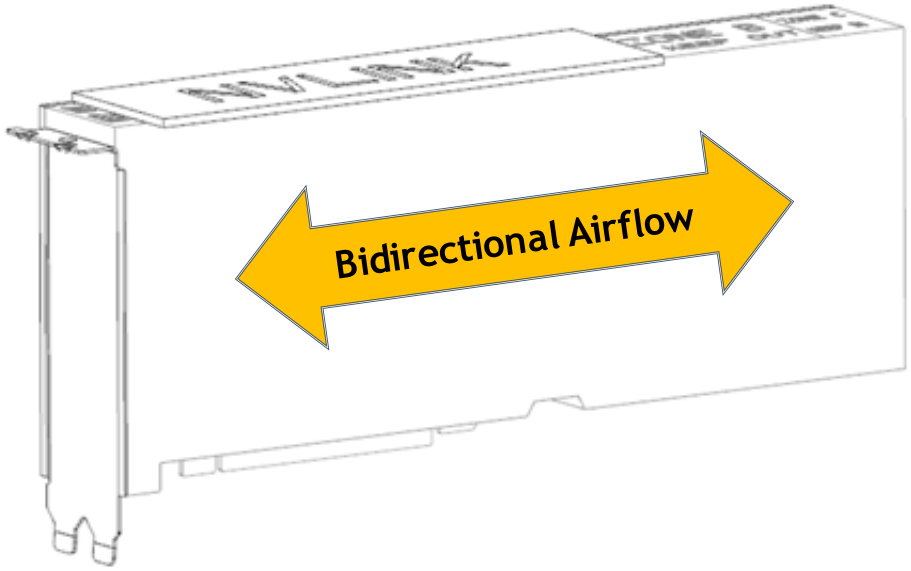
[1]Some environmental stress with limited maintenance (GF35).

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

Airflow Direction Support

The NVIDIA A40 PCIe card employs a bidirectional heat sink, which accepts airflow either left-to-right or right-to-left directions.

Figure 2. NVIDIA A40 Airflow Directions



Product Features

PCI Express Interface Specifications

The following sub-sections describe the PCIe interface specifications for the NVIDIA A40 PCIe card.

PCIe Speed Support

The A40 card supports PCIe Gen4 and PCIe Gen link training. For optimal graphics processing unit (GPU) performance, a Gen4 ×16 connection is recommended, but a Gen4 ×8 or Gen3 ×16 link connection is supported as well. Use of a Gen3 ×8 link connection is not recommended.

Polarity Inversion and Lane Reversal Support

Lane Polarity Inversion, as defined in the PCIe specification, is supported on the A40 PCIe card.

CEC Hardware Root of Trust

The NVIDIA A40 provides secure boot capability via CEC. Implementing code authentication, rollback protection and key revocation, the CEC device authenticates the contents of the GPU firmware ROM before permitting the GPU to boot from its ROM.

It also provides out-of-band (OOB) secure firmware update, secure application processor recovery, and remote attestation.

Display

This section details the operating modes for NVIDIA A40.

Display On/Off

The A40 PCIe card supports three operating modes:

- Display Off Mode (default)
- Display On, 8 GB BAR1 Mode
- Display On, 256 MB BAR1 Mode

Display Off Mode

The default mode on NVIDIA A40 is Display Off Mode. This supports SR-IOV and is required to run NVIDIA Virtual GPU software. NVIDIA A40 with NVIDIA RTX™ Virtual Workstation (vWS) software enables the user to tackle massive datasets, large 3D models, and complex designs with scaled memory and performance. NVIDIA A40 supports all four editions of NVIDIA virtual GPU software: NVIDIA vWS, NVIDIA Virtual Applications (vApps), NVIDIA Virtual PC (vPC), and NVIDIA Virtual Compute Server (vCS).

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 A40 card requires a BAR1 size of 8GB and can drive up to four VESA® DisplayPort™ monitors via the integral DisplayPort (DP) connectors on the card's bracket.

Synchronizing content across multiple monitors driven from different A40 cards is accomplished by use of the NVIDIA® Quadro® Sync II card.

Display On 256MB BAR1 Mode

The Display On, 256 MB BAR1 mode is the recommended configuration for professional desktop systems. In this mode, the NVIDIA A40 card can drive up to four DisplayPort monitors via the integral DisplayPort (DP) connectors on the card's bracket.

Synchronizing content across multiple monitors driven from different A40 cards is accomplished by use of the NVIDIA Quadro Sync II card.

Switching Operating Modes

For running NVIDIA A40 for Broadcast and Virtual production with Display ON, register on NVIDIA Developer Zone. System requirements should be checked prior to switching modes.

After switching modes, the system must be rebooted, after which the configured mode takes effect.

Frame Lock

The NVIDIA A40 supports frame lock by use of the NVIDIA Quadro Sync II board. The A40 Frame Lock and Stereo connectors are on the north edge near the NVLink Bridge interface.

Max-Q Operation

Configuring for Max-Q operation optimizes for GPU performance per watt. Max-Q operation is applicable to either operating mode of the A40 card and 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 300 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 with a power limit of 140 W is shown:

```
nvidia-smi -pm 1
nvidia-smi -pl 140
```

To restore the A40 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, 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 5. SMBPBI Commands

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

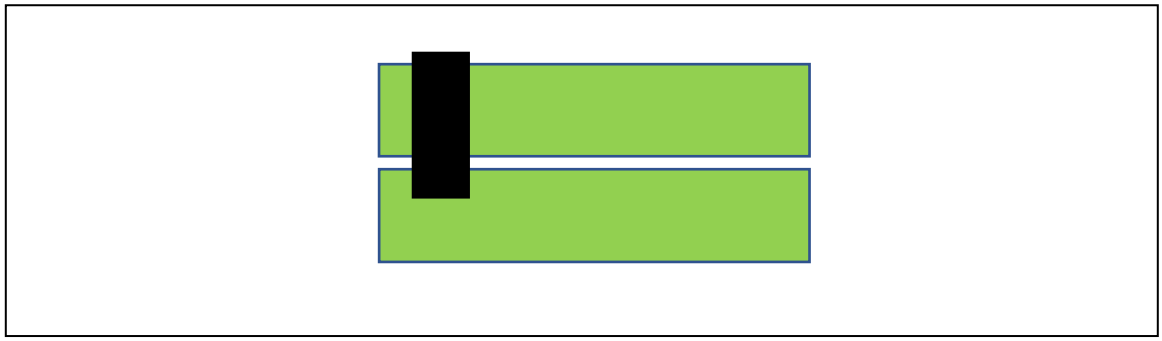
The operator is given the option to configure this power setting to be persistent across driver reloads or to revert to default power settings upon driver unload.

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 A40 card supports NVLink bridge connection with a single adjacent A40 card.

The single attached bridge spans two PCIe slots. Figure 3 illustrates A40 NVLink connection.

Figure 3. A40 NVLink Connection – Top View



For systems that feature multiple CPUs, both A40 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. There are exceptions, for example in a system with dual CPUs wherein each CPU has a single A40 PCIe card under it; in that case, the two A40 PCIe cards in the system may be bridged together.

A40 NVLink speed and bandwidth are given in the following table.

NVLink Connector Placement

Figure 4 shows the connector keep-out area for the NVLink bridge support of the A40 PCIe card.

Figure 4. NVLink Connector Keep Out Area – Top View

The A40 PCIe card supports the 2-slot span NVLink bridge. NVIDIA A40 NVLink speed and bandwidth are given in the following table.

Table 6. NVLink Speed and Bandwidth

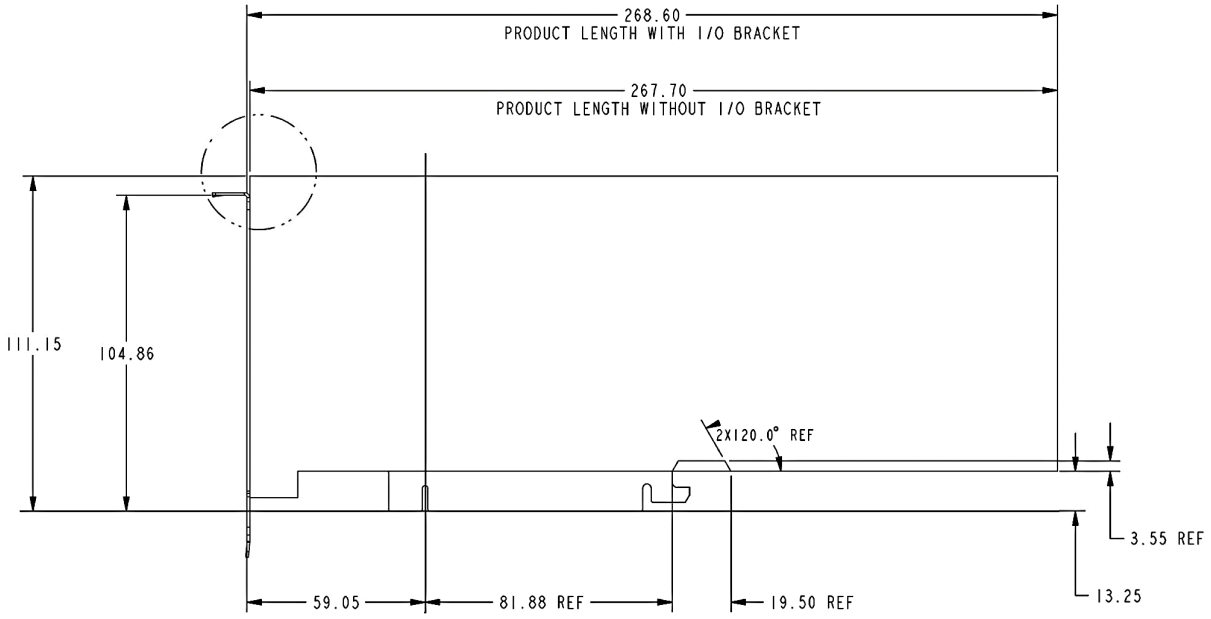
Parameter	Value
Total NVLink bridges supported by NVIDIA A40	1
NVLink links per bridge	4
Lanes per link	4
Data rate per NVIDIA A40 NVLink lane (each direction)	28.125 Gbps
Total maximum NVLink bandwidth (bi-directional)	112.5 GB/s

Sufficient clearance must be provided both above the north edge of the card and behind the backside of the card's PCB to accommodate an NVIDIA A40 NVLink bridge. The clearance above the card's north edge should meet or exceed 2.5 mm. The backside clearance (from the PCB's rear surface) should meet or exceed 2.67 mm.

Form Factor

In this product brief, nominal dimensions are shown.

Figure 5. NVIDIA A40 PCIe Card Dimensions



Power Connector Placement

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

Figure 6. CPU 8-Pin Power Connector

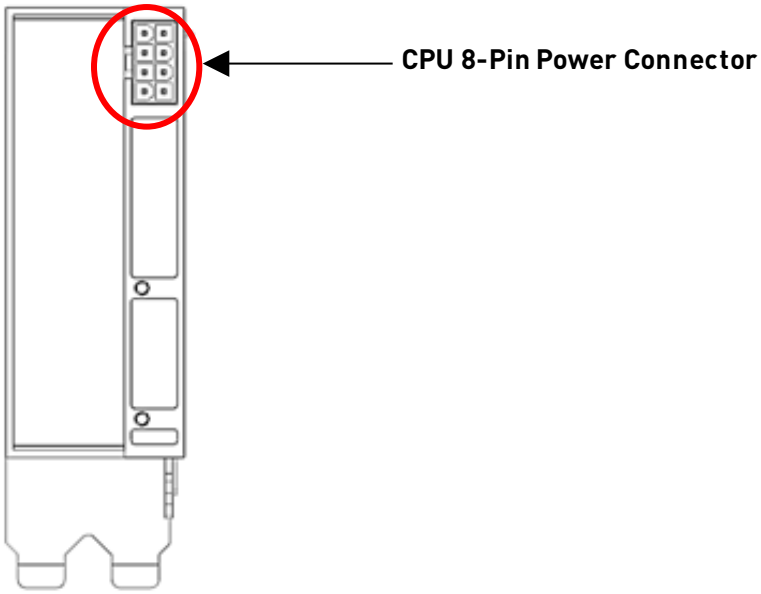


Table 8 lists supported auxiliary power connections for the NVIDIA A40 PCIe card.

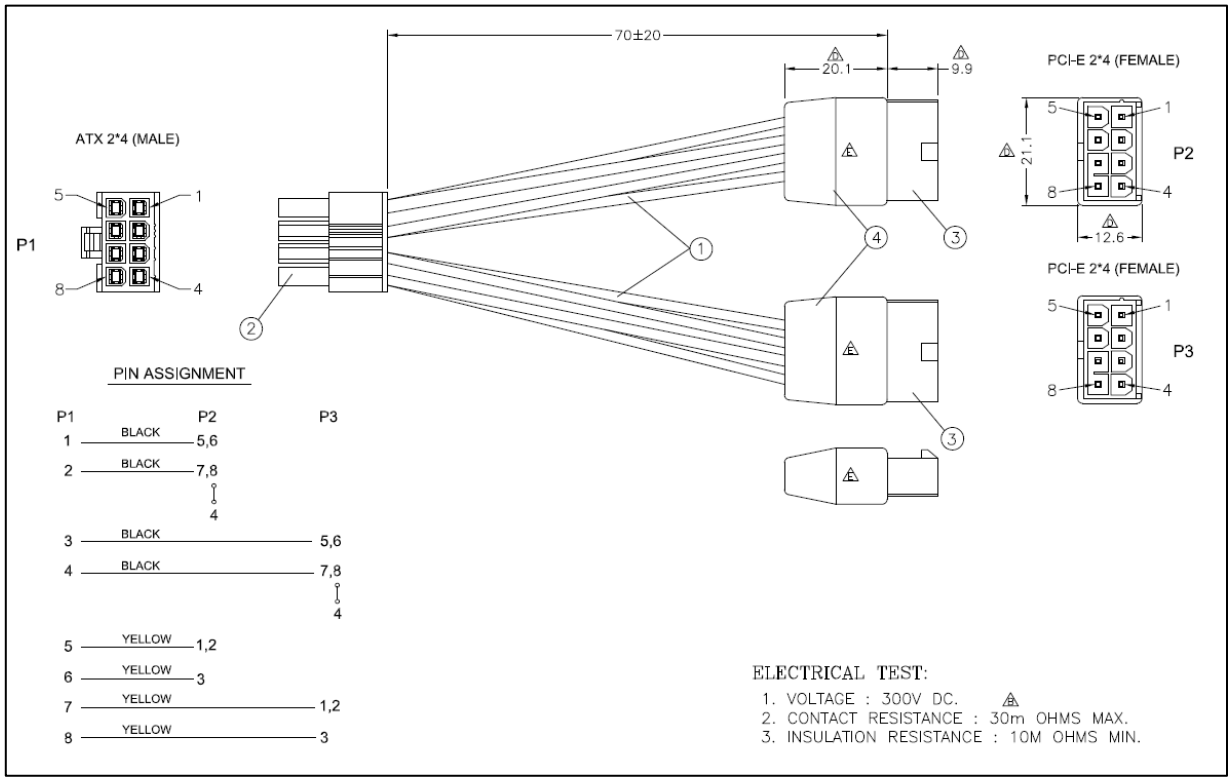
Table 7. Supported Auxiliary Power Connections

Board Connector	PSU Cable
CPU 8-pin	1x CPU 8-pin cable
CPU 8-pin	2x PCIe 8-pin cable

CPU 8-Pin to PCIe 8-Pin Power Adapter

Figure 8 lists the pin assignments of the power adapter. Consult NVIDIA Applications Engineering for qualified suppliers of the power adapter.

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



Extenders

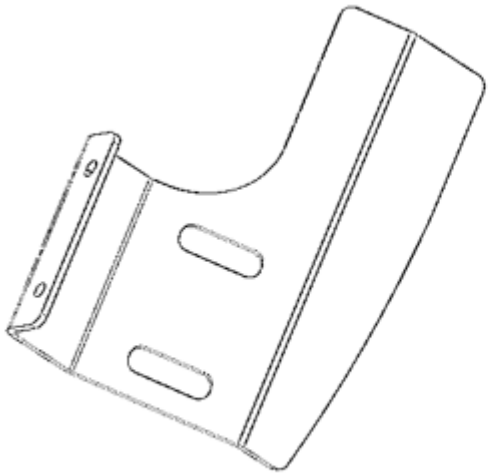
The NVIDIA A40 PCIe card provides two extender options shown in Figure 9.

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

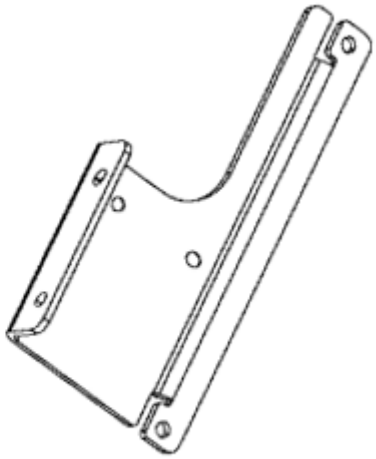
Using the standard NVIDIA extender ensures greatest forward compatibility with future NVIDIA product offerings.

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

Support Information

Certifications

- Windows Hardware Quality Lab (WHQL):
  - Certified Windows Server 2016, Windows Server 2019
  - 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 9 lists the languages supported for the NVIDIA A40 card.

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 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.

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.

Trademarks

NVIDIA, the NVIDIA logo, CUDA, NGC-Ready, NVIDIA GRID, 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.

Copyright

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