



NVIDIA A100 80GB PCIe GPU

Product Brief

Document History

PB-10577-001_v03

Version	Date	Authors	Description of Change
01	June 28, 2021	AS, SM	Initial Release
02	June 28, 2021	AS, SM	Removed confidential and preliminary markings
03	March 17, 2022	VK, SM	<ul style="list-style-type: none">• Removed Secure Boot from Table 1• Updated Secure Boot and CEC description in Table 3• Updated short term ambient operating temperature in Table 4• Updated “Root of Trust” section• Added Table 6 “Root of Trust Feature Set”

Table of Contents

- Overview 1
- Specifications 3
 - Product Specifications 3
 - Environmental and Reliability Specifications 5
- Airflow Direction Support 6
- Product Features 7
 - Form Factor 7
 - NVLink Bridge Support 8
 - NVLink Bridge 8
 - NVLink Connector Placement 9
 - Multi-Instance GPU Support 9
 - Programmable Power 9
 - Root of Trust 10
 - Power Connector 11
 - Power Connector Placement 11
 - CPU 8-Pin to PCIe 8-Pin Power Adapter 12
 - Extenders 12
- Support Information 14
 - Certifications 14
 - Agencies 14
 - Languages 15

List of Figures

Figure 1.	NVIDIA A100 80GB with NVLink Bridge Volumetric.....	2
Figure 2.	A100 80GB Airflow Directions	6
Figure 3.	NVIDIA A100 80GB PCIe Card Dimensions	7
Figure 4.	NVLink Topology – Top Views	8
Figure 5.	NVLink Connector Placement – Top View.....	9
Figure 6.	CPU 8-Pin Power Connector	11
Figure 7.	CPU 8-Pin to PCIe 8-Pin Power Adapter	12
Figure 8.	Long Offset and Straight Extenders	13

List of Tables

Table 1.	Product Specifications	3
Table 2.	Memory Specifications	4
Table 3.	Software Specifications.....	4
Table 4.	Board Environmental and Reliability Specifications	5
Table 5.	A100 80GB PCIe Card NVLink Speed and Bandwidth	8
Table 6.	Root of Trust Feature Set.....	10
Table 7.	Supported Auxiliary Power Connections.....	11
Table 8.	Languages Supported	15

Overview

The NVIDIA® A100 80GB PCIe card delivers unprecedented acceleration to power the world's highest-performing elastic data centers for AI, data analytics, and high-performance computing (HPC) applications. NVIDIA A100 Tensor Core technology supports a broad range of math precisions, providing a single accelerator for every compute workload. The NVIDIA A100 80GB PCIe supports double precision (FP64), single precision (FP32), half precision (FP16), and integer (INT8) compute tasks.

The NVIDIA A100 80GB card 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 airflow to properly operate the card within its thermal limits. The NVIDIA A100 80GB PCIe operates unconstrained up to its maximum thermal design power (TDP) level of 300 W to accelerate applications that require the fastest computational speed and highest data throughput. The latest generation A100 80GB PCIe doubles GPU memory and debuts the world's highest PCIe card memory bandwidth up to 1.94 terabytes per second (TB/s), speeding time to solution for the largest models and most massive data sets.

The NVIDIA A100 80GB PCIe card features Multi-Instance GPU (MIG) capability, which can be partitioned into as many as seven isolated GPU instances, providing a unified platform that enables elastic data centers to dynamically adjust to shifting workload demands. When using MIG to partition an A100 GPU into up to seven smaller instances, A100 can readily handle different-sized acceleration needs, from the smallest job to the biggest multi-node workload. A100 80GB versatility means IT managers can maximize the utility of every GPU in their data center.

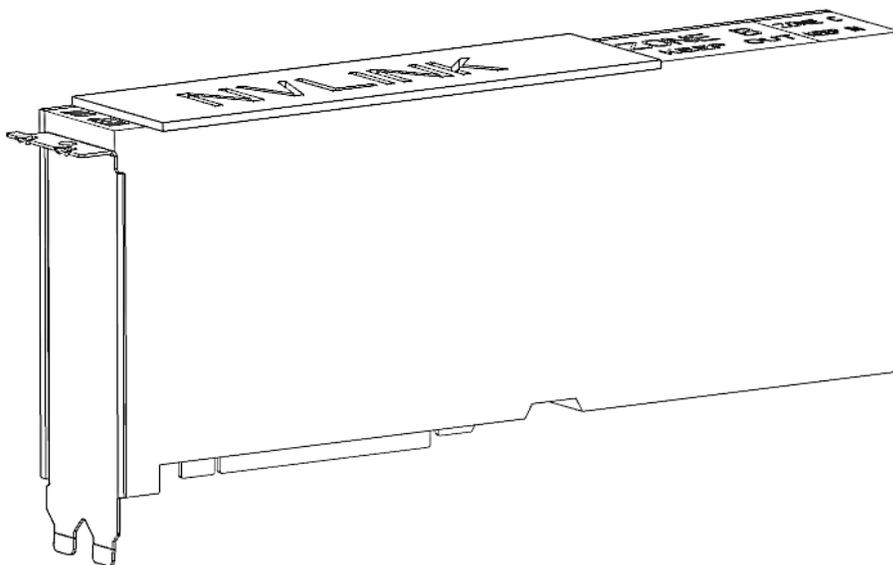
NVIDIA A100 80GB PCIe cards use three NVIDIA® NVLink® bridges that allow multiple A100 80GB PCIe cards to be connected together to deliver 600 GB/s bandwidth or 10x the bandwidth of PCIe Gen4, in order to maximize application throughput with the larger workloads.

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>

Refer to the following website for the latest list of qualified A100 80GB servers:

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

Figure 1. NVIDIA A100 80GB with NVLink Bridge Volumetric



Specifications

Product Specifications

Table 1 through Table 3 provide the product, memory, and software specifications for the NVIDIA A100 80GB PCIe card.

Table 1. Product Specifications

Specification	NVIDIA A100 80GB
Product SKU	P1001 SKU 230 NVPN: 699-21001-0230-xxx
Total board power	300 W default 300 W maximum 150 W minimum
Thermal solution	Passive
Mechanical Form Factor	Full-height, full-length (FHFL) 10.5", dual-slot
GPU SKU	GA100-893FF, GA100-893FFF, GA100-893HH, GA100-893HHH
PCI Device IDs	Device ID: 0x20B5 Vendor ID: 0x10DE Sub-Vendor ID: 0x10DE Sub-System ID: 0x1533
GPU clocks	Base: 1065 MHz Boost: 1410 MHz
Performance States	P0
VBIOS	EEPROM size: 8 Mbit UEFI: Supported
PCI Express interface	PCI Express 4.0 ×16 Lane and polarity reversal supported
Multi-Instance GPU (MiG)	Supported (seven instances)
Power connectors and headers	One CPU 8-pin auxiliary power connector

Specification	NVIDIA A100 80GB
Weight	Board: 1170 grams (excluding bracket, extenders, and bridges) NVLINK Bridge: 20.5 grams per bridge (x 3 bridges) Bracket with screws: 20 grams Long offset extender: 48 grams Straight extender: 32 grams

Table 2. Memory Specifications

Specification	Description
Memory clock	1512 MHz
Memory type	HBM2e
Memory size	80 GB
Memory bus width	5120 bits
Peak memory bandwidth	Up to 1.94 TB/s

Table 3. Software Specifications

Specification	Description ¹
SR-IOV support	Supported -- 20 VF (virtual functions)
BAR address (physical function)	BAR0: 16 MiB ¹ BAR1: 128 GiB ¹ BAR3: 32 MiB ¹
BAR address (virtual function)	BAR0: 5 MiB, (256 KiB per VF) ¹ BAR1: 80 GiB, 64-bit (4 GiB per VF) ¹ BAR3: 640 MiB, 64-bit (32 MiB per VF) ¹
Message signaled interrupts	MSI-X: Supported MSI: Not supported
ARI Forwarding	Supported
Driver support	Linux: R470.12 or later Windows: R470.37 or later
Secure Boot	Supported (See "Root of Trust" section)
CEC firmware	Version 6.01 or later (for CEC-enabled cards)
NVFlash	Version 5.695 or later
NVIDIA® CUDA® support	CUDA 11.4 or later
Virtual GPU Software Support	Supports vGPU 13/R470 or later: NVIDIA Virtual Compute Server Edition
NVIDIA AI Enterprise	Supported with VMWare

Specification	Description ¹
NVIDIA Certification	NVIDIA-Certified Systems™ 2.3 or later
PCI class code	0x03 – Display Controller
PCI sub-class code	0x02 – 3D Controller
ECC support	Enabled
SMBus (8-bit address)	0x9E (write), 0x9F (read)
Reserved I2C addresses ²	0xAA, 0xAC
SMBus direct access	Supported
SMBPBI SMBus Post-Box Interface)	Supported

Note:

¹The KiB, MiB and GiB notation emphasizes the “power of two” nature of the values. Thus,

- 256 KiB = 256 x 1024
- 16 MiB = 16 x 1024²
- 64 GiB = 64 x 1024³

²See “Root of Trust” section in this product specification.

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 A100 80GB 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: ² 945,568 hours at 35 °C Controlled environment: ³ 1,303,691 hours at 35 °C

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

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

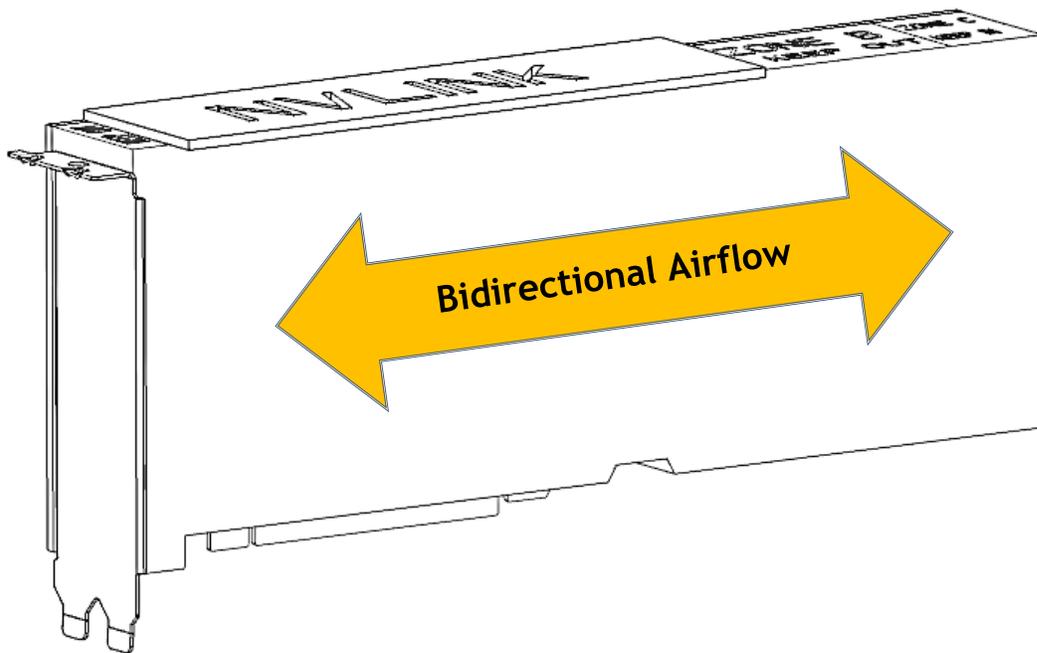
²Some environmental stress with limited maintenance (GF35).

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

Airflow Direction Support

The NVIDIA A100 80GB PCIe card employs a bidirectional heat sink, which accepts airflow either left-to-right or right-to-left directions.

Figure 2. A100 80GB Airflow Directions



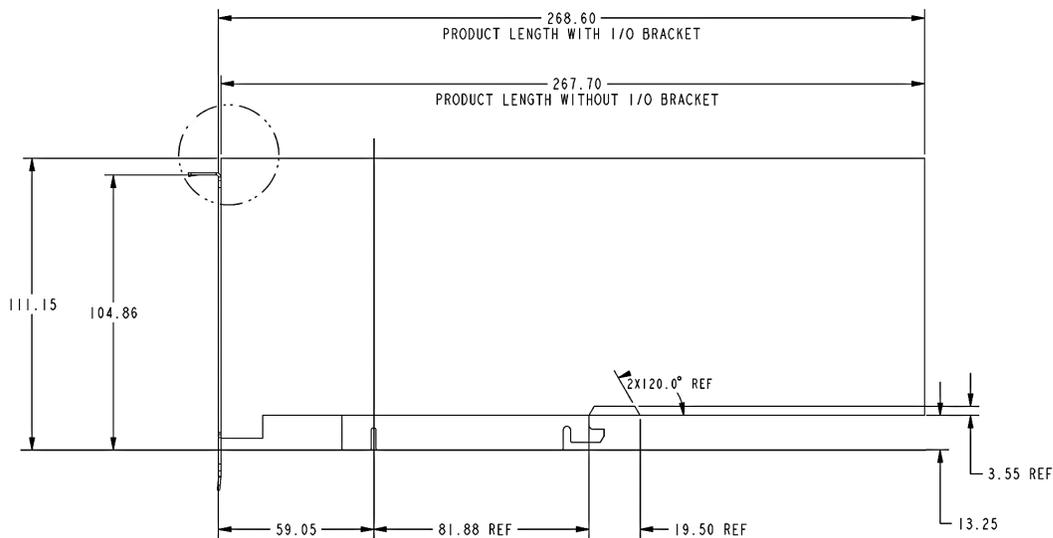
Product Features

Form Factor

The NVIDIA A100 80GB 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 for Server PCIe Products Specification* (NVOnline reference number 1052306).

In this product specification, 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 80GB PCIe Card Dimensions

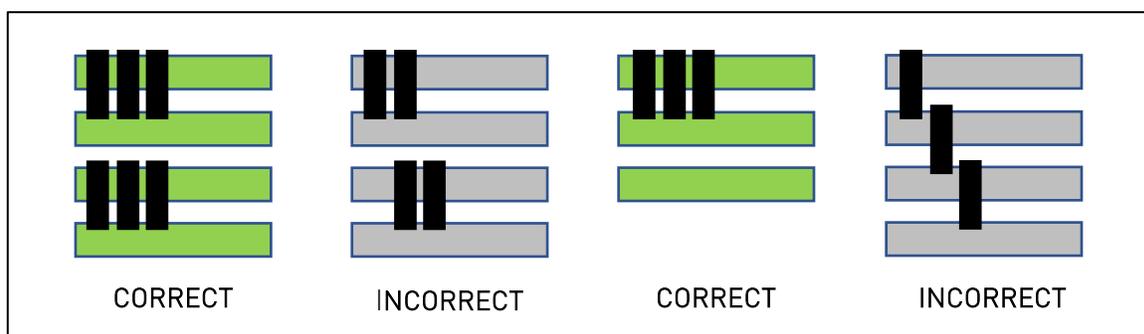


NVLink Bridge Support

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 80GB card supports NVLink bridge connection with a single adjacent A100 80GB 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 80GB 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 80GB 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 A100 80GB PCIe card under it; in that case, the two A100 80GB PCIe cards in the system may be bridged together.

A100 80GB PCIe card NVLink speed and bandwidth are given in the following table.

Table 5. A100 80GB PCIe Card NVLink Speed and Bandwidth

Parameter	Value
Total NVLink bridges supported by NVIDIA A100 80GB	3
Total NVLINK Rx and Tx lanes supported	96
Data rate per NVIDIA A100 80GB NVLink lane (each direction)	50 Gbps
Total maximum NVLink bandwidth	600 Gbytes per second

NVLink Bridge

The 2-slot NVLink Bridge for the NVIDIA A100 80GB PCIe card has the following NVIDIA part number: 900-53651-0000-000

NVLink Connector Placement

Figure 5 shows the connector keep-out area for the NVLink bridge support of the A100 80GB.

Figure 5. NVLink Connector Placement – Top View



Sufficient clearance must be provided both above the card's north edge and behind the backside of the card's PCB to accommodate NVIDIA A100 80GB NVLink bridges. The clearance above the 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 details.

NVLink bridge interfaces of the A100 80GB PCIe card include removable caps to protect the interfaces in non-bridged system configurations.

Multi-Instance GPU Support

The A100 80GB PCIe card supports Multi-Instance GPU (MIG) capability by providing up to seven 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>

Programmable Power

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

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

- ▶ nvidia-smi (power cap adjustment must be re-established after each new driver load)
- ▶ SMBPBI (power cap adjustment remains in force across driver loads and system boots)

Root of Trust

The NVIDIA A100 80GB 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 A100 80GB 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 A100 80GB 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.

- ▶ 900-21001-XXXX-1XX A100 80GB GPUs without CEC1712 (secondary root of trust)
- ▶ 900-21001-XXXX-0XX A100 80GB 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 6. 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.		

Power Connector

This section details the power connector for the NVIDIA A100 80GB PCIe card.

Power Connector Placement

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

Figure 6. CPU 8-Pin Power Connector

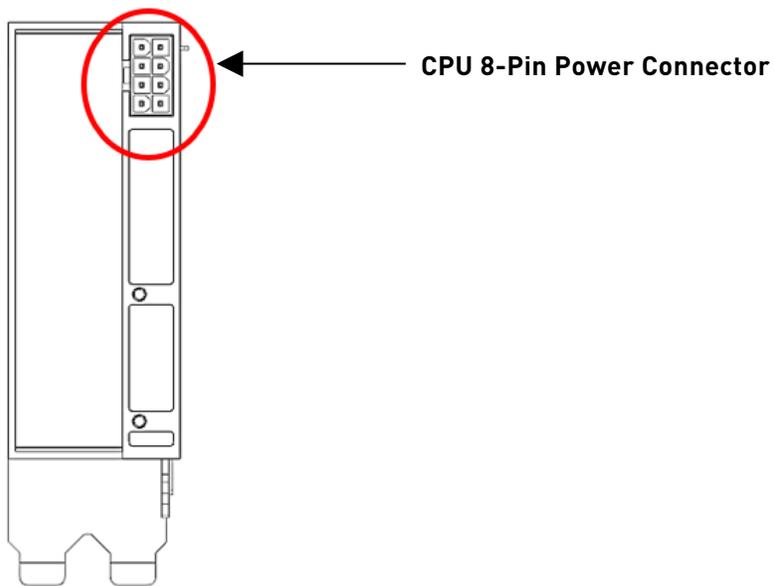


Table 7 lists supported auxiliary power connections for the NVIDIA A100 80GB GPU card.

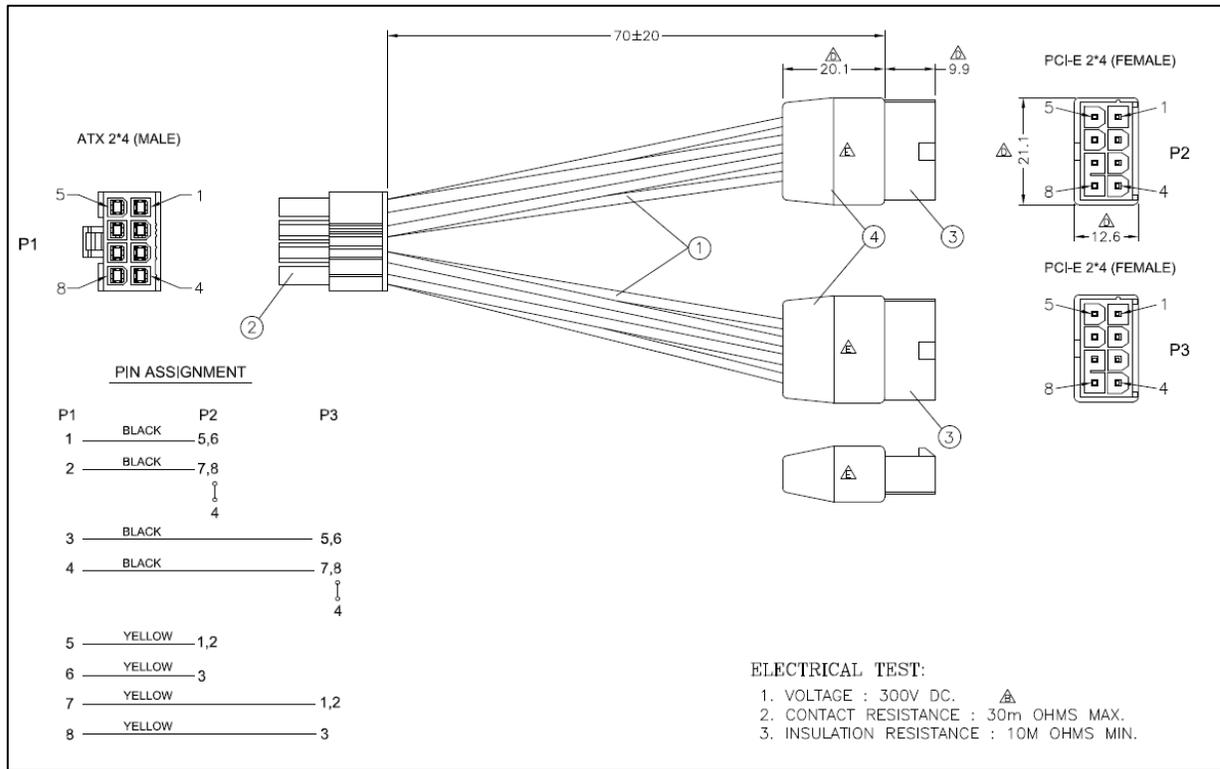
Table 7. Supported Auxiliary Power Connections

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

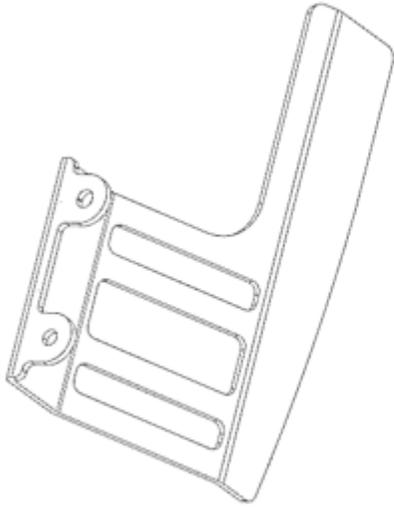
The NVIDIA A100 80GB PCIe card provides two extender options, shown in Figure 8.

- ▶ NVPN: 682-00007-5555-000 – Long offset extender
 - Card + extender = 339 mm
- ▶ NVPN: 682-00007-5555-001 – Straight extender
 - Card + extender = 312 mm

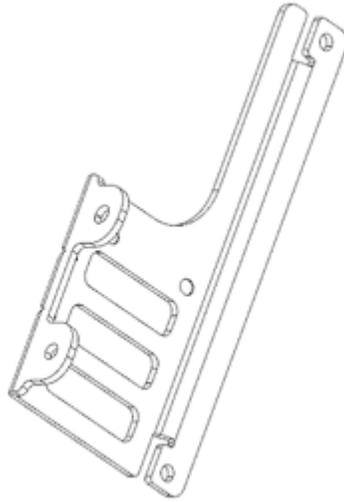
Using the standard NVIDIA extender ensures greatest forward compatibility with future NVIDIA product offerings.

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. Long Offset and Straight Extenders



LONG OFFSET EXTENDER



STRAIGHT EXTENDER

Support Information

Certifications

- ▶ Windows Hardware Quality Lab (WHQL):
 - Certified Windows 7, Windows 8.1, 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:

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-Certified Systems, and NVLink 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

© 2021, 2022 NVIDIA Corporation. All rights reserved.