NVIDIA H100 PCIe GPU

Product Brief

PB-11133-001_v02 | November 2022

Document History

PB-11133-001_v02

Version	Date	Authors	Description of Change
01	September 30, 2022	FL, SM	Initial release
02	November 30, 2022	SM	Document template modification

Table of Contents

List of Figures

List of Tables

Overview

The NVIDIA® H100 Tensor Core GPU delivers unprecedented acceleration to power the world's highest-performing elastic data centers for AI, data analytics, and high-performance computing (HPC) applications. NVIDIA H100 Tensor Core technology supports a broad range of math precisions, providing a single accelerator for every compute workload. The NVIDIA H100 PCIe supports double precision (FP64), single-precision (FP32), half precision (FP16), and integer (INT8) compute tasks.

NVIDIA H100 Tensor Core graphics processing units (GPUs) for mainstream servers comes with an NVIDIA AI Enterprise five-year software subscription and includes enterprise support, simplifying AI adoption with the highest performance. This ensures organizations have access to the AI frameworks and tools needed to build H100 accelerated AI workflows such as conversational AI, recommendation engines, vision AI, and more.

Activate NVIDIA AI Enterprise license for H100 at: https://www.nvidia.com/activate-h100/

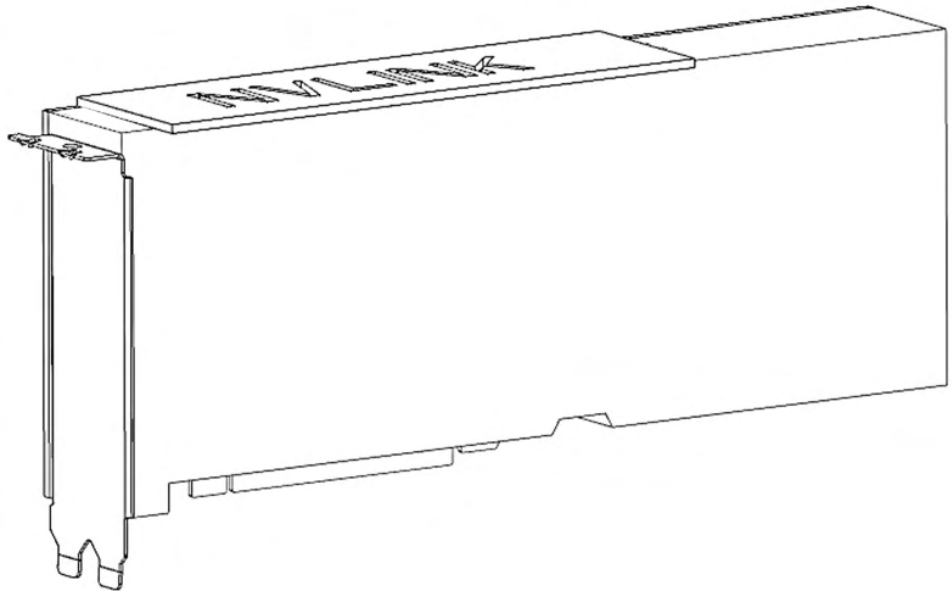
The NVIDIA H100 card is a dual-slot 10.5 inch PCI Express Gen5 card based on the NVIDIA Hopper™ architecture. It uses a passive heat sink for cooling, which requires system airflow to operate the card properly within its thermal limits. The NVIDIA H100 PCIe operates unconstrained up to its maximum thermal design power (TDP) level of 350 W to accelerate applications that require the fastest computational speed and highest data throughput. The NVIDIA H100 PCIe debuts the world's highest PCIe card memory bandwidth greater than 2,000 gigabytes per second (GBps). This speeds time to solution for the largest models and most massive data sets.

The NVIDIA H100 PCIe card features Multi-Instance GPU (MIG) capability. This can be used to partition the GPU into as many as seven hardware-isolated GPU instances, providing a unified platform that enables elastic data centers to adjust dynamically to shifting workload demands. As well as one can allocate the right size of resources from the smallest to biggest multi-GPU jobs. NVIDIA H100 versatility means that IT managers can maximize the utility of every GPU in their data center.

NVIDIA H100 PCIe cards use three NVIDIA® NVLink® bridges. They are the same as the bridges used with NVIDIA A100 PCIe cards. This allows two NVIDIA H100 PCIe cards to be connected to deliver 900 GB/s bidirectional bandwidth or 5x the bandwidth of PCIe Gen5, to maximize application performance for large workloads.

The list of qualified H100 servers is TBD.

Figure 1. NVIDIA H100 with NVLink Bridge Volumetric



Specifications

Product Specifications

Table 1 through Table 3 the product, memory, and software specifications for the NVIDIA H100 PCIe card.

Table 1. Product Specifications

Specification	NVIDIA H100
Product SKU	P1010 SKU 200 NVPN: 699-21010-0200-xxx
Total board power	PCIe 16-pin 450 W or 600 W power mode: • 350 W default • 350 W maximum • 200 W minimum PCIe 16-pin 300 W power mode: • 310 W default • 310 W maximum • 200 W minimum
Thermal solution	Passive
Mechanical form factor	Full-height, full-length (FHFL) 10.5", dual-slot
GPU SKU	GH100-200
PCI Device IDs	Device ID: 0x2331 Vendor ID: 0x10DE Sub-Vendor ID: 0x10DE Sub-System ID: 0x1626
GPU clocks	Base: 1,125 MHz Boost: 1,755 MHz
Performance states	P0
VBIOS	EEPROM size: 8 Mbit UEFI: Supported

Specification	NVIDIA H100
PCI Express interface	PCI Express Gen5 x16; Gen5 x8; Gen4 x16 Lane and polarity reversal supported
Multi-Instance GPU (MIG)	Supported (seven instances)
Secure Boot (CEC)	Supported
Zero Power	Not supported
Power connectors and headers	One PCIe 16-pin auxiliary power connector
Weight	Board: 1200g grams (excluding bracket, extenders, and bridges) NVLink bridge: 20.5 grams per bridge (x 3 bridges) Bracket with screws: 20 grams Enhanced straight extender: 35 grams Long offset extender: 48 grams Straight extender: 32 grams

Table 2. Memory Specifications

Specification	Description
Memory clock	1,593 MHz
Memory type	HBM2e
Memory size	80 GB
Memory bus width	5,120 bits
Peak memory bandwidth	2,000 GB/s

Table 3. Software Specifications

Specification	Description[1]
SR-IOV support	Supported -- 32 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: R520 or later Windows: R520 or later

Specification	Description[1]
Secure Boot	Supported
CEC Firmware	Version 2.0025 or later
NVFlash	Version 5.792 or later
NVIDIA® CUDA® support	x86: CUDA 11.8 or later Arm: CUDA 12.0 or later
Virtual GPU software support	Supports vGPU 15.0 or later: NVIDIA Virtual Compute Server Edition
NVIDIA AI Enterprise	Supported with VMWare
NVIDIA certification	NVIDIA-Certified Systems™ TBD or later
PCI class code	0x03 – Display Controller
PCI subclass code	0x02 – 3D Controller
ECC support	Enabled
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:

[1]The KiB, MiB, and GiB notation emphasize 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

Thermal Specifications

Table 4 provides the PCIe reported temperatures and Table 5 provides the thermal specifications for the NVIDIA H100 PCIe card.

Table 4. H100 PCIe Reported Temperatures

Specification	Units	Description
T_{AVG}	°C	Average temperature of all internal GPU sensors
T_{LIMIT}	°C	GPU and HBM temperature limit – current distance in degrees C from software slowdown event
T_{HBM}	°C	Maximum temperature of all HBM sensors

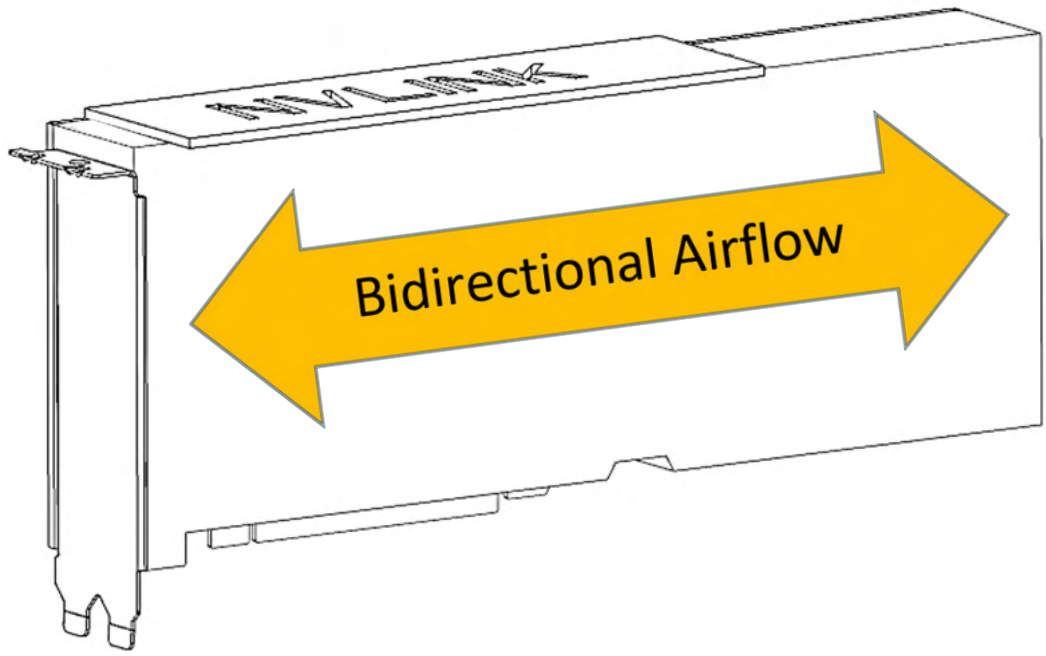
Table 5. Thermal Specifications

Specification	Applies to	Thermal Parameter Value
Thermal qualification temperature	GPU	T_{AVG} = 87°C
	HBM	T_{HBM} = 95°C
Maximum operating temperature	GPU	T_{LIMIT} = 0°C
Hardware slowdown temperature (50% clock slowdown)	GPU	T_{LIMIT} = -2°C
Hardware shutdown temperature	GPU	T_{LIMIT} = -5°C

Airflow Direction Support

The NVIDIA H100 PCIe card employs a bidirectional heat sink, which accepts airflow either left-to-right or right-to-left directions.

Figure 2. NVIDIA H100 Airflow Directions

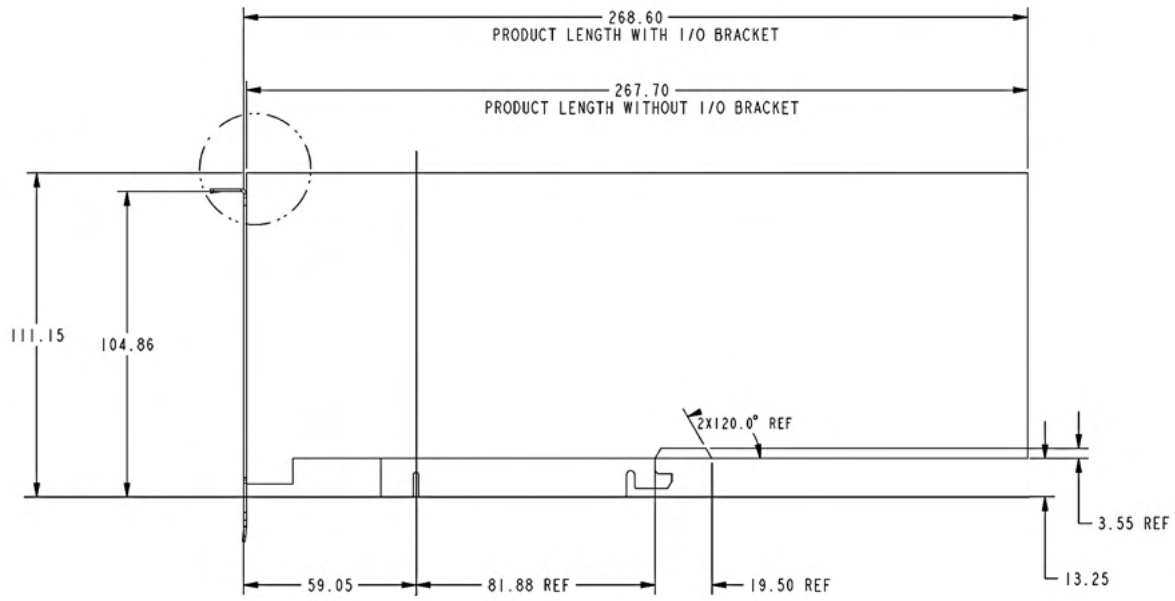


Product Features

Form Factor

The NVIDIA H100 PCIe card conforms to NVIDIA Form Factor 5.5 specification for a full-height, full-length (FHFL) dual-slot PCIe card. For details refer to the *NVIDIA Form Factor 5.5 Specification for Enterprise PCIe Products Specification* (NVOnline: 1063377).

Figure 3. NVIDIA H100 PCIe Card Dimensions



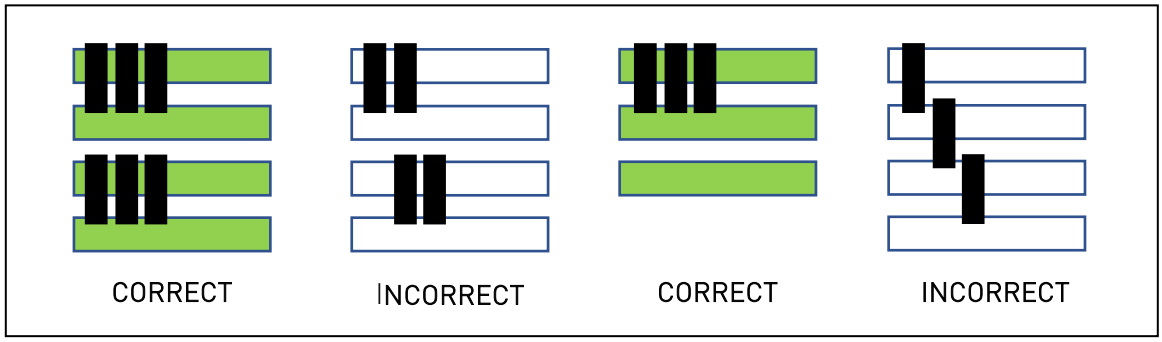
NVLink Bridge Support

NVIDIA NVLink is a high-speed point-to-point (P2P) peer transfer connection. Where one GPU can transfer data to and receive data from one other GPU. The NVIDIA H100 card supports NVLink bridge connection with a single adjacent NVIDIA H100 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 NVIDIA H100 card must incorporate all three NVLink bridges. Wherever an adjacent pair of NVIDIA H100 cards exists in

the server, for best bridging performance and balanced bridge topology, the NVIDIA H100 pair should be bridged. Figure 4 illustrates correct and incorrect NVIDIA H100 NVLink connection topologies.

Figure 4. NVLink Topology – Top View



For systems that feature multiple CPUs, both NVIDIA H100 cards of a bridged card pair, should be within the same CPU domain. That is, under the same CPU's topology, and ensuring this benefits workload application performance. There are exceptions, for example, in a system with dual CPUs wherein each CPU has a single NVIDIA H100 PCIe card under it. In that case, the two NVIDIA H100 PCIe cards in the system may be bridged together. See Section "PCIe and NVLink Topology."

NVIDIA H100 PCIe card, NVLink speed, and bandwidth are given in the following table.

Table 6. H100 PCIe Card NVLink Speed and Bandwidth

Parameter	Value
Total NVLink bridges supported by NVIDIA H100	3
Total NVLink Rx and Tx lanes supported	48
Data rate per NVIDIA H100 NVLink lane (each direction)	100 Gbps
Total maximum NVLink bandwidth	600 Gbytes per second

NVLink Bridge

The 2-slot NVLink bridge for the NVIDIA H100 PCIe card (the same NVLink bridge used in the NVIDIA Ampere Architecture generation, including the NVIDIA A100 PCIe card), has the following NVIDIA part number: 900-53651-0000-000.

NVLink Connector Placement

Figure 5 shows the connector keepout area for the NVLink bridge support of the NVIDIA H100.

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 H100 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 5.5 Specification for Enterprise PCIe Products Specification* (NVOnline: 1063377) for more detailed information.

NVLink bridge interfaces of the H100 PCIe card include removable caps to protect the interfaces in non-bridged system configurations.

PCIe and NVLink Topology

As stated, it is strongly recommended that both NVIDIA H100 PCIe cards of a bridged card pair should be within the same CPU topology domain. Unless a dual CPU system has only two H100 PCIe cards each of which is under its own CPU. Full NVLink connection topology guidance is as follows:

- **Best NVLink Topology (Recommended):**
  - Bridge two GPUs under the same CPU or PCIe switch
  - GPU count in a system should be in powers of two (1, 2, 4, 8, and so on)
  - Locate the same (even) number of GPUs under each CPU socket
  - Maintain a balanced configuration: same count of CPU:GPU:NIC for each grouping
- **Good NVLink Topology:**
  - Bridge two GPUs under different PCIe switches but under the same CPU
  - Same number of GPUs and NICs under each CPU socket, but not powers of 2
- **Allowed but Not Recommended:**
  - Bridge two GPUs under two different CPUs
  - Odd number of GPUs under each CPU
  - Unbalanced configurations: Different ratios of CPU:GPU:NIC for each grouping

Power Connector

This section details the power connector for the NVIDIA H100 PCIe card.

Power Connector Placement

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

Figure 6. PCIe 16-Pin Power Connector

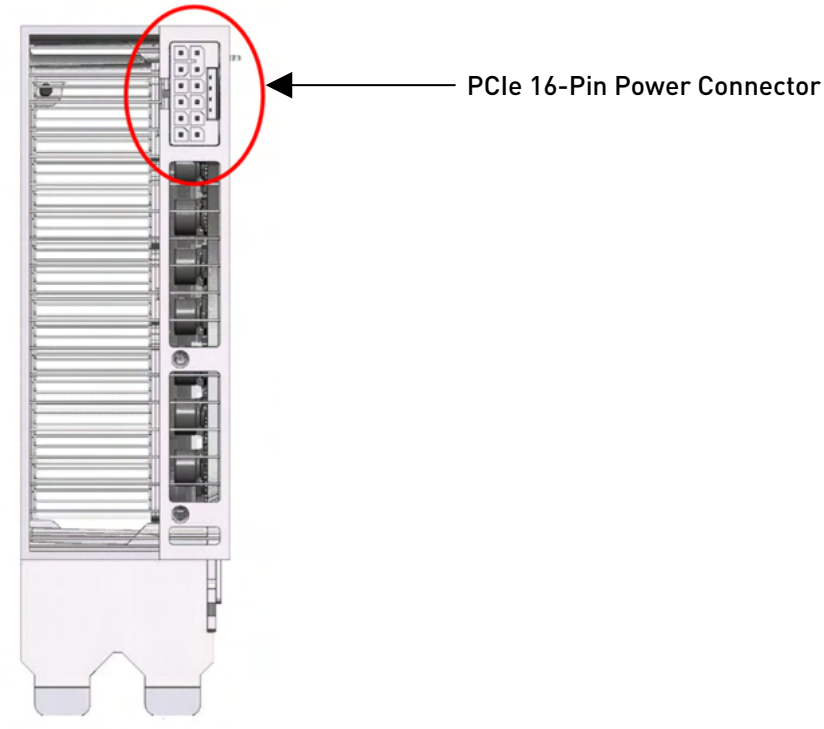


Table 7 lists the power level options identifiable by the PCIe 16-pin power connecter per CEM5 PSU, and the corresponding Sense0 and Sense1 logic. The NVIDIA card senses the Sense0 and Sense1 levels and recognizes the power available to the NVIDIA card from the power connector. If the power level identified 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 H100 requires up to 350 W from the 16-pin auxiliary power connector. Table 7 shows the supported auxiliary power connector sense pin logic and maximum supported TGP per power level.

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

Power Level	Sideband 3 (Sense0)	Sideband 4 (Sense1)	Maximum TGP
451 - 600 W	0	0	350 W
301 - 450 W	1 (float)	0	350 W
151 - 300 W	0	1 (float)	310 W
Up to 150 W	1 (float)	1 (float)	Not supported. Insufficient power

Table 8 lists supported auxiliary power connections for the NVIDIA H100 GPU card.

Table 8. Supported Auxiliary Power Connections

Board Connector	PSU Cable
PCIe 16-pin	PCIe 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 PCIe 16-pin power adapter is available for systems that do not have native PCIe 16-pin power connectors. Figure 7 illustrates the power adapter. The power adapter provided by NVIDIA can only support 310 W TGP operation. Partners are advised to build their own power adapters (if necessary) to support the 301 W-450 W power sense option to enable full 350 W TGP operation of the H100 PCIe card.

- NVPN: 030-1546-000 – CPU 8-pin to PCIe 16-Pin Power Adapter
- Astron MFN: DAMAF01041-H

Figure 7. CPU 8-Pin to PCIe 16-Pin Power Adapter

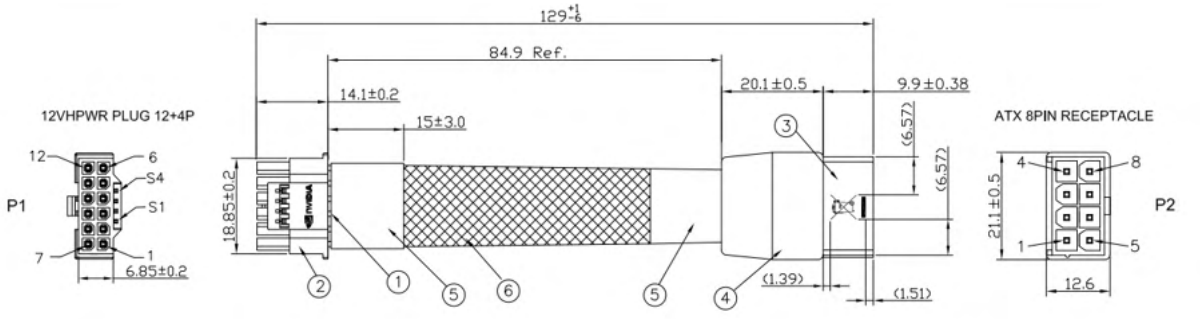
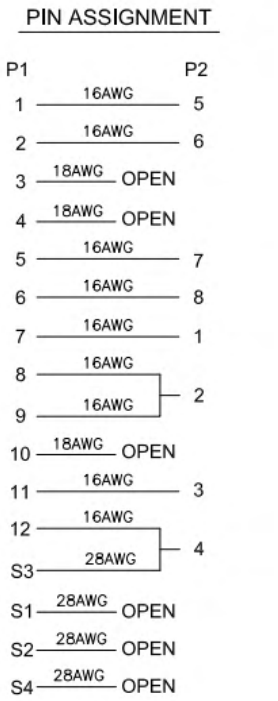


Figure 8 shows the CPU 8-pin to PCIe 16-pin power adapter pin assignments.

Figure 8. CPU 8-Pin to PCIe 16-Pin Power Adapter Pin Assignments



Note: The power adapter supports the four Sideband signals, hardware-strapped per Row 3. This strapping corresponds to the "151 – 300 W" power level PCIe CEM 5.0 specification (shown in Row 3). As a result, it supports only a 310 W TGP for the H100 PCIe card. To support a TGP of 350 W, a power adapter or cable strapped as in Row 1 or Row 2 should be used.

Power Adapter Availability

The power adapter is provided with sample NVIDIA H100 PCIe cards only. For production cards, consult NVIDIA applications engineering for qualified suppliers of a power adapter.

Extenders

The NVIDIA H100 PCIe card provides two extender options, shown in Figure 9 and Figure 10.

- NVPN: 151-0398-000 -- Enhanced Straight Extender
  - Card + extender = 312 mm
- NVPN: 682-00007-5555-001 – Straight extender
  - Card + extender = 312 mm
- NVPN: 682-00007-5555-000 – Long offset extender
  - Card + extender = 339 mm

Using a 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 9. Enhanced Straight Extender

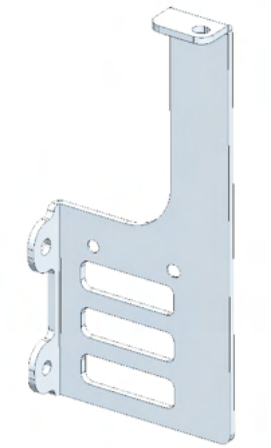
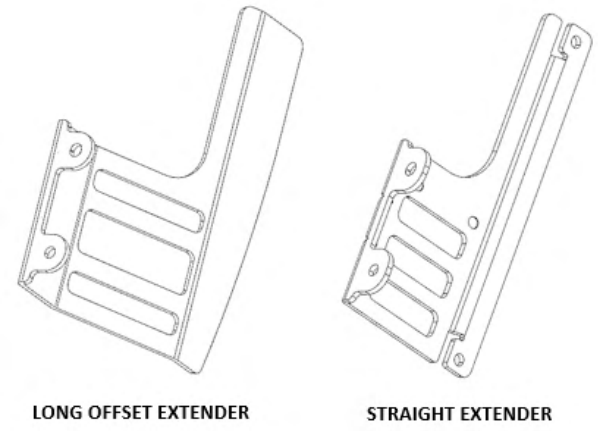


Figure 10. Legacy Long Offset and Straight Extenders



NVIDIA AI Enterprise Software Suite

H100 for mainstream servers comes with a five-year subscription. It includes enterprise support to the NVIDIA AI Enterprise software suite and simplifying AI adoption with the highest performance. This ensures organizations have access to the AI frameworks and tools they must build H100-accelerated AI workflows such as AI chatbots, recommendation engines, vision AI, and more.

Customers can activate their licenses at: https://www.nvidia.com/activate-h100/

The OS of the NVIDIA AI platform, NVIDIA AI Enterprise is essential for production and support of applications built with the extensive NVIDIA library of frameworks and pre-trained models such as NVIDIA® Riva for speech AI, NVIDIA Merlin™ for recommendation engines, NVIDIA Clara™ for medical imaging and more. Certified to deploy NVIDIA-Certified Systems from leading server vendors.

Optimize every step of the AI workflow including data prep, model training, inference, and deployment at scale with NVIDIA AI tools and frameworks.

- Accelerate Data Prep with NVIDIA RAPIDS™
- Train at Scale with the NVIDIA TAO Toolkit
- Optimized for Inference NVIDIA® TensorRT™
- Deploy to Scale NVIDIA Triton™ Inference Server

A broad ecosystem of certified partner integrations reduces deployment risk.

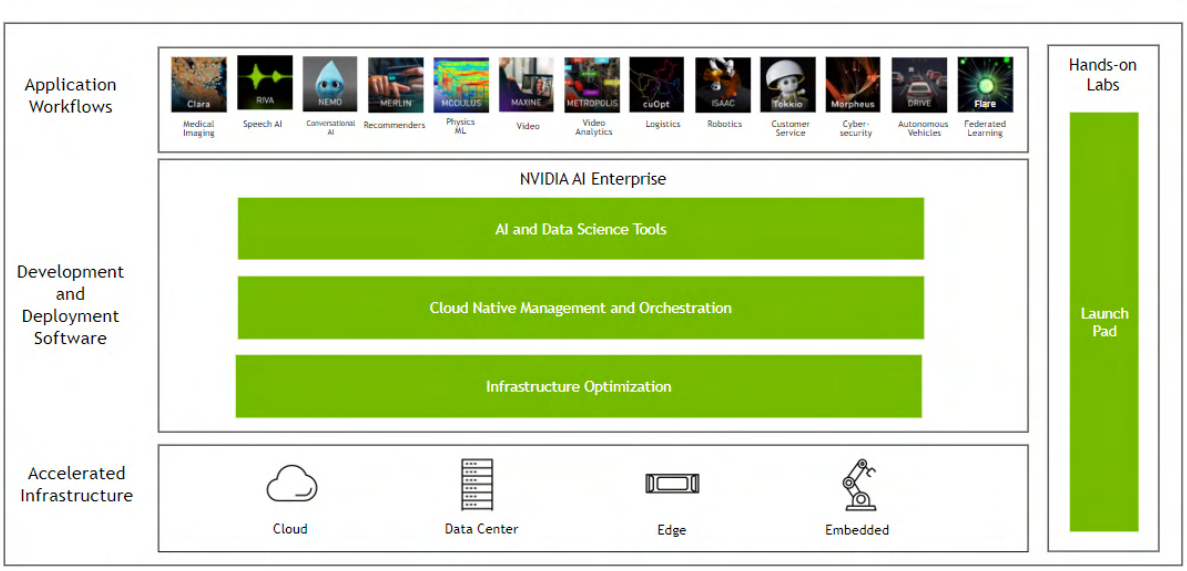
- MLOps solution providers for collaboration and productivity,
- VMware vSphere, VMware Cloud Foundation and VMware Cloud Director to scale in virtualized environments
- Red Hat OpenShift certification

Support for Production AI

Organizations get the transparency of open source with the assurance of support from NVIDIA when they move from development to production.

- Full enterprise support for their AI workflows.
  - Conversational AI
  - Recommendation engines
  - Medical imaging and more.
- Access to NVIDIA AI experts and engineering teams.
- Priority notifications related to the latest security fixes and maintenance releases.
- Enterprise Training Services included so developers, data scientists, and IT professionals get the most out of NVIDIA AI Enterprise.
- Long-term support (LTS) for up to three-years for designated software branches.
- Customized support upgrade options, including a designated technical account manager (TAM) and Business Critical support for 24x7 live agent access.

Figure 11. NVIDIA AI Enterprise Software Stack



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 9. 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.

Trademarks

NVIDIA, the NVIDIA logo, CUDA, NVIDIA-Certified System, NVIDIA Clara, NVIDIA Hopper, NVIDIA Merlin, NVIDIA RAPIDS, NVIDIA Triton, NVLink, and TensorRT 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