NVIDIA L40S GPU Accelerator

Product Brief

PB-11470-001_v02 | August 2023

Document History

PB-11470-001_v02

Version	Date	Authors	Description of Change
01	August 2, 2023	VNK, SM	Initial release
02	August 8, 2023	VNK, NM, SM	Updated power adapter section title to “3x PCIe 8-Pin to PCIe 16-Pin Power Adapter”

Table of Contents

List of Figures

List of Tables

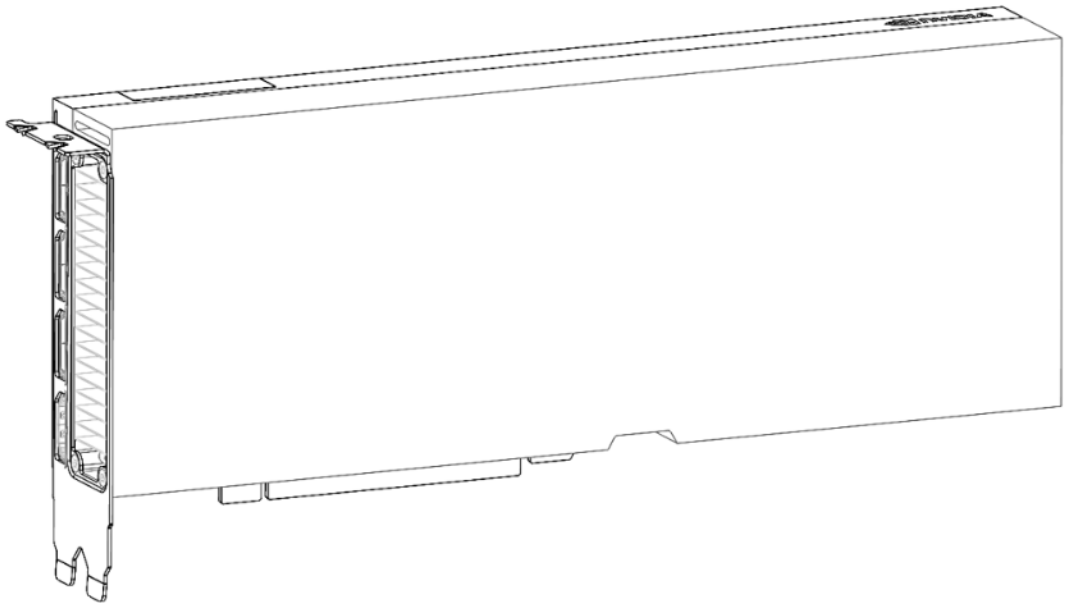
Overview

The NVIDIA L40S GPU Accelerator is a full height, full-length (FHFL), dual-slot 10.5 inch PCI Express Gen4 graphics solution based on the NVIDIA Ada Lovelace architecture. The card is passively cooled and capable of 350 W maximum board power.

The NVIDIA L40S 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 deep learning inference as well as training workloads enabling generative AI workloads, batch and real-time rendering, virtual workstations, and cloud gaming.

As part of the NVIDIA OVX™ server platform, L40S 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.

Figure 1. NVIDIA L40S PCIe Card



Specifications

Product Specifications

Table 1 through Table 3 provide the product, memory, and software specifications for the NVIDIA L40S PCIe cards.

Table 1. Product Specifications

Specifications	NVIDIA L40S
Product SKU	PG133 SKU 242 NVPN: 669-2G133-242-xxx
Total board power	350 W default 350 W maximum TBD W minimum
Thermal solution	Passive
Mechanical form factor	Full-height, full-length (FHFL) 10.5”, dual-slot
PCI Device IDs	Device ID: 26B9 Vendor ID: 0x10DE Sub-Vendor ID: 0x10DE Sub-System ID: 0x1851
GPU clocks	Base: 1,065 MHz Boost: 2,520 MHz
Performance States	P0, P8
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
Connectors	One PCIe 16-pin auxiliary power connector Four VESA® DisplayPort® connectors
Weight	Board: 1,052 grams (excluding bracket, and extenders)

Specifications	NVIDIA L40S
	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	9,001 MHz
Memory type	GDDR6
Memory size	48 GB
Memory bus width	384 bits
Peak memory bandwidth	864 GB/s

Table 3. Software Specifications

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: R535TRD1 or later Windows: R535TRD1 or later
Secure boot	Supported (See Section “CEC Hardware Root of Trust”)
CEC firmware	v2.0134 or later
NVFlash	Version 5.814 or later
NVIDIA® CUDA® support	CUDA 12.2 or later
Virtual GPU software support	Supports vGPU 16.1 (R535 GA6) or later: NVIDIA Virtual Compute Server Edition

Specification	Description[1]
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, 0xA0, 0x40
SMBus direct access	Supported
SMBPBI SMBus Post-Box Interface	Supported

Note: [1]The KiB, MiB, and GiB notations emphasize the "power of two" nature of the values. Thus,

- 256 KiB = 256 × 1024
- 16 MiB = 16 × 1024[2]
- 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 L40S 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] TBD hours at 35°C Controlled environment:[3] TBD hours at 35°C

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

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

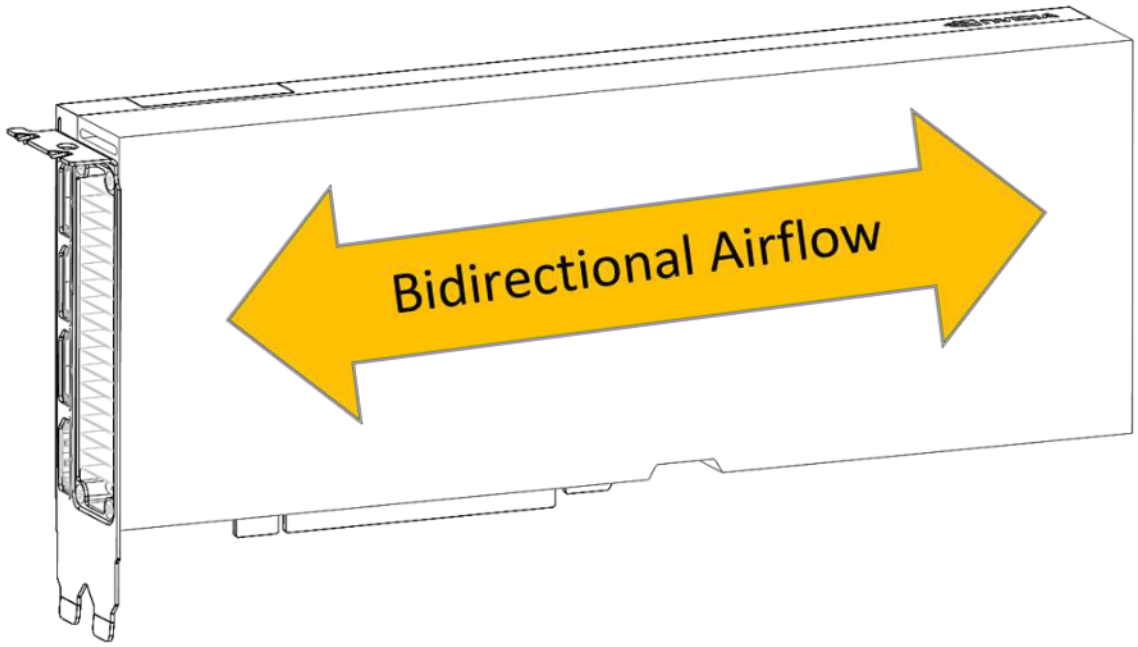
[2]Some environmental stress with limited maintenance (GF35).

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

Airflow Direction Support

The NVIDIA L40S PCIe card employs a bidirectional heat sink, which accepts airflow either left-to-right or right-to-left directions.

Figure 2. L40S Airflow Directions



Product Features

PCI Express Interface Specifications

The following subsections describe the PCIe interface specifications for the NVIDIA L40S PCIe card.

PCIe Support

The NVIDIA L40S GPU card supports PCIe Gen4. Either a Gen4 x16, Gen4 x8, or Gen3 x16 interface should be used when connecting to the NVIDIA L40S 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 L40S 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 can only 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 L40S 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 L40S 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, as defined in the PCIe specification, is supported on the NVIDIA L40S PCIe card.

Lane Reversal, as defined in the PCIe specification, is supported on the NVIDIA L40S PCIe card. When reversing the order of the PCIe lanes, the order of both the Rx lanes and the Tx lanes must be reversed.

CEC Hardware Root of Trust

The NVIDIA L40S provides secure boot capability through 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.

The hardware 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.

Display

This section details the operating modes for NVIDIA L40S.

Display On and Off

The L40S 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)
Scalable visualization	BAR1: 8 GiB (Display On, 8 GB BAR1 mode)
Display enabled	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 L40S with NVIDIA® Quadro® vDWS software enables the user to tackle massive datasets, large 3D models, and complex designs with scaled memory and performance. NVIDIA L40S supports all four editions of NVIDIA virtual GPU software:

- NVIDIA RTX™ Virtual Workstation vDWS
- NVIDIA GRID® Virtual Applications (GRID vApps)
- NVIDIA GRID Virtual PC (GRID vPC)
- NVIDIA Virtual Compute Server (vCS)

Display On 8GB BAR1 Mode

The Display On, 8 GB BAR1 mode is the recommended configuration for scalable visualization system deployments. In this mode, the NVIDIA L40S 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 L40S cards is accomplished by use of the NVIDIA Quadro Sync II card. The *Quadro Sync II User's Guide* (DU-08348-001) provides specifications and usage guidance for this technology.

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 L40S 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 L40S cards is accomplished by use of the NVIDIA Quadro Sync II card. The *Quadro Sync II User's Guide* (DU-08348-001) provides specifications and usage guidance for this technology:

Switching Operating Modes

The NVIDIA GPUModeSwitch will be provided to switch modes. An example command to switch GPU mode is shown:

```
sudo ./gpumodeswitch --gpumode <mode_name>
```

To display the list of gpumodes available, the following command may be used:

```
sudo ./gpumodeswitch --listgpumodes
```

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

Frame Lock

The NVIDIA L40S 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 L40S frame lock and stereo connectors are on the card's north edge.

Display Bracket

The L40S card provides a display bracket that may be removed for system configurations that do not use the NVIDIA bracket. See the attached bracket mechanical collateral for more specifics on bracket design.

Programmable Power

The Programmable Power feature provides partners the general 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 L40S 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 L40S 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 350
```

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 is loaded for full functionality. The power cap can be adjusted through the following asynchronous command:

Table 6. SMBPBI Commands

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 Guide* (DG-06034-002) for full implementation details.

IPMI FRU EEPROM

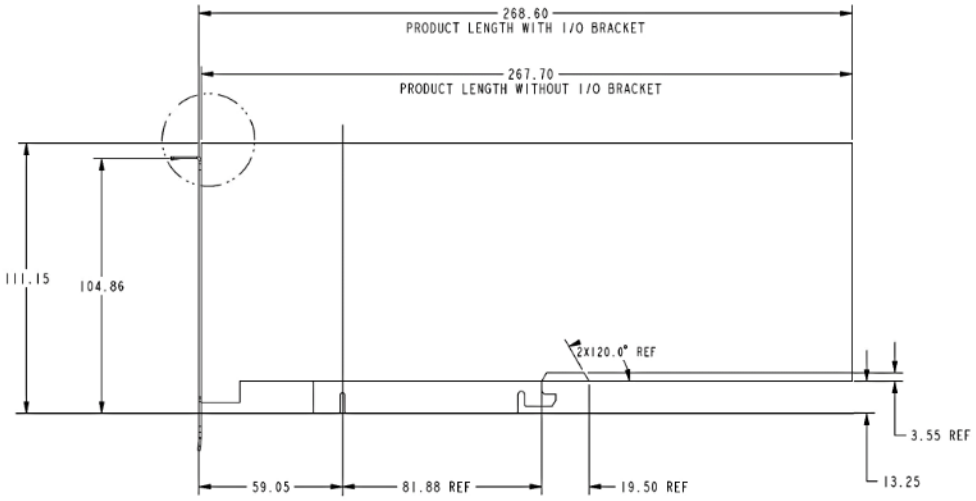
The NVIDIA L40S supports the *Intelligent Platform Management Interface (IPMI) FRU Specification* v1.0 r1.3. See Table 3 for the FRU EEPROM I2C address.

Form Factor

The NVIDIA L40S 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).

In this product brief, nominal dimensions are shown.

Figure 3. NVIDIA L40S PCIe Card Dimensions



Power Connector

This section details the power connector for the NVIDIA L40S 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

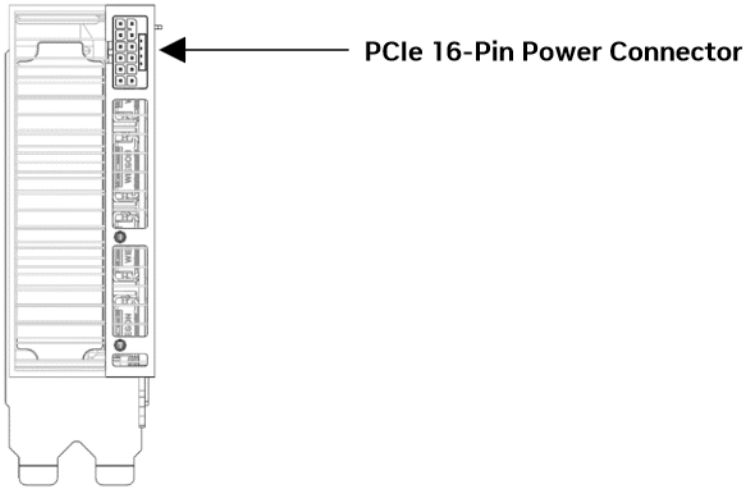
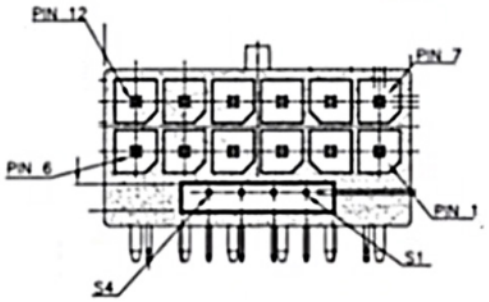


Figure 4 3 shows the pin assignments for the PCIe 16-pin power connector, per PCIe CEM 5.0 specification.

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 L40S requires up to 350 W from the 16-pin power connector. Thus, the top two rows of Table 7 are acceptable. The configurations [Sense0=1, Sense1=1] and [Sense0=0, Sense1=1] will prevent the NVIDIA L40S from booting.

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

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 L40S GPU card.

Table 8. Supported Auxiliary Power Connections

Board Connector	PSU Cable
PCIe 16-pin	PCIe 16-pin
PCIe 16-pin	3x PCIe 8-pin to PCIe 16-pin

3x PCIe 8-Pin to PCIe 16-Pin Power Adapter

A 3x PCIe 8-pin to PCIe 16-pin power adapter for systems that do not have native PCIe 16-pin power connectors may be used. The L40S is a 350 W product and will not boot if insufficient power is indicated by the system. Power delivery cabling must be strapped for the 450 W or 600 W power level for the L40S to boot.

Extenders

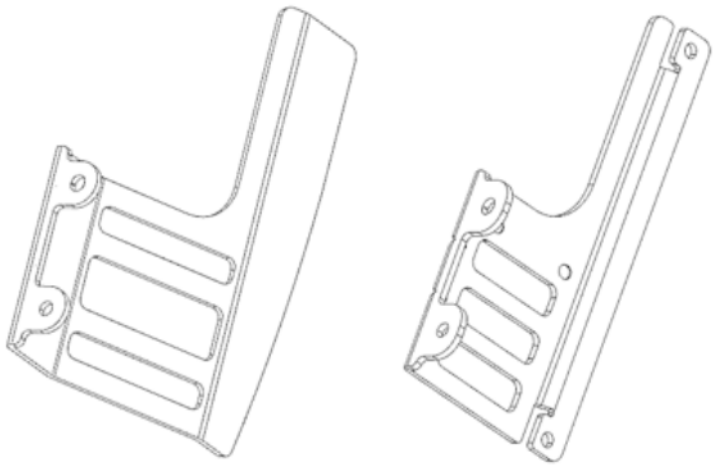
The L40S PCIe card provides two extender options, shown in Figure 6.

- 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 6. Long Offset and Straight Extenders



Support Information

Certification

- Windows Hardware Quality Lab (WHQL):
  - Windows 10, Windows 11
  - 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

The following table lists the languages supported by the L40S GPU Accelerator.

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]Windows 10, Windows 11, Windows Server 2019 R2, and Windows Server 2021 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 GPU Boost, NVIDIA GRID, NVIDIA Omniverse, NVIDIA OVX, 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

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