NVIDIA L4 GPU Accelerator

Product Brief

PB-11316-001_v01 | March 2023

Document History

PB-11316-001_v01

Version	Date	Authors	Description of Change
01	March 9, 2023	AV, SM	Initial release

Table of Contents

List of Figures

List of Tables

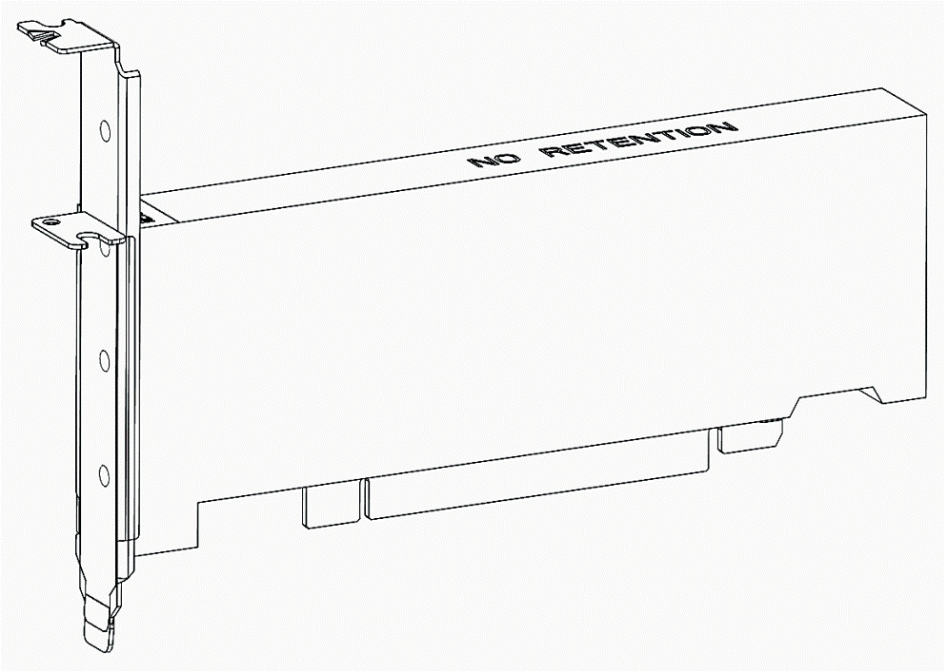
Overview

The NVIDIA L4 Tensor Core GPU delivers a versatile platform to accelerate Deep Learning, Graphics and Video processing applications in the Cloud and at the Edge. It is a half-height (low profile), half-length, single slot card featuring 24 GB of GDDR6 memory, x16 PCIe Gen4 connectivity at a 72 W maximum power envelope. It is a passively cooled card with a superior thermal design-requiring system airflow to operate and handles challenging ambient environments with ease (NEBS-3 capable).

Powered by the NVIDIA Ada Lovelace architecture, L4 provides revolutionary multi-precision performance to accelerate deep learning and machine learning training and inference, video transcoding, AI audio (AU) and video effects, rendering, data analytics, virtual workstations, virtual desktop, and many other workloads.

As part of NVIDIA AI, the L4 supports all AI frameworks and neural network models, delivering dramatic performance and efficiency that maximizes the utility of at-scale deployments.

Figure 1. NVIDIA L4 NVFF 5.5 HHHL with Full Height Bracket



Specifications

Product Specifications

Table 1 through Table 3 the product, memory, and software specifications for the NVIDIA L4 PCIe card.

Table 1. Product Specifications

Specification	NVIDIA L4
Product SKU	PG193 SKU 200 NVPN: 699-2G193-0200-xxx
Total board power	72 W default 72 W maximum 40 W minimum
Thermal solution	Passive
Mechanical form factor	HHHL-SS (half-height, half-length, single-slot)
PCI Device IDs	Device ID: 0x27B8 Vendor ID: 0x10DE Sub-Vendor ID: 0x10DE Sub-System ID: 0x16CA
Four-part ID (VID:DEVID:SVID:SSID)[1]	10DE:27B8:10DE:16CA
GPU clocks	Base: 795 MHz Boost: 2,040 MHz
VBIOS	EEPROM size: 16 Mbit UEFI: Supported
Drivers	Linux: R525 or later Windows: R525 or later
PCI Express interface	Physical x16 PCIe lanes PCIe Gen4 x16, x8; Gen3 x16 Lane and polarity reversal supported
Performance states	P0, P8
Zero Power	Not supported

Specification	NVIDIA L4
Weight	Board: 270 Grams (excluding bracket) Bracket (Full height) with screws: 14 Grams Bracket (Half height) with screws: 9 Grams

Note:
[1]The NVIDIA L4 is uniquely identified by its complete four-part ID.

Table 2. Memory Specifications

Specification	Description
Memory clock	6,251 MHz
Memory type	GDDR6
Memory size	24 GB
Memory bus width	192 bits
Peak memory bandwidth	300 GB/sec

Table 3. Software Specifications

Specification	NVIDIA L4
SR-IOV support	Supported: 32 VF (virtual functions)
BAR address (physical function)	BAR0: 16 MiB[1] BAR1: 32 GiB[1] BAR3: 32 MiB[1]
BAR address (virtual function)	BAR0: 8 MiB (256 KiB per VF)[1] BAR1: 64 GiB, 64-bit (2 GiB per VF)[1] BAR3: 1 GiB, 64-bit (32 MiB per VF)[1]
Message signaled interrupts	MSI-X: Supported MSI: Not supported
ARI Forwarding	Supported
Secure Boot	Supported (See “Root of Trust” section)
NVIDIA® CUDA® support	CUDA 12.0 or later
Virtual GPU software support	Supports vGPU 15.2 or later
PCI class code	0x03 – Display controller
PCI sub-class code	0x02 – 3D controller
ECC support	Enabled (by default); can be disabled using software
SMBus (8-bit address)	0x9E (write), 0x9F (read)
IPMI FRU EEPROM I2C address	0x50 (7-bit), 0xA0 (8-bit)
Reserved I2C addresses	0xA0, 0xAA, 0xAC
SMBus direct access	Supported

Specification	NVIDIA L4
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

Environmental and Reliability Specifications

Table 4 provides the environment conditions specifications for the L4 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] 2,147,604 hours at 35°C Controlled environment:[3] 2,785,669 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.

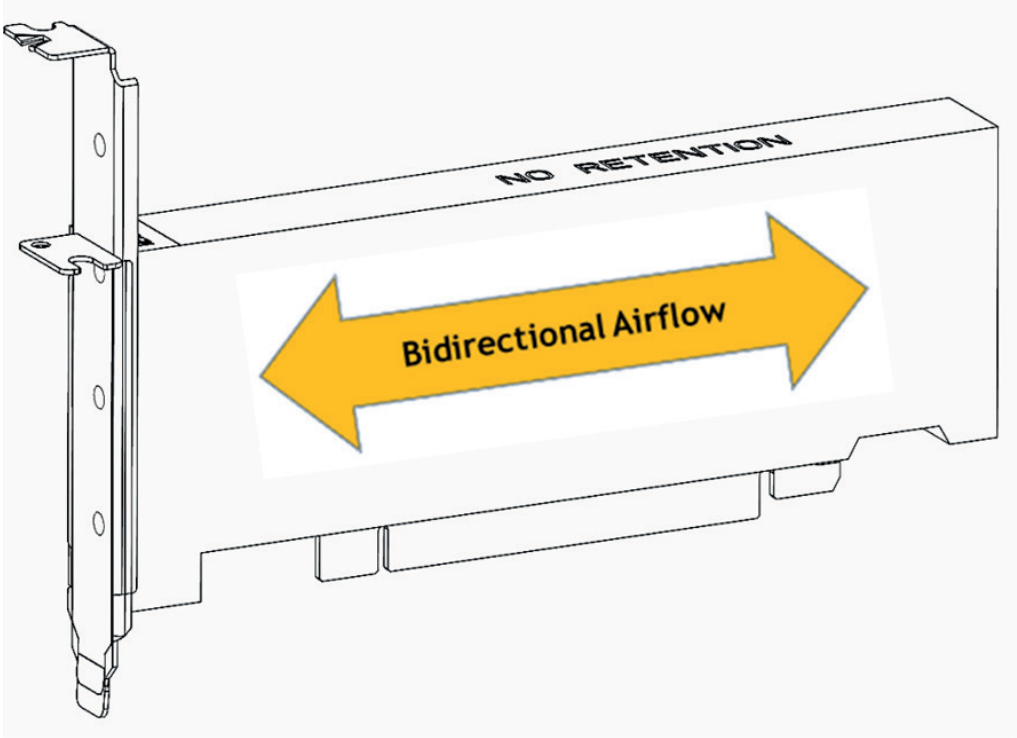
[2]Some environmental stress with limited maintenance (GF35).

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

Airflow Direction Support

The NVIDIA L4 PCIe card employs a bidirectional heat sink, which accepts airflow either left-to-right or right-to-left directions.

Figure 2. NVIDIA L4 Airflow Direction



Product Features

PCI Express Interface Specifications

The following subsections describe the PCIe interface specifications for the L4 PCIe card.

PCIe Support

The L4 card supports PCIe Gen4. Gen4 x16 interface should be used when connecting to the L4 PCIe card.

Single Root I/O Virtualization Support

Single Root I/O (SR-IOV) Virtualization 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 L4 PCIe card. The number of VFs supported is given in Table 3.

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 L4 requires that SBIOS and software support in the OS instance or hypervisor is configured to enable support SR-IOV. The OS instance or hypervisor must be able to detect and initialize PFs and VFs.

Interrupt Messaging

The L4 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 L4 PCIe card.

Lane Reversal, as defined in the PCIe specification, is supported on the L4 PCIe card. When reversing the order of the PCIe lanes, the order of both the Rx lanes and the Tx lanes must be reversed.

Root of Trust

The NVIDIA L4 provides a primary root of trust within the GPU that provides the following:

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

Form Factor

The NVIDIA L4 PCIe card conforms to NVIDIA Form Factor 5.5 specification for a half-height (low profile) half-length (HHHL) single slot PCIe card. For details refer to the *NVIDIA Form Factor 5.5 for Enterprise PCIe Products Specification* (NVOnline reference number 106337).

In this product brief, nominal dimensions are shown.

Figure 3. NVIDIA L4 PCIe Card Dimensions with Full Height Bracket

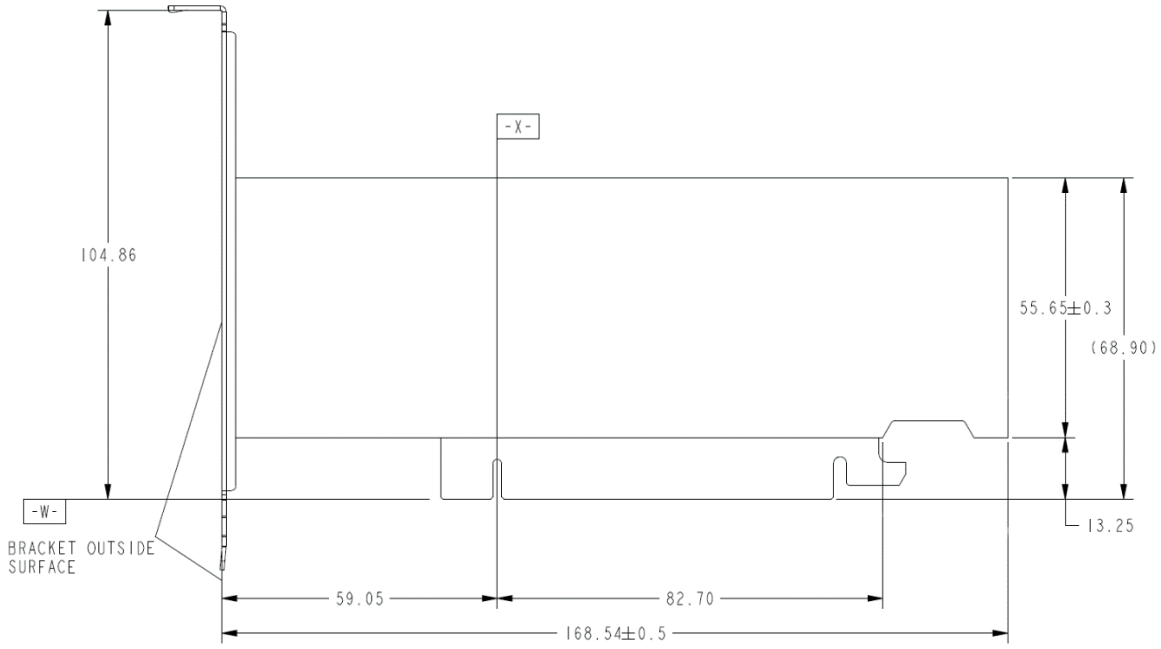
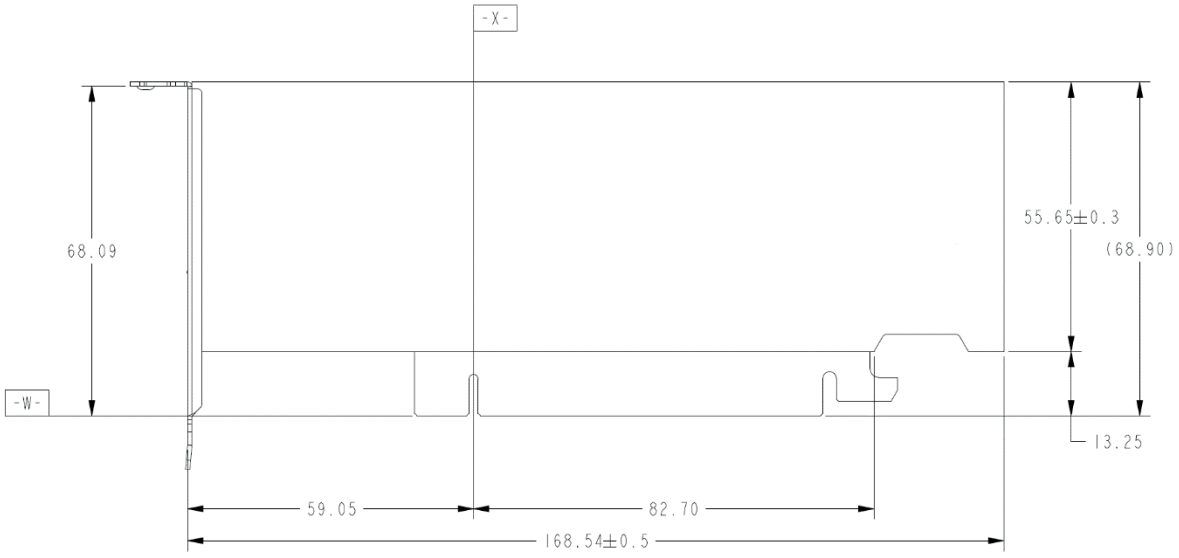


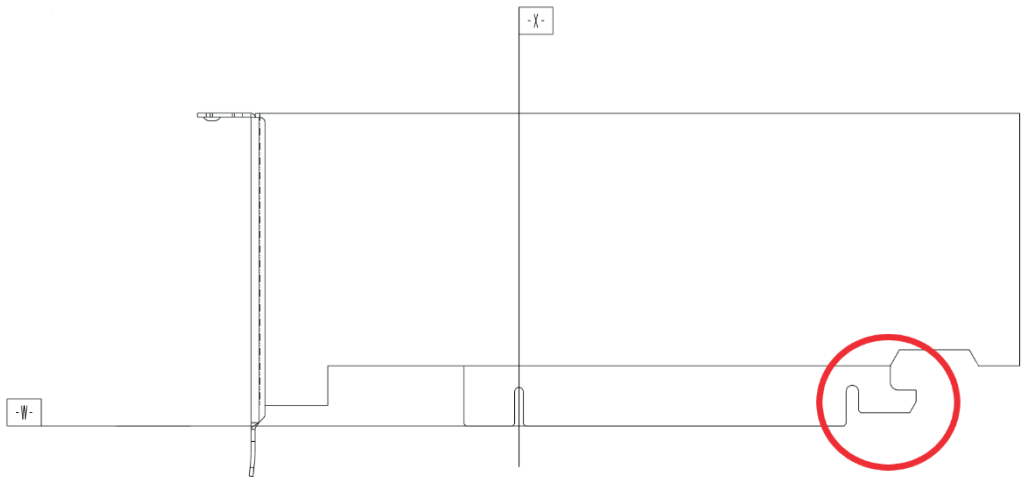
Figure 4. NVIDIA L4 PCIe Card Dimensions with Low Profile Bracket



Hockey Stick Board Retention

The NVDIA L4 enables south edge board retention using a “hockey stick” tab located to the east of the PCIe card fingers, as shown in Figure 5.

Figure 5. NVIDIA L4 Hockey Stick Tab



Languages

Table 5. 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 10, Windows 11, Windows Server 2019, and Windows Server 2022 Windows are supported.**

www.openzeka.com/l4

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051