TESLA V100 PCIe GPU ACCELERATOR

PB-08744-001_v03 | October 2017

Product Brief

DOCUMENT CHANGE HISTORY

PB-08744-001_v03

Version	Date	Authors	Description of Change
01	September 19, 2017	GG, SM	Initial Release
02	September 19, 2017	TC, SM	Removed graphics mode support in Table 3
03	October 6, 2017	GG, SM	Note added after “Extender” section

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

OVERVIEW

The NVIDIA® Tesla® V100 GPU Accelerator for PCIe is a dual-slot 10.5 inch PCI Express Gen3 card with a single NVIDIA Volta GV100 graphics processing unit (GPU). It uses a passive heat sink for cooling, which requires system air flow to properly operate the card within its thermal limits. The Tesla V100 PCIe supports double precision (FP64), single precision (FP32) and half precision (FP16) compute tasks, unified virtual memory and page migration engine.

For performance optimization, NVIDIA GPU Boost™ feature is supported. By automatically adjusting the GPU clock dynamically, maximum performance is achieved within the power cap limit.

Tesla V100 PCIe boards are shipped with ECC enabled by default to protect the GPU's memory interface and the on-board memories. ECC protects the memory interface by detecting any single, double, and all odd-bit errors. The GPU will retry any memory transaction that has an ECC error until the data transfer is error-free. ECC protects the DRAM content by fixing any single-bit errors and detecting double-bit errors. The Tesla V100 PCIe with 16GB of HBM2 memory has native support for ECC and has no ECC overhead, both in memory capacity and bandwidth.

Tesla V100 PCIe supports Maximum Performance (Max-P) and Maximum Efficiency (Max-Q) modes. In Max-P mode, the Tesla V100 PCIe Accelerator will operate unconstrained up to its thermal design power (TDP) level of 250 W to accelerate applications that require the fastest computational speed and highest data throughput.

Max-Q mode allows data center managers to tune power usage of their Tesla V100 PCIe Accelerators to operate with optimal performance per watt. A power limit can be set via software across all GPUs in a rack, reducing power consumption dramatically, while still obtaining excellent rack performance for target applications. Max-Q is not tied to a specific power number. The data center manager can set the Tesla V100 to a power budget (as long as it is below 250 W) that delivers the best perf/watt for the target

workload. Max-Q gives the data center manager the flexibility to optimize the throughput at a node-level, rack-level or data center-level based on the power budget.

Note: All occurrences of Volta refer to the NVIDIA project code name.

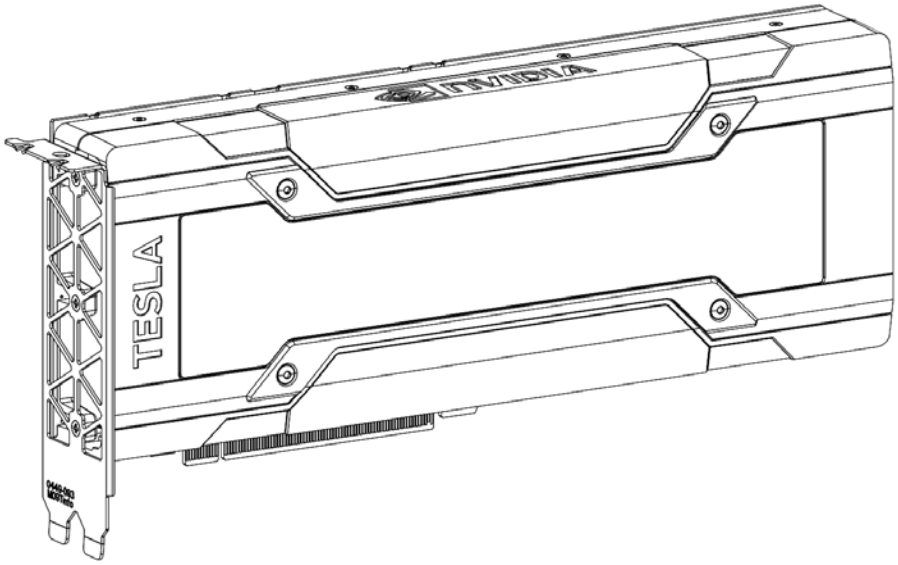

Figure 1. Tesla V100 PCIe Board (With Optional I/O Bracket)

SPECIFICATIONS

PRODUCT SPECIFICATIONS

Table 1 provides the product specifications for the Tesla V100 PCIe board.

Table 1. Product Specifications

Specification		Tesla V100 PCIe 16GB
Product SKUs		NVPN: 699-2G500-0200-XXX
Total board power		250 W
GPU SKUs		GV100-893-A1
PCI Device IDs		Device ID: 0x1DB4 Vendor ID: 0x10DE Sub-Vendor ID: 0x10DE Sub-System ID: 0x1214
GPU clocks	Base	1245 MHz
	Maximum boost	1380 MHz
VBIOS	EEPROM size	8 Mbit
	UEFI	Supported
PCI Express interfaces		PCI Express 3.0 ×16, Lane and polarity reversal supported
Power connectors and headers		One CPU 8-pin auxiliary power connector
Weight	Board	1196 Grams
	Bracket with screws	21 Grams
	Long offset extender	52 Grams
	Straight extender	42 Grams

Table 2 provides the memory specifications for the Tesla V100 PCIe board.

Table 2. Memory Specifications

Specification	Tesla V100 PCIe 16GB
Maximum memory clock	877 MHz
Memory size	16 GB HBM2
Memory bus width	4096-bit
Peak memory bandwidth	Up to 900 GB/s

Table 3 provides the software specifications.

Table 3. Software Specifications

Specification	Description
Compatibility mode supported	Compute only
Base address	BAR0: 16 MB BAR1: 16 GB BAR3: 32 MB
PCI class code	0x03 - Display Controller
PCI sub-class code	0x02 – 3D Controller
ECC support	Supported (Enabled by default)
SMBus (8-bit address)	0x9E (write), 0x9F (read)
SMBus direct access	Supported
SMBPBI (SMBus Post Box Interface)	Supported
Max customer boost clock	Supported
Zero Power	Not supported

Table 4 provides the environment conditions specifications for the Tesla V100 PCIe board.

Table 4. Board Environment and Reliability Specifications

Specification	Condition
Operating temperature	0 °C to 45 °C
Storage temperature	-40 °C to 75 °C
Operating humidity	5% to 90% relative humidity
Storage humidity	5% to 95% relative humidity
Mean time between failures (MTBF)	Uncontrolled environment: TBD hours at 35 °C Controlled environment: TBD hours at 35 °C

Note: MTBF data is currently being measured and will be published in a later revision of this product brief.

THERMAL SPECIFICATIONS

Table 5 provides the thermal specifications for the Tesla V100 PCIe board.

Table 5. Thermal Specifications

Parameter	Value	Units
Total board power	250	W
GPU thermal qualification temperature	80	°C
GPU maximum operating temperature	83	°C
HBM maximum operating temperature	85	°C
GPU slowdown temperature (50% clock slowdown)	87	°C
GPU shutdown temperature	90	°C

SYSTEM AIRFLOW REQUIREMENTS

Airflow Direction Support

The Tesla V100 PCIe board employs a bidirectional heat sink, which accepts airflow either left-to-right or right-to-left directions.

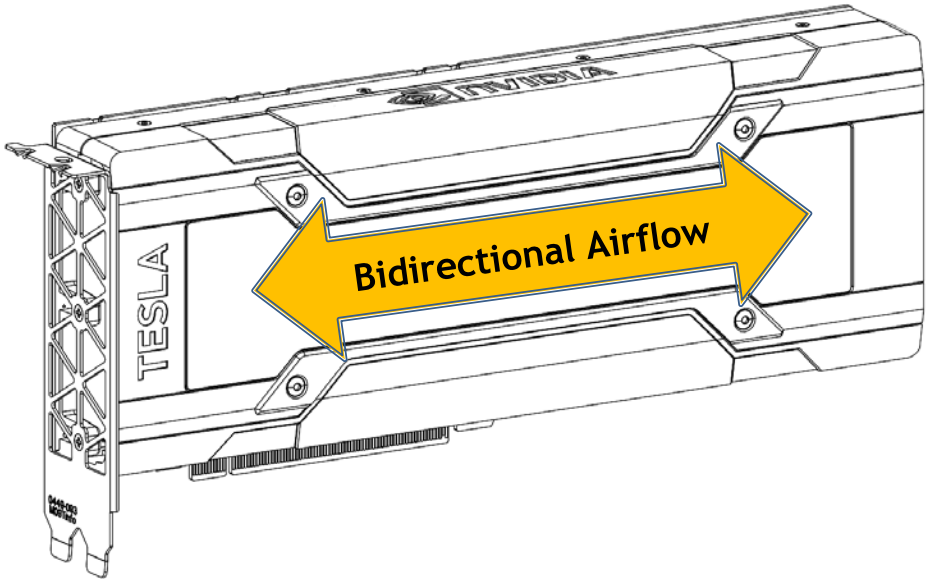

Figure 2. Tesla V100 PCIe Airflow Directions (With Optional I/O Bracket)

MAX-Q MODE

Max-Q is defined as the point that delivers the best performance/watt for a given workload. Different workloads may have different Max-Q points.

Data center managers can tune power usage of their Tesla V100 PCIe Accelerators via nvidia-smi to any value below 250 W. For example, when the workload does not need all 250 W or the rack is power constrained, the board power can be set to a lower level.

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 is shown (power limit 180 W):

```
nvidia-smi -pm 1
```

```
nvidia-smi -pl 180
```

To restore the Tesla V100 back to its default TDP power consumption, you can either unload the driver module and reload, or use the following command:

```
nvidia-smi -pl 250
```

DESIGN DISCUSSION

FORM FACTOR

The Tesla V100 PCIe board conforms to NVIDIA Form Factor 3.0 specification. In this product brief, nominal dimensions are shown.

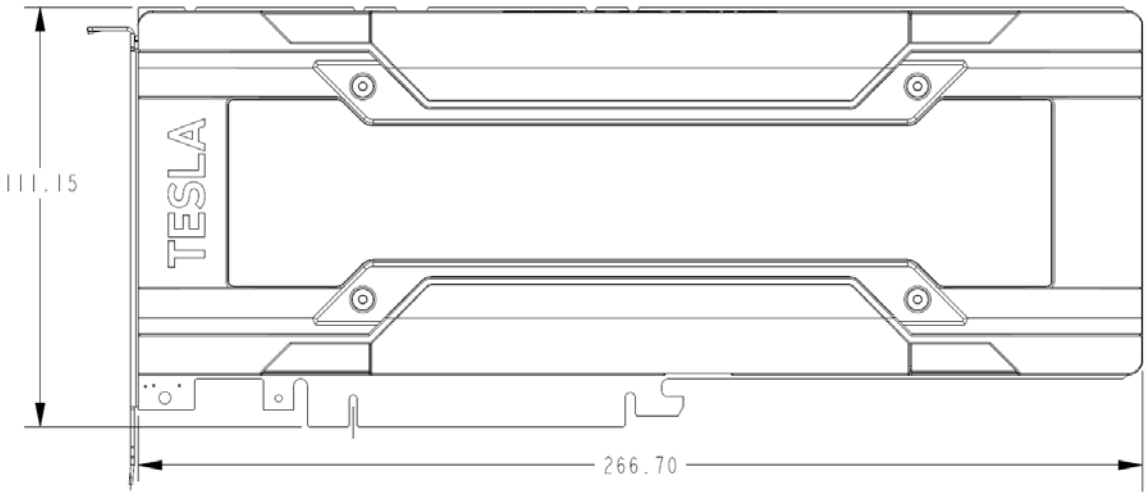

Figure 3. Tesla V100 PCIe Board Dimensions (With Optional I/O Bracket)

POWER CONNECTOR PLACEMENT

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

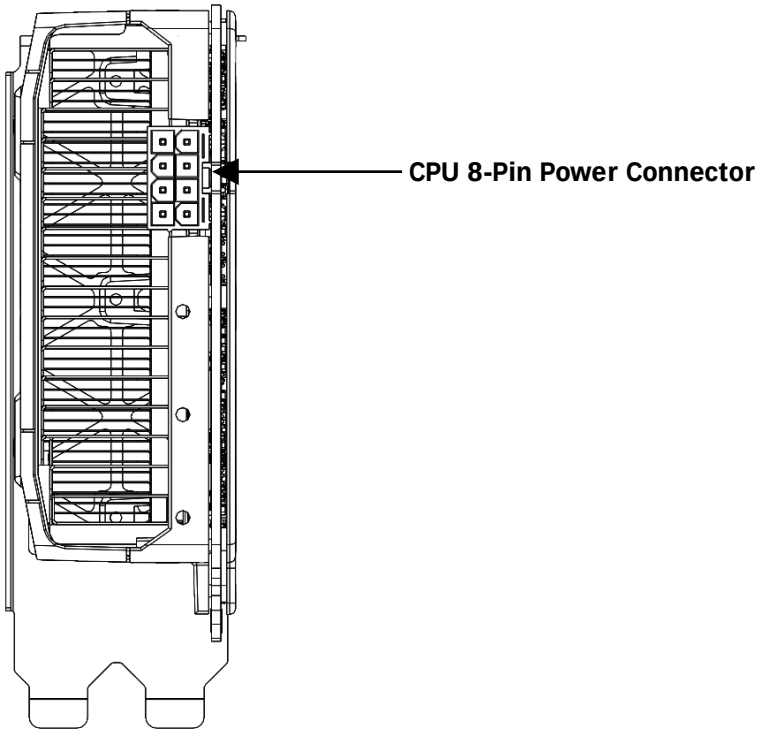

Figure 4. CPU 8-Pin Power Connector (With Optional I/O Bracket)

Table 6. Supported Auxiliary Power Connections

Board Connector	PSU Cable
CPU 8-pin	1x CPU 8-pin cable
CPU to PCIe 8-pin dongle	2x PCIe 8-pin cable 2x PCIe 6-pin cable[1] 1x PCIe 8-pin cable and 1x PCIe 6-pin cable[1]

Notes:

[1]The PCIe 6-pin cable must be capable of carrying up to 120 W.

CPU 8-Pin to PCIe 8-Pin Dongle

Figure 5 lists the pin assignments of the dongle. The part number for the dongle is NVPN: 030-0571-000.

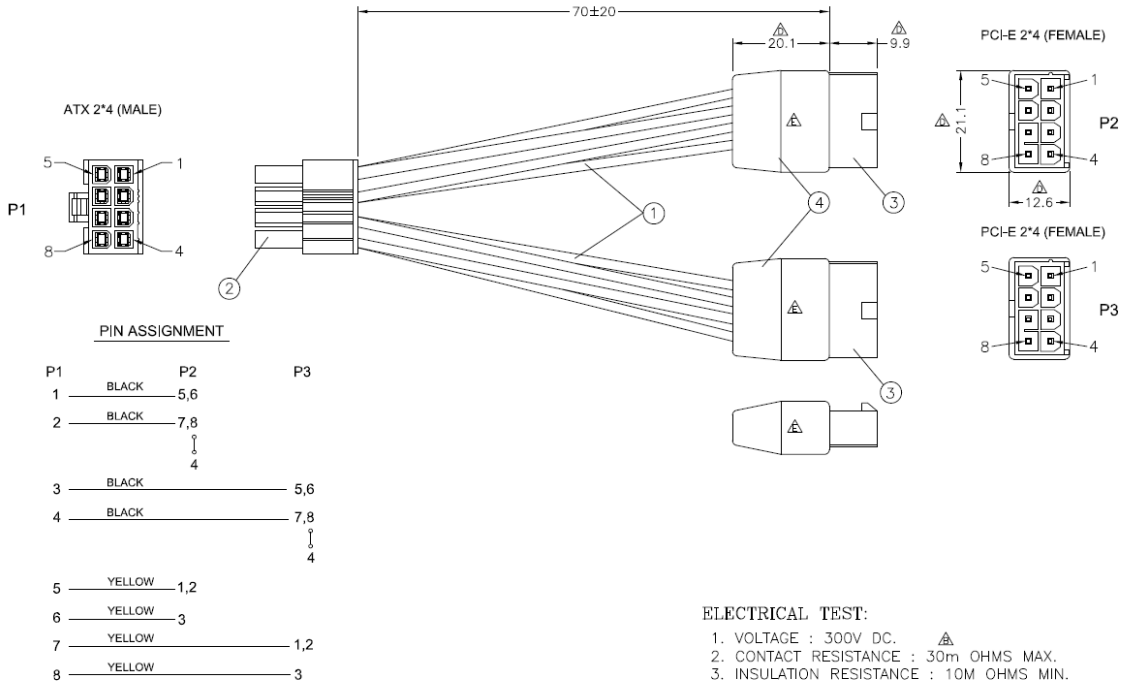

Figure 5. CPU 8-Pin to PCIe 8-Pin Dongle

EXTENDERS

The Tesla V100 PCIe board provides two extender options as shown in the following figures.

- NVPN: 682-00003-5555-002 –Long offset extender (Figure 6)
  - Card + extender = 339 mm
- NVPN: 682-00003-5555-000 – Straight extender (Figure 7)
  - Card + extender = 312 mm

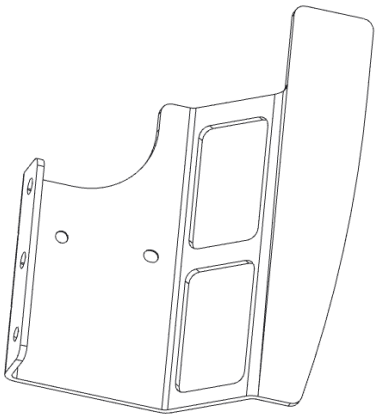

Figure 6. Long Offset Extender

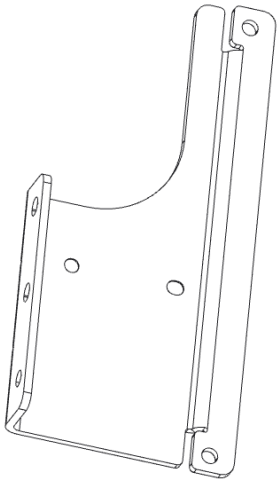

Figure 7. Straight Extender

- 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 heat sink baseplate. The extender mounting holes will vary among NVIDIA products, so designing for flexibility is recommended.

Note: Use this device only with UL Listed ITE Personal Computer (PC)/Server. The device must be installed with the bracket facing the side of the equipment.

SUPPORT INFORMATION

LANGUAGES

Table 7. 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	

Languages	Windows[1]	Linux
Russian	Yes	
Slovak	Yes	
Slovenian	Yes	
Spanish (European)	Yes	
Spanish (Latin America)	Yes	
Swedish	Yes	
Thai	Yes	
Turkish	Yes	

Note:

[1]Windows 7, Windows 8, Windows 8.1, Windows 10, Windows Server 2008 R2, Windows Server 2012 R2, and Windows Server 2016 are supported.