TESLA K80 GPU ACCELERATOR

BD-07317-001_v05 | January 2015

Board Specification

DOCUMENT CHANGE HISTORY

BD-07317-001_v05

Version	Date	Authors	Description of Change
01	June 23, 2014	GG, SM	Preliminary Information (Information contained within this board specification is subject to change)
02	October 8, 2014	GG, SM	•Updated product name •Minor change to Table 2
03	October 31, 2014	GG, SM	•Added “8-Pin CPU Power Connector” section •Updated Figure 2
04	November 14, 2014	GG, SM	•Removed preliminary and NDA •Updated boost clocks •Minor edits throughout document
05	January 30, 2015	GG, SM	Updated Table 2 with MTBF data

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

OVERVIEW

The NVIDIA® Tesla® K80 graphics processing unit (GPU) is a PCI Express, dual-slot computing module in the Tesla (267 mm length) form factor comprising of two Tesla K80 GPUs. The Tesla K80 GPU Accelerator is designed for servers and offers a total of 24 GB of GDDR5 on-board memory (12 GB per GPU) and supports PCI Express Gen3. The Tesla K80 is only available with a passive heat sink, which requires externally generated airflow for cooling.

The Tesla K80 GPU Accelerator boards ship with ECC enabled by default protecting the register files, cache and DRAM. With ECC enabled, some of the memory is used for the ECC bits, so the user available memory is reduced by ~6.25%. On the Tesla K80 the total available memory with ECC turned on will be ~22.5 GB.

The following figure shows the block diagram of the Tesla K80. It has two identical Tesla K80 GPUs, connected via an on-board PLX switch. Both the GPUs have access to 12 GB of GDDR5. The board supports PCI Express Gen3. The board is designed for a maximum input power consumption of 300 W.

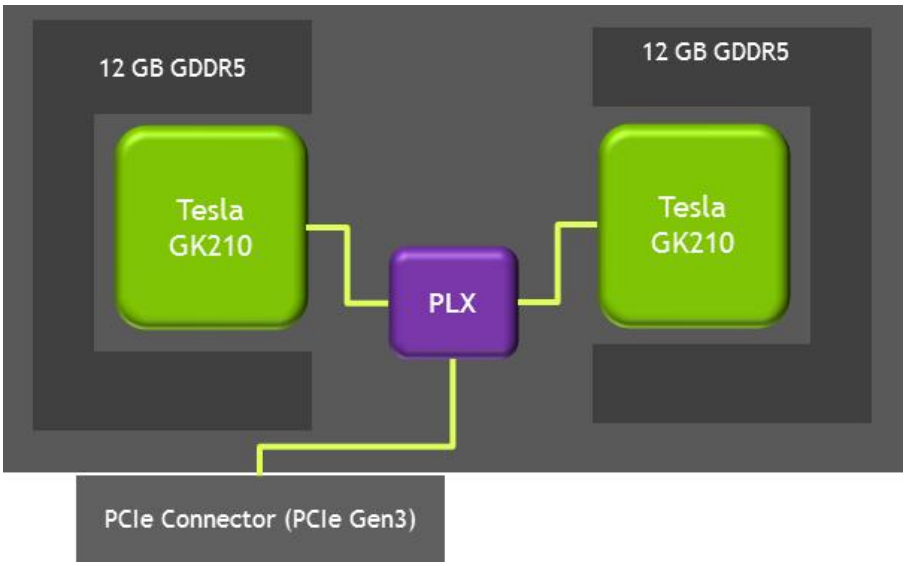

Figure 1. Tesla K80 Block Diagram

KEY FEATURES

GPU

The Tesla K80 Accelerator has two Tesla GK210 GPUs. Characteristics for both GPUs are as follows:

- Number of processor cores: 2496
- Base core clock: 560 MHz
- Boost clocks: 562 MHz to 875 MHz
- Package size: 45 mm × 45 mm 2397-pin ball grid array (S-FCBGA)

Note: All boards ship with core clock set to the base clock value. The GPU clock will start at base clock and will boost automatically until the 300 W power cap limit or thermal limit is reached.

Board

- PCI Express Gen3 ×16 system interface
- Physical dimensions: 111.15 mm (height) × 267 mm (length), dual-slot

Thermal Solution

- Passive heat sink

Display Connectors

- None

Power Connectors

- One 8-pin CPU power connector

Memory

- Memory clock: 2.5 GHz
- Memory bandwidth: 480 GB/sec (cumulative)
- Interface: 384-bit
  - Total board memory: 24 GB
  - 48 pieces of 256M ×16 GDDR5, SDRAM

BIOS

- 2Mbit serial ROM
- BAR1 size: 16 GB per GPU

NVIDIA GPU BOOST ON TESLA K80

The NVIDIA GPU Boost™ feature makes use of any power headroom by raising the core clock to a higher frequency. When an application is being run and the GPU has thermal headroom, the driver will automatically raise the clocks to ensure maximum utilization and performance.

The Tesla K80 ships with Autoboost enabled by default. Autoboost mode means that when the end user starts using the Tesla K80 for the first time, the GPUs will start at base clock and raise the core clock to higher levels automatically as long as the boards stays within the 300 W power limit.

If the end user does not want the Tesla K80 clocks to boost automatically, the end-user can disable this feature and lock the module to a clock supported by the GPU.

Having the boards boost automatically will be useful in scenarios where the workloads have a lot of headroom, as each GPU works independently and is not required to run in lock step with all the GPUs in the cluster.

For more information on NVIDIA GPU Boost and dynamic clock management, refer to the *NVIDIA GPU Boost for Tesla Application Note* (DA-06767-001).

Note: The memory clock remains constant at 2.5 GHz. It's likely that the effective memory bandwidth utilization will change depending on the core clock frequency.

ENVIRONMENTAL CONDITIONS

Table 1 lists the environmental operating and storage conditions for the Tesla K80 board.

Table 1. Board Environmental Conditions

Specifications	Conditions
Operating temperature	0 °C to 45 °C
Storage temperature	-40 °C to 75 °C
Operating humidity	5% to 90% RH
Storage humidity	5% to 95% RH

CONFIGURATION

The Tesla K80 board is available in the following configuration.

Table 2. Board Configuration

Specifications	Tesla GK210
Generic SKU reference	699-22080-0200-xxx
Number of GPUs	2× Tesla GK210B
Core clocks	•Base clock: 560 MHz •Boost clocks: 562 – 875 MHz
Memory clock	2.5 GHz
Memory size/board	•24 GB (per board) •12 GB (per GPU)
Memory I/O	384-bit GDDR5
Memory bandwidth	•480 GB/s (per board) •240GB/s (per GPU)
Memory configuration	48 pieces of 256M × 16 GDDR5 SDRAM
Display connectors	None
Power connectors	8-pin CPU power connector (ships with a 2× 8-pin PCIe to single 8-pin CPU convertor)
Board power	300 W
Power cap level	•150 W per GPU •300 W per board
BAR1 size	16 GB (per GPU)
Extender options	Straight extender or long offset extender
Idle power	TBD
Thermal cooling solution	Passive heat sink
Mean time between failures (MTBF)	Controlled environment: 151377.2164 hours at 35 °C
ASPM	Off

MECHANICAL SPECIFICATIONS

PCI EXPRESS SYSTEM

The Tesla K80 board (Figure 2) conforms to the PCI Express full height form factor.

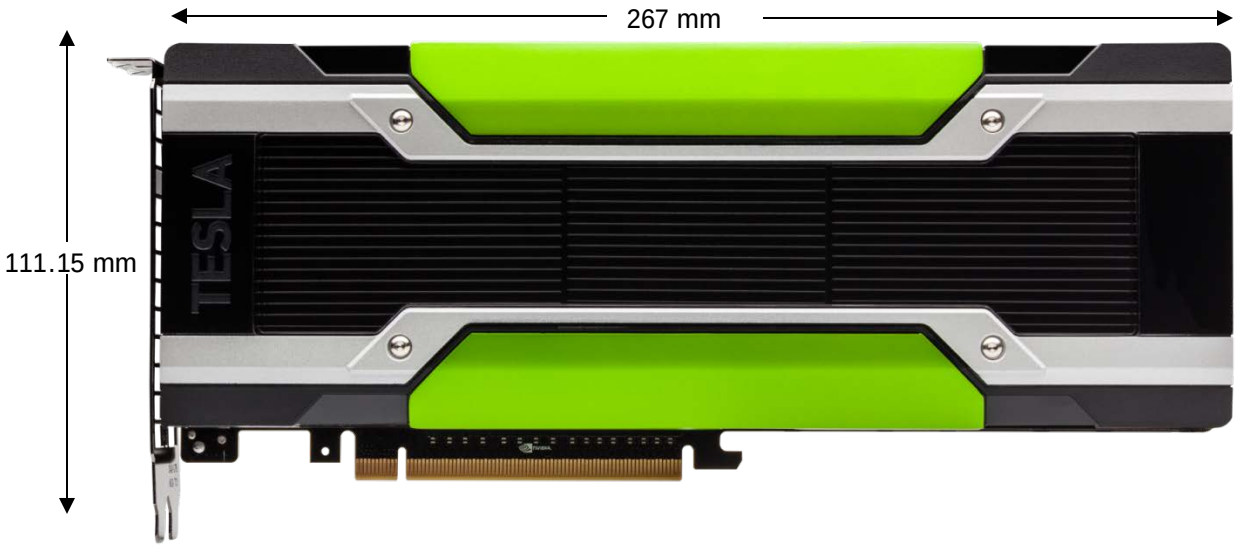

Figure 2. Tesla K80 GPU Accelerator

TESLA K80 BRACKET

As shown in Figure 3, the Tesla K80 includes a vented bracket. If you are an OEM who qualifies for bracket modifications, you have the option of receiving your module with no bracket installed.

Figure 3. Tesla K80 Bracket

8-PIN CPU POWER CONNECTOR

Figure 4 is a diagram of the 8-pin CPU power connector including pin assignments. The 8-pin CPU power connector ships with a 2× 8-pin PCIe to single 8-pin CPU convertor.

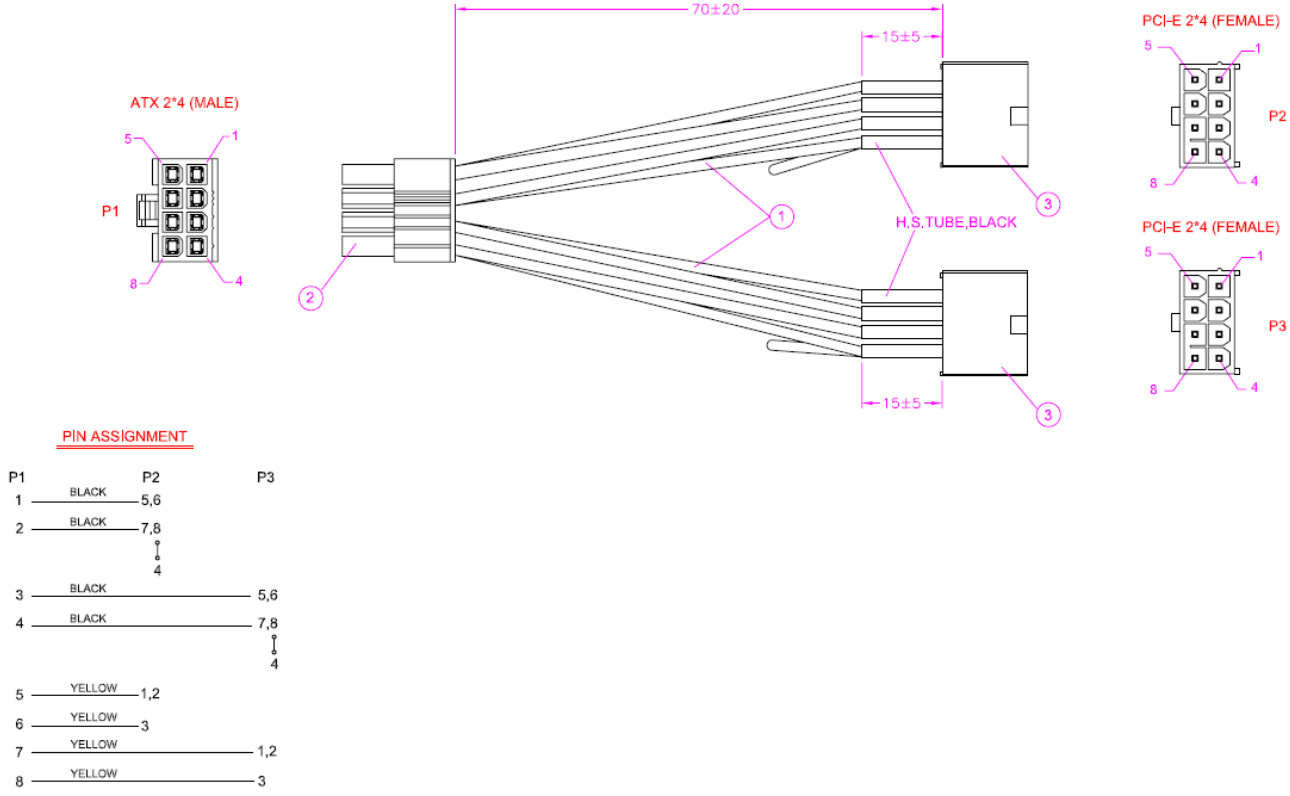

Figure 4. 8-Pin CPU Power Connector

EXTENDERS

A straight extender (NVPN: 320-0867-003) and a long offset extender (NVPN: 320-0866-003) are available for all NVIDIA Form Factor 2.0 compliant boards. These extenders are shown in Figure 5 and Figure 6.

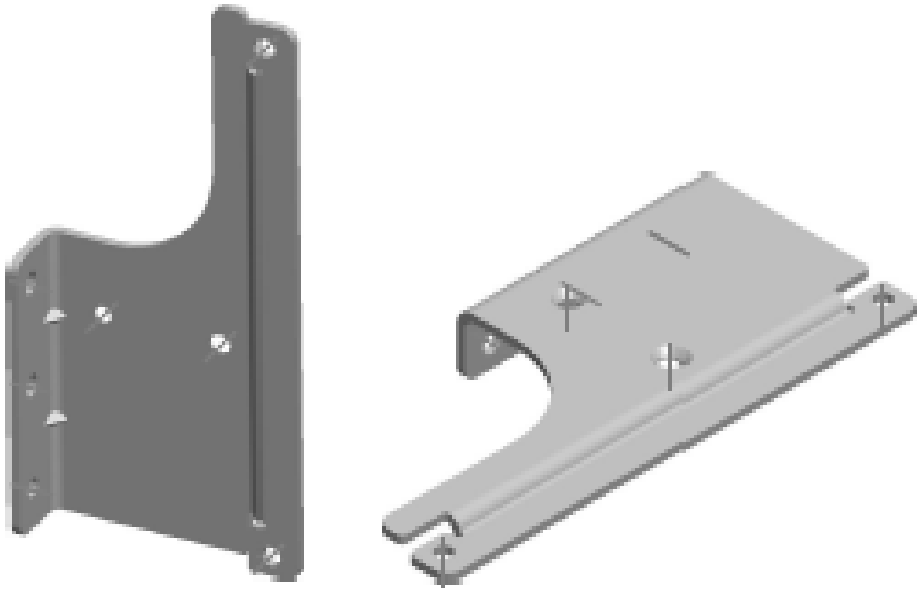

Figure 5. Straight Extender

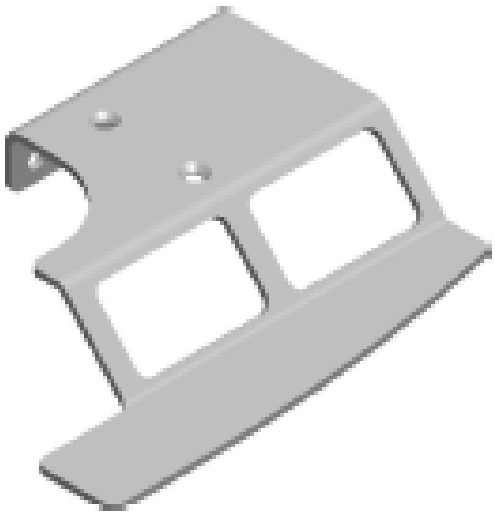

Figure 6. Long Offset Extender

POWER SPECIFICATIONS

The board provides a single EPS12V CPU 8-pin power connector on the "east" edge of the board. The Tesla K80 no longer uses the PCI Express auxiliary connectors.

For backward compatibility with existing systems, NVIDIA will provide a power dongle that converts the CPU 8-pin to two PCI Express 8-pin connectors. The two PCI Express cables must be from a common rail on the system power supply and together must be able to supply sufficient power as specified in Table 3.

Table 3. Power Consumption

Cable Attachments	Support	Comments
8-pin EPS-12V auxiliary power cable attached	Required (unless the power dongle is used)	The CPU 8-pin cable must be able to provide 225 W.
Power dongle: •PCIe 8-pin + PCIe 8-pin cables •PCIe 8-pin + PCIe 6-pin cables •PCIe 6-pin + PCIe 6-pin cables	Required (unless the CPU 8-pin is used)	Any of these PCIe power cable combinations can be used as long as both cables are from a common rail on the power supply and together provide 225 W total.

SUPPORT INFORMATION

CERTIFICATES AND AGENCIES

Agencies

- Australian Communications Authority and Radio Spectrum Management Group of New Zealand (C-Tick)
- 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)
- Voluntary Control Council for Interference (VCCI)

LANGUAGES

Table 4. Languages Supported

	Windows Server 2008 and Windows Server 2008 R2	Linux
English (US)	X	X
English (UK)	X	
Arabic	X	
Chinese, Simplified	X	
Chinese, Traditional	X	
Danish	X	
Dutch	X	
Finnish	X	
French	X	
French (Canada)	X	
German	X	
Italian	X	
Japanese	X	
Korean	X	
Norwegian	x	
Portuguese (Brazil)	X	
Russian	X	
Spanish	X	
Spanish (Latin America)	X	
Swedish	X	
Thai	X	

Note: NVIDIA® CUDA® software is only supported in English (U.S.)