TESLA K40 GPU ACTIVE ACCELERATOR

BD-06949-001_v03 | November 2013

Board Specification

DOCUMENT CHANGE HISTORY

BD-06949-001_v03

Version	Date	Authors	Description of Change
01	July 26, 2013	GG, SM	Preliminary Information (Information contained within this board specification is subject to change)
02	September 19, 2013	GG, SM	•Added new section: "NVIDIA GPU Boost on Tesla K40" •Updated Table 2 •Updated "Energy Star Report" section •Added GPU block diagram
03	November 11, 2013	GG, DV	•Updated Figure 1: Tesla K40 GPU Active Accelerator Board •Added MTBF numbers for Table 2: Board Configuration

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

OVERVIEW

The NVIDIA® Tesla® K40 graphics processing unit (GPU) active accelerator board is a PCI Express, dual-slot full height (4.376 inches by 10.5 inches) form factor computing module comprised of a single GK110B GPU. The Tesla K40 active accelerator is designed for workstations and servers, offers a total of 12 GB of GDDR5 on-board memory, and supports PCI Express Gen3.

Tesla K40 active board ships with ECC enabled by default protecting the register files, cache and DRAM. The Tesla K40 active can be configured by the OEM or by the end user to enable or disable ECC, which can fix single-bit errors and detect double-bit errors. Enabling ECC will cause some of the memory to be used for the ECC bits, so the user available memory is reduced by ~6.25%. On the Tesla K40 Active the total available memory with ECC turned on will be ~11.25 GB.

Figure 1. Tesla K40 GPU Active Accelerator Board

KEY FEATURES

GPU

- Number of processor cores: 2880
- Core clocks
  - Base clock: 745 MHz
  - Boost clocks: 810 MHz and 875 MHz
- Package size: 45 mm × 45 mm 2397-pin flip chip ball grid array (S-FCBGA)

Board

- PCI Express Gen3 ×16 system interface
- Physical dimensions: 4.376 inches × 10.5 inches, dual-slot

Display Connectors

- None

Power Connectors

- One 6-pin PCI Express power connector
- One 8-pin PCI Express power connector

Memory

- Memory clock: 3.0 GHz
- Memory bandwidth: 288 GB/sec
- Interface: 384-bit
  - Total board memory: 12 GB
  - 24 pieces of 256M × 16 GGDR5, SDRAM

BIOS

- 2Mbit serial ROM
- BAR1 size: 256 MB

Note: All boards ship with core clock set to the base clock value. Boost clocks can be selected using NVML or NVSMI. Refer to the NVML/NVSMI documentation for more details.

NVIDIA GPU BOOST ON TESLA K40 ACTIVE

NVIDIA GPU Boost™ is a feature available on Tesla K40 active. It makes use of any power headroom to run the core clock to a higher frequency. Application workloads that have power headroom can run at high GPU clocks to boost application performance.

Note: The memory clock remains constant at 3 GHz. It's likely that the effective memory bandwidth utilization will change depending on the core clock frequency.

NVIDIA GPU Boost for HPC Workloads

NVIDIA GPU Boost for Tesla K40 active is optimized to deliver a robust and deterministic boost behavior for a wide range of HPC workloads.

Tesla K40 active gives full control to end-users to select the core clock frequency that fits their workload the best. The workload may have one or more of the following characteristics.

- Problem set is spread across multiple GPUs and requires periodic synchronization.
- Problem set spread across multiple GPUs and runs independent of each other.
- Workload has “compute spikes.” For example, some portions of the workload are extremely compute intensive pushing the power higher and some portions are moderate.
- Workload is compute intensive through-out without any spikes.
- Workload requires fixed clocks and is sensitive to clocks fluctuating during the execution.
- Workload runs in a cluster where all GPUs need to start, finish, and run at the same clocks.
- Workload or end user requires predictable performance and repeatable results.
- Datacenter is used to run different types of workload at different hours in a day to better manage the power consumption.
- Some boards in a cluster have access to better cooling than others.

By default the Tesla K40 active ships with the core clock set to the base clock. HPC workloads can have one or more characteristics as described. When selecting one of the supported boost clocks a good strategy is to characterize the workload with the available boost clocks. For example, DGEMM/Linpack are extremely demanding on power. Therefore, the “base clock” may be the correct choice when running Linpack. Some workloads in life sciences, manufacturing, CFD, CAD, etc., may have power headroom and can take advantage of one of the boost clocks.

API FOR NVIDIA GPU BOOST ON TESLA

The Tesla K40 active gives full control to end-users to select the core clock frequency via NVML or `nvidia-smi`. NVML is a C-based API for monitoring and managing the various states of Tesla products. It provides a direct access to submit queries and commands via `nvidia-smi`. NVML documentation is available at https://developer.nvidia.com/nvidia-management-library-nvml

Table 1 gives a summary of the `nvidia-smi` commands for using NVIDIA GPU Boost on Tesla.

Table 1. nvidia-smi Commands

Usage	Command
View the clocks the Tesla board supports	nvidia-smi –q –d SUPPORTED_CLOCKS
Set one of the supported clocks	nvidia-smi -ac <MEM clock, Graphics clock>
Make the clock settings persistent across driver unload	nvidia-smi -pm 1
Make the clock settings revert to base clocks after driver unloads (or turn off the persistent mode)	nvidia-smi -pm 0
To view the clock in use, use the command	nvidia-smi -q –d CLOCK
To reset clocks back to the base clock (as specified in the board specification)	nvidia-smi –rac
To allow “non-root” access to change graphics clock	nvidia-smi -acp 0

When using non-default applications clocks, driver persistence mode should be enabled. Persistence mode ensures that the driver stays loaded even when no NVIDIA® CUDA® or X applications are running on the GPU. This maintains current state, including requested applications clocks. If persistence mode is not enabled, and no applications are using the GPU, the driver will unload and any current user settings will revert back to default for the next application. To enable persistence mode run '`sudo nvidia-smi -pm 1`'.

The driver will attempt to maintain requested applications clocks whenever a CUDA context is running on the GPU. However, if no contexts are running the GPU will revert back to idle clocks to save power and will stay there until the next context is created. Thus, if the GPU is not busy, you may see idle current clocks even though requested applications clocks are much higher.

Note: By default changing the application clocks requires root access. If the user does not have root access, the user can request his or her cluster manager to allow non-root control over application clocks. Once changed, this setting will persist for the life of the driver before reverting back to root-only defaults. Persistence mode should always be enabled whenever changing application clocks, or enabling non-root permissions to do so.

TESLA K40 ACTIVE BLOCK DIAGRAM

Figure 1 is the block diagram for the Tesla K40 active dual-slot computing processor module.

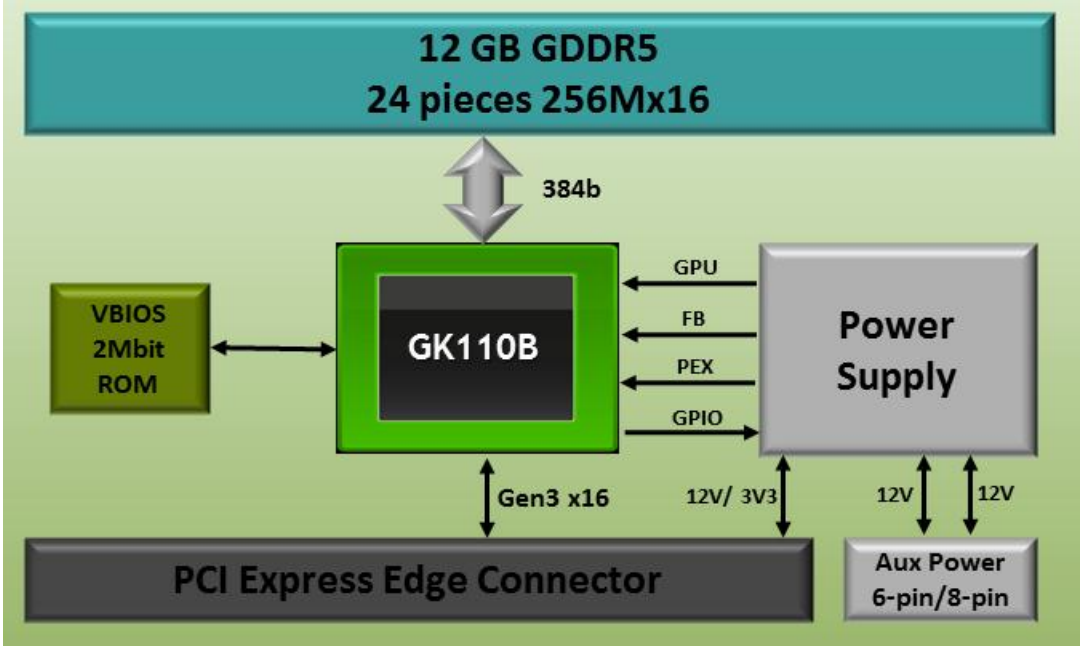

Figure 2. Tesla K40 GPU Active Block Diagram

CONFIGURATION

The Tesla K40 GPU active accelerator board is available in the following configuration.

Table 2. Board Configuration

Specifications	Tesla K40
Generic SKU reference	699-22081-0206-xxx
Chip	GK110B
Package size GPU	45 mm × 45 mm 2397-pin S-FCBGA
Core clocks	•Base clock: 745 MHz •Boost clocks: 810 MHz and 875 MHz
Memory clock	3.0 GHz
Memory size	12 GB
Memory I/O	384-bit GDDR5
Memory configuration	24 pieces of 256M ×16 GDDR5 SDRAM
Display connectors	None
Power connectors	•One 6-pin PCI Express power connector •One 8-pin PCI Express power connector
Board power	235 W
Thermal cooling solution	Active fan sink
Idle power	20.57 W
Form factor	Tesla Form Factor B (10.5 inches in length or 267 mm)
Weight without extender	826 grams
Available extenders	•Long Extender with Offset •Straight Extender
Meantime between failures (MTBF)	GB@ 35C : 282,847 hours GF@ 35C : 252,222 hours
ASPM	Off

MECHANICAL SPECIFICATIONS

PCI EXPRESS SYSTEM

The Tesla K40 GPU active accelerator board (Figure 3) conforms to the PCI Express full height (4.376 inches) and has a board length of 10.5 inches (267 mm). Total board length with the straight extender is 312 mm and 339 mm with the long offset extender.

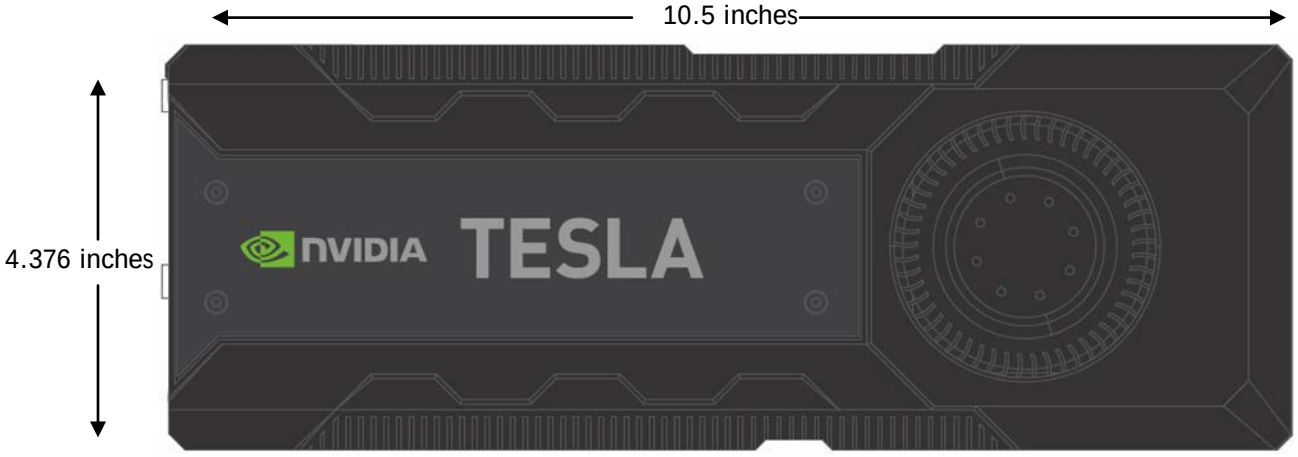

Figure 3. Tesla K40 GPU Active Accelerator

TESLA K40 ACTIVE BRACKET

As shown in Figure 4, the Tesla K40 active includes a vented bracket.

Figure 4. Tesla K40 Active Bracket

POWER CONNECTORS

The Tesla K40 active is a performance optimized, high-end product and uses power from the PCI Express connector as well as external power connectors.

Figure 5 and Figure 6 show the specifications and Table 3 and Table 4 show the pinouts for the 6-pin and 8-pin PCI Express power connectors.

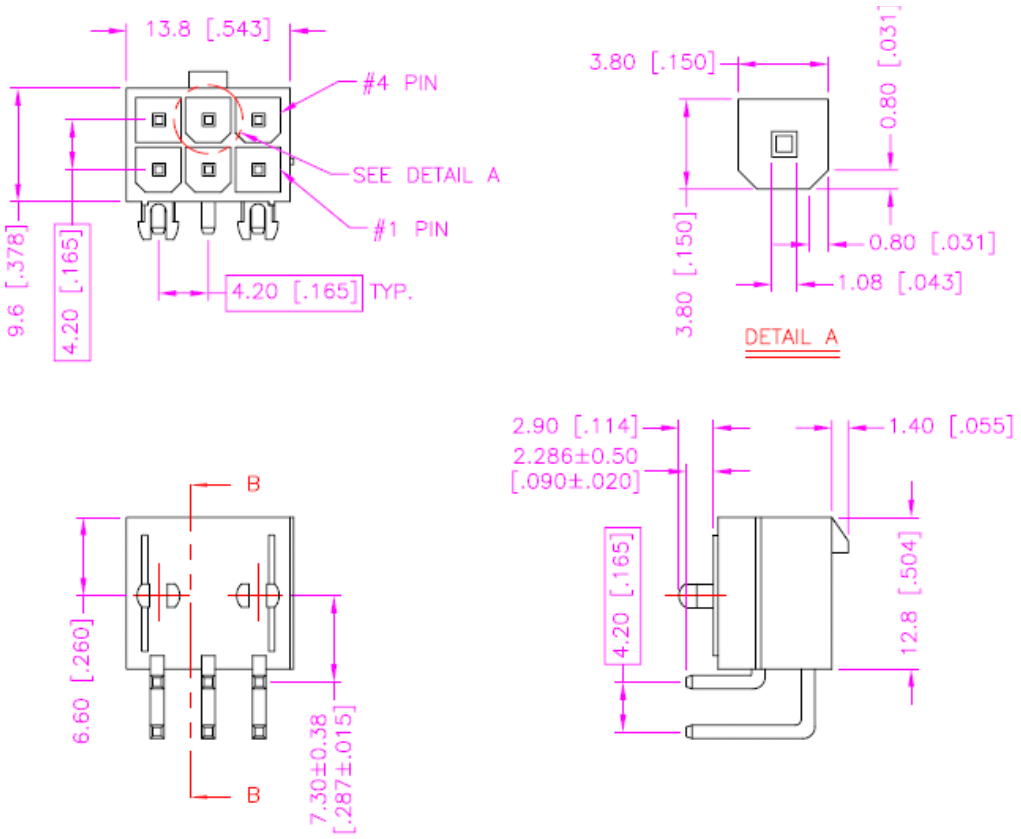

Figure 5. 6-Pin PCI Express Power Connector

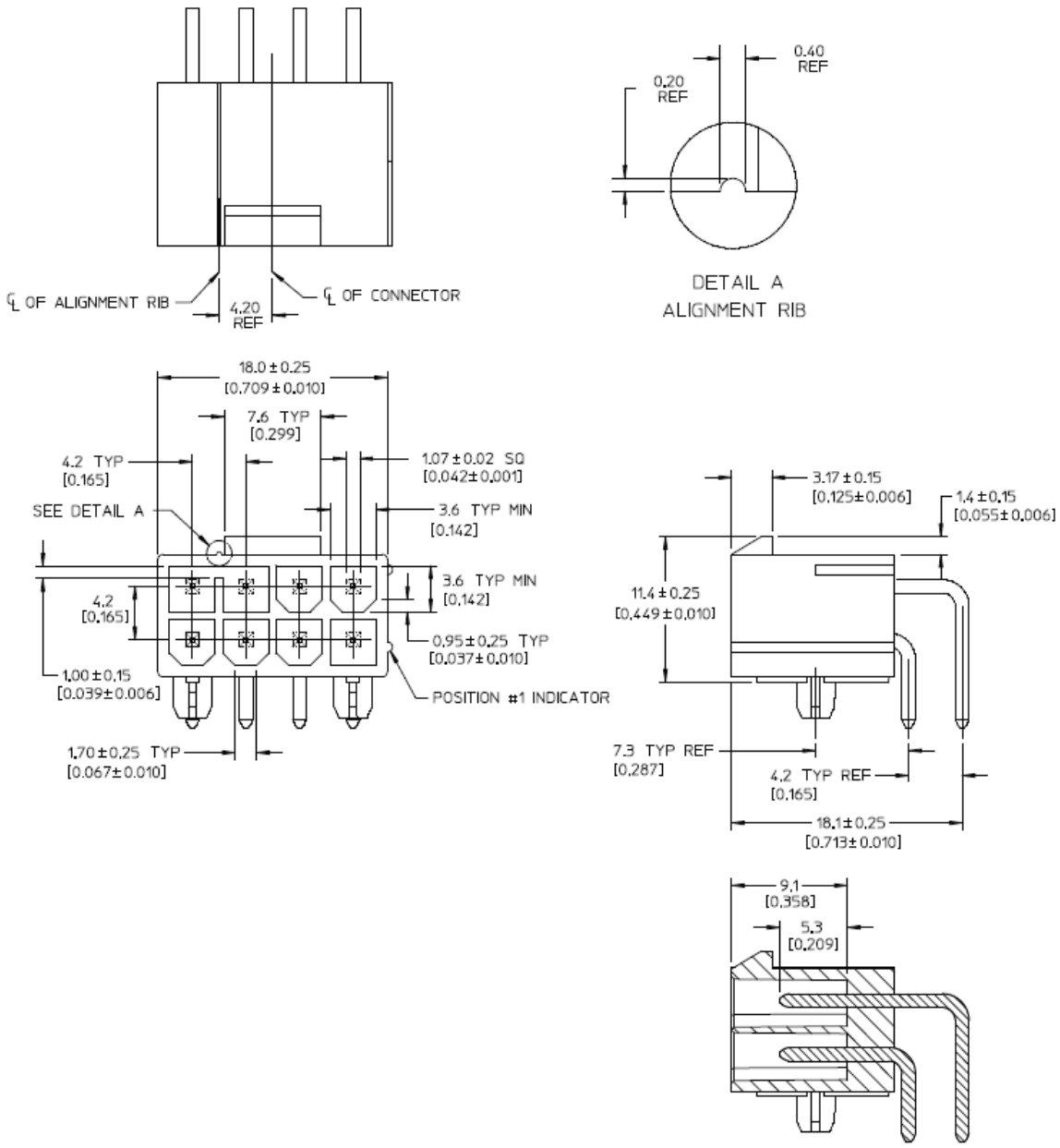

Figure 6. 8-Pin PCI Express Power Connector

Table 3. 6-Pin PCI Express Power Connector Pinout

Pin Number	Description
1	+12 V
2	+12 V
3	+12 V
4	GND
5	Sense
6	GND

Table 4. 8-Pin PCI Express Power Connector Pinout

Pin Number	Description
1	+12 V
2	+12 V
3	+12 V
4	Sense1
5	GND
6	Sense0
7	GND
8	GND

EXTENDERS

The Tesla K40 active board provides three extender options as shown in the following figures.

- NVPN: 320-0866-000 – Long Offset Extender (Figure 7)
  - Card + Extender = 339 mm
- NVPN: 320-0867-000 – Straight Extender (Figure 8)
  - Card + Extender = 312 mm PCIe full length

Figure 7. Long Offset Extender

Figure 8. Straight Extender

- Ensure the system can accommodate standard mounting of “Type B” form factor boards using the NVIDIA extenders.
- 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 Tesla products, so designing for flexibility is recommended.

POWER SPECIFICATIONS

The Tesla K40 GPU active accelerator board requires power from the PCI Express connector as well as one or two auxiliary power connectors.

Table 5. PCI Express Power Connectors

8-Pin Header	6-Pin Header	Support	Notes
Connect 8-pin cable	Connect 6-pin cable	Yes	
Connect 8-pin cable	No cable installed	Yes	8-pin cable must supply 175 W
Connect 6-pin cable	Connect 6-pin cable	No	8-pin connector should always be connected

POWER BY RAIL

Table 6 lists the average power (in watts) by connector for the Tesla K40 active while running DGEMM.

Table 6. Power by Rail

PEX12V	PEX3V3	EXT12V	EXT12V	Total Board Power
55 W	3 W	115 W	62 W	235 W

ENERGY STAR REPORT

Table 7 lists the energy star report for the Tesla K40 active graphics board.

Table 7. Energy Star Report

Specification	Description
Brand	Tesla K40
GPU	GK110B
Board	P2081
SKU	206
Clocks	•Base clock: 745 MHz •Boost clocks: 810 MHz and 875 MHz •Memory clock: 3.0 GHz
Frame buffer	384-bit
Memory	GDDR5 SDRAM
PEX	×16
Idle power: ASPM OFF	20.57 W
Maximum power: ASPM OFF	215 W

THERMAL SPECIFICATIONS

COOLING SOLUTION

NVIDIA has designed an active fan sink (Figure 9) to cool the GPU, memories and power components. For fan and environmental specifications refer to Table 8.

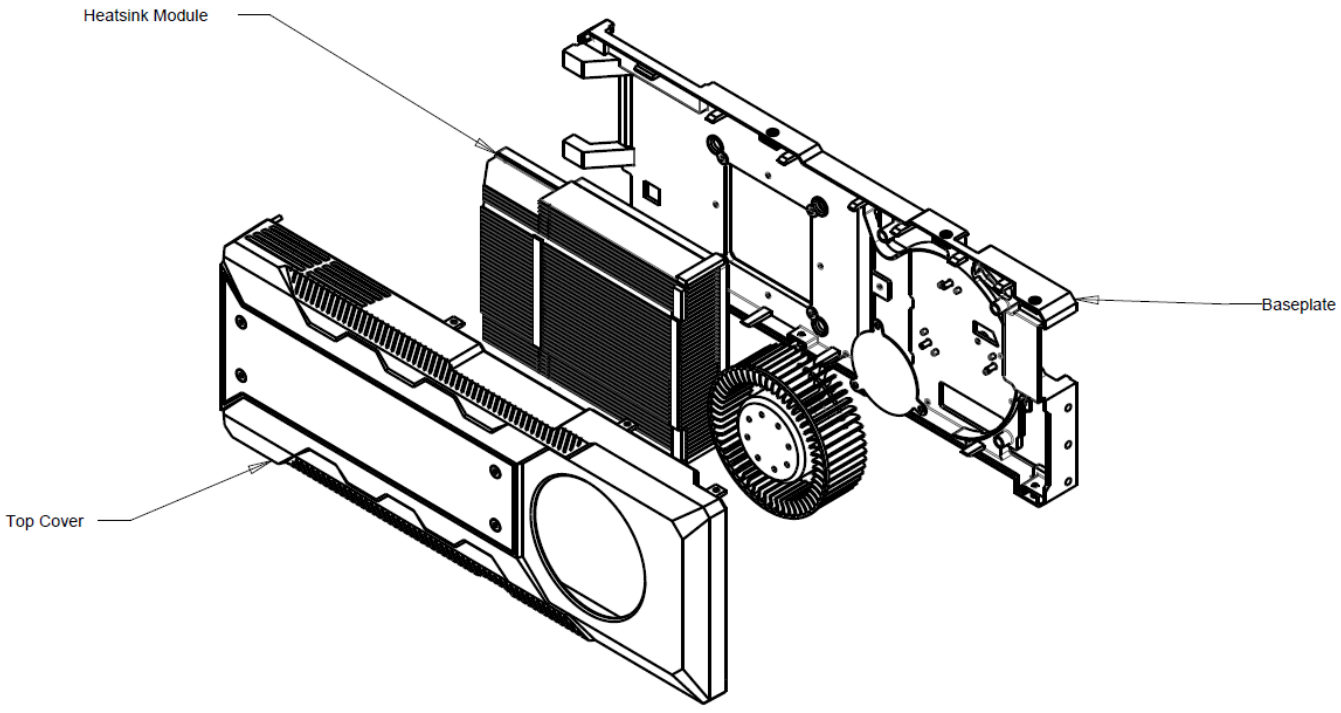

Figure 9. Active Fan Sink

Table 8. Fan Environmental Specifications and Conditions

Specifications	Conditions
Rated voltage	12 V
Operating voltage	10.8 to 12.6 V DC
Rated current	0.86 Amp (maximum 1.80 Amp)
Rated power	10.32 W (maximum 21.60 W)
Speed	5000 RPM
Life expectancy	50,000 hours continuous operation at 55 °C with 15 to 65% RH
Operating temperature	-10 °C to 70 °C
Storage temperature	-40 °C to 75 °C
Operating humidity	5% to 90% RH
Storage humidity	5% to 95% RH

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 9. Languages Supported

	Windows 7 (64-Bit), 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®'s CUDA® software is only supported in English (U.S.)