Overview

Models

NVIDIA Tesla K40 Compute Processor F4A88AA

Introduction

The NVIDIA Tesla™ K40 represents the latest and most advanced Kepler GPU architecture from NVIDIA. Based on the new GK110B GPU, new levels of supercomputing processing capability is now available in a wide range of HP Z-workstations. Based on the massively parallel re-worked CUDA architecture, it features many "must have" HPC capabilities including 12GB ECC memory and an L1/L2 cache design for improved bandwidth and data sharing. The Tesla K40 features innovative technologies like SMX, Hyper-Q, and Dynamic Parallelism.

Features and benefits

2880 CUDA Cores

Delivers up to 1.5 Tflops of double-precision peak performance, enabling a single workstation to deliver ~3 Tflops with two K40's. The K40 single precision peak performance is 3.6 Tflops.

SMX (Streaming multiprocessor)

New CUDA core architecture delivers up to 3x more performance per watt compared to the previous generation Fermi GPU based Tesla cards.

GPU Boost Technology

When enabled, converts power headroom to higher clocks and achieve even greater acceleration for various HPC workloads on Tesla K40.

Dynamic Parallelism

Enables GPU threads to automatically spawn new threads. This allows elimination of unnecessary program control transfers between CPU and GPU and enables GPU acceleration of a broader set of algorithms.

Hyper-Q

This feature enables multiple CPU cores to simultaneously utilize the CUDA cores on a single Kepler GPU. The result is a dramatically increased level of average GPU utilization.

ECC Memory*

Meets a critical requirement for computing accuracy and reliability for workstations. Offers protection of data in memory to enhance data integrity and reliability for applications. Register files, L1/L2 caches, shared memory, and DRAM all are ECC protected.

12GB of GDDR5 Memory

Doubles the size of datasets/models that can be loaded into the Tesla K40 onboard memory. Maximizes performance and reduces data transfers by keeping larger data sets in local memory that is attached directly to the GPU.

API's

Use OpenACC, CUDA, or OpenCL development environments for C, C++, or Fortran to express application parallelism and take advantage of the Kepler GPU's innovative architecture.

*Enabling ECC will cause some of the memory to be used for the ECC bits so the user available memory will decrease by ~6.25%.

Overview

Compatibility

The NVIDIA Tesla K40 Computing Processor is supported on the following HP Personal Workstation:

- Z820 with 1125W power supply, Z620, Z420

The NVIDIA Tesla K40 is supported with 64-bit operating systems only.

Subject to configuration restrictions.

Service and Support

The NVIDIA Tesla K40 has a one-year limited warranty or the remainder of the warranty of the HP product in which it is installed. Technical support is available seven days a week, 24 hours a day by phone, as well as online support forums. Parts and labor are available on-site within the next business day. Telephone support is available for parts diagnosis and installation. Certain restrictions and exclusions apply.

Technical Specifications

Form Factor	Size: 4.376 inches by 10.5 inches Slots: Dual Slot Power Connectors: One 6-pin and one 8-pin Weight: ~826 grams
System Interface	PCI Express Gen3 ×16
Video Outputs	None.
Memory	12GB GDDR5, memory path: 384-bit memory clock: 3Ghz
Peak Memory Bandwidth	288 GB/s
Supported APIs	CUDA, OpenACC, OpenCL 1.2 API support includes: C, C++, Java, Python, and Fortran
Supported Operating Systems	Windows 8 (64-bit) Genuine Windows 7 Professional (64-bit) Red Hat Enterprise Linux (RHEL) 5, 6 Desktop/Workstation (64-bit) SUSE Linux Enterprise Desktop 11 (64-bit) HP qualified drivers may be preloaded or available from the HP support Web site: http://welcome.hp.com/country/us/en/support.html Novell SUSE Linux Enterprise drivers may also be obtained from: ftp://download.nvidia.com/novell or http://www.nvidia.com
Processor Cores	GK110B GPU Base Clock: 745 MHz Boost Clock: up to 875 Mhz 2888 CUDA cores
Power Consumption	~235 Watts **Note 1**: A 1125W PSU is required for any K40 configuration on the Z820
Tesla K40 GPU Boost	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.