NVIDIA® TESLA® M40 GPU ACCELERATOR

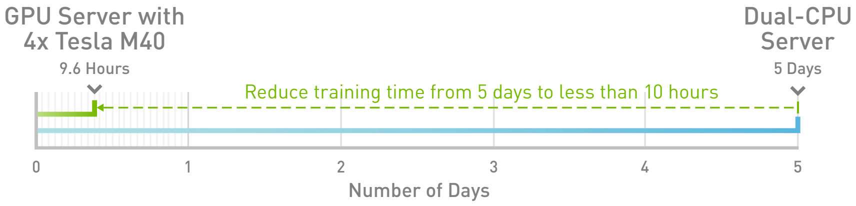
Power your data center with the world's fastest deep learning training accelerator.

Deep learning is redefining what's possible, from image recognition and natural language processing to neural machine translation and image classification. From early-stage startups to large web service providers, deep learning has become the fundamental building block in delivering amazing solutions for end users.

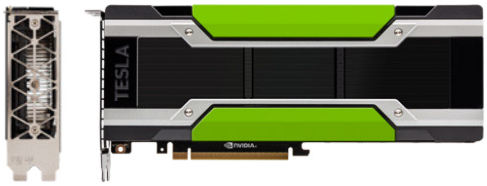
Deep learning models typically take days to weeks to train, forcing scientists to make compromises between accuracy and time to deployment. The NVIDIA Tesla M40 GPU accelerator, based on the ultra-efficient NVIDIA Maxwell™ architecture, is designed to deliver the highest single precision performance. Together with its high memory density, this makes the Tesla M40 the world's fastest accelerator for deep learning training.

Running Caffe and Torch on the Tesla M40 delivers the same model within hours versus days on CPU-based compute systems:

13x FASTER TRAINING



Note: Caffe benchmark with AlexNet, training 1.3M images with 90 epochs | CPU server uses 2x Xeon E5-2699v3 CPU, 128 GB System Memory, Ubuntu 14.04



FEATURES

- NVIDIA GPU Boost™ delivering up to 7 Teraflops of single-precision performance
- 24 GB of GDDR5 memory for training large deep learning models
- Server-qualified to deliver maximum uptime in the data center

SPECIFICATIONS

GPU Architecture	**NVIDIA Maxwell**
NVIDIA CUDA® Cores	**3072**
Single-Precision Performance	**7 Teraflops with NVIDIA GPU Boost**
Double-Precision Performance	**0.2 Teraflops**
GPU Memory	**24 GB GDDR5**
Memory Bandwidth	**288 GB/s**
System Interface	**PCI Express 3.0 x16**
Max Power Consumption	**250 W**
Thermal Solution	**Passive**
Form Factor	**4.4" H × 10.5" L, Dual Slot, Full Height**
Compute APIs	**CUDA, DirectCompute, OpenCL™, OpenACC**

TESLA M40 FEATURES THE LARGEST MEMORY CAPACITY PER GPU

Researchers and developers are building bigger, more sophisticated neural networks to increase detection and prediction accuracy. Training these bigger networks demands more GPU memory, and the M40 is purpose-built to handle these workloads.

This accuracy improves performance in a variety of applications:

- More accurate speech recognition
- More accurate image identifying of objects like street signs, pedestrians, etc.
- Deeper understanding in video and natural language content
- Better detection of anomalies in medical images, improving medical diagnosis

DEEP LEARNING ECOSYSTEM BUILT FOR TESLA PLATFORM

The Tesla M40 accelerator provides a powerful foundation for customers to leverage best-in-class software and solutions for deep learning. NVIDIA cuDNN, DIGITS™ and various deep learning frameworks are optimized for the NVIDIA Maxwell™ architecture and Tesla M40 to power the next generation machine learning applications.

Frameworks

Caffe
Chainer
DL4J
julia
KERAS
MatConvNet
Microsoft CNTK
MINERVA
mxnet
OpenDeep
Purine
Pylearn2
TensorFlow
theano
torch

Deep Learning SDK

NVIDIA cuDNN

cuDNN provides GPU-accelerated deep neural network primitives, low memory overhead, flexible data layouts, and support for:

- 2D and 3D datasets
- Forward and backward convolution routines
- Arbitrary dimension ordering, striding, and sub- regions for 4d tensors means, allowing for easy integration into any neural net implementation
- Tensor transformation functions
- Neuron activations forward and backward (Rectified Linear, Sigmoid, Hyperbolic Tangent)
- Context-based API for easy multithreading
- Automatic best algorithm selection for convolutions
- The latest NVIDIA GPU architectures

NVIDIA DIGITS

DIGITS is an interactive deep neural network development environment that allows data scientists to:

- Design and visualize deep neural networks
- Schedule, monitor, and manage DNN training jobs
- Manage GPU resources, allowing users to train multiple models in parallel
- Visualize accuracy and loss in real time while training
- Track datasets, results, and trained neural networks
- Automatically scale training jobs across multiple GPUs

GPUltima

A Petaflop-in-a-Rack Networked GPU Cluster, the GPUltima has 10 times more cores, 90% less power and 95% less space* than other petaflop compute solutions. OSScan provide subsets of the GPUltima depending on customer needs.

*Versus traditional 1 petaflop clusters; based on HPC 500 listing/data.

ONE STOP SYSTEMS

NVIDIA SOLUTION PROVIDER

One Stop Systems

One Stop Systems (OSS) produces high-density, GPU-accelerated appliances for a variety of performance-intensive applications in the HPC market. A leader in PCIe expansion, OSS provides scalable clusters of petaflop compute performance in a single rack.

www.onestopsystems.com | +1 (877) 438-2724 | sales@onestopsystems.com