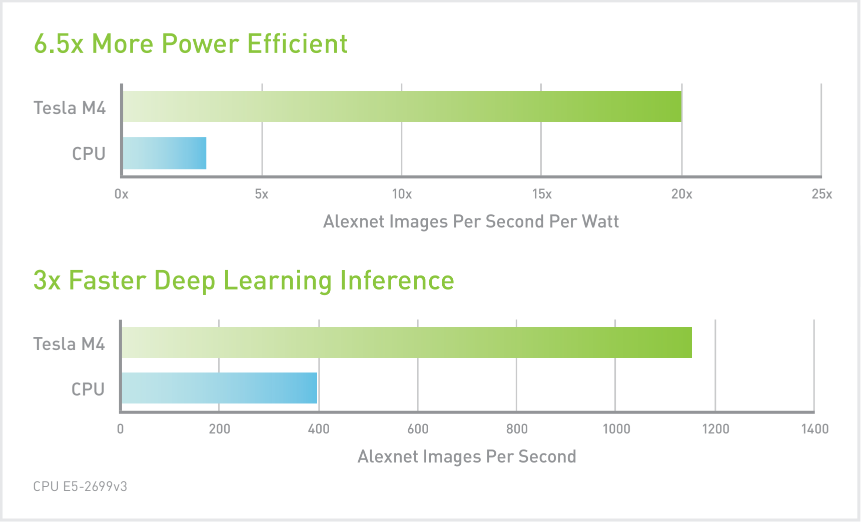
NVIDIA® TESLA® M4 GPU ACCELERATOR

The World's First Accelerator for the Hyperscale Data Center

Exploding volumes of user-generated data are redefining what's required for hyperscale data centers. Today's cloud applications harness valuable data to deliver smarter, real-time experiences using modern video and image processing and deep learning techniques. These applications can benefit greatly from GPU acceleration in the data center.

The NVIDIA Tesla M4 is the world's first accelerator designed for hyperscale servers, enabling customers to keep up with ever-growing amount of data. It's engineered to accelerate application throughput in a small, low-power design, slashing data center costs by half and deliver up to 7x more power-efficient processing than CPUs for deep learning inference at 20 images/sec/watt and video workloads.

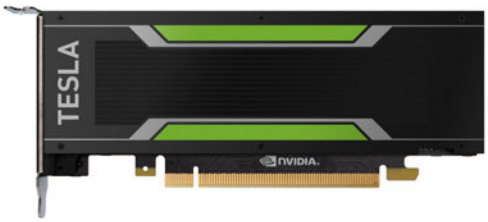
Hyperscale Application Advantage:



6.5x More Power Efficient

3x Faster Deep Learning Inference

CPU E5-2699v3



FEATURES

- NVIDIA GPU Boost™, which delivers up to 2.2 Teraflops of single-precision performance
- Small, low-power design for hyperscale servers
- Server qualification to deliver maximum uptime in the data center

SPECIFICATIONS

GPU Architecture	**NVIDIA Maxwell™**
NVIDIA CUDA® Cores	**1024**
Single-Precision Performance	**2.2 Teraflops with NVIDA GPU Boost**
Double-Precision Performance	**.07 Teraflops with NVIDA GPU Boost**
GPU Memory	**4 GB GDDR5**
Memory Bandwidth	**88 GB/s**
System Interface	**PCIe Gen3**
Max Power Consumption	**50W-75W**
Thermal Solution	**Passive**
Form Factor	**Low Profile**
Compute APIs	**NVIDIA 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 | K 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

NVIDIA.