



NVIDIA® TESLA® P40 INFERENCE ACCELERATOR

EXPERIENCE MAXIMUM INFERENCE THROUGHPUT

In the new era of AI and intelligent machines, deep learning is shaping our world like no other computing model in history. GPUs powered by the revolutionary NVIDIA Pascal™ architecture provide the computational engine for the new era of artificial intelligence, enabling amazing user experiences by accelerating deep learning applications at scale.

The NVIDIA Tesla P40 is purpose-built to deliver maximum throughput for deep learning deployment. With 47 TOPS (Tera-Operations Per Second) of inference performance and INT8 operations per GPU, a single server with 8 Tesla P40s delivers the performance of over 140 CPU servers.

As models increase in accuracy and complexity, CPUs are no longer capable of delivering interactive user experience. The Tesla P40 delivers over 30X lower latency than a CPU for real-time responsiveness in even the most complex models.



FEATURES

The world's fastest processor for inference workloads

47 TOPS of INT8 for maximum inference throughput and responsiveness

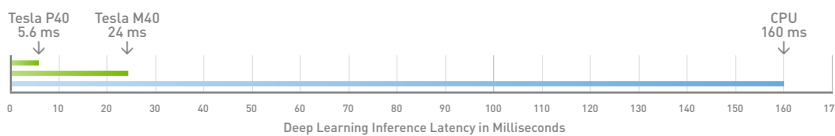
Hardware-decode engine capable of transcoding and inferencing 35 HD video streams in real time

SPECIFICATIONS

GPU Architecture	NVIDIA Pascal™
Single-Precision Performance	12 TeraFLOPS*
Integer Operations (INT8)	47 TOPS* (Tera-Operations per Second)
GPU Memory	24 GB
Memory Bandwidth	346 GB/s
System Interface	PCI Express 3.0 x16
Form Factor	4.4" H x 10.5" L, Dual Slot, Full Height
Max Power	250 W
Enhanced Programmability with Page Migration Engine	Yes
ECC Protection	Yes
Server-Optimized for Data Center Deployment	Yes
Hardware-Accelerated Video Engine	1x Decode Engine, 2x Encode Engine

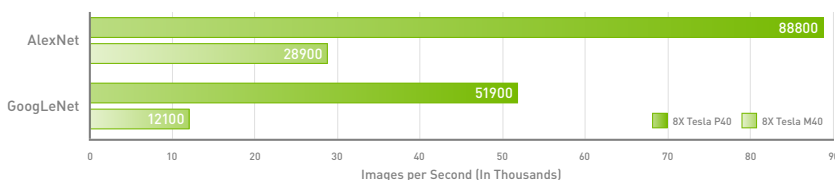
* With Boost Clock Enabled

Reduce Application Latency by Over 30X



CPU: 22-Core Intel Xeon E5-2699V4, MKL2017 IntelCaffe+VG619, Batch Size: 4 | GPU Tesla M4 (TensorRT + FP32) and P4 (TensorRT + Int8), nvCaffe + VG619, Bbatch Size: 4

Achieve Over 4X the Inference Throughput



Note: GPU: Tesla M40 (TensorRT + FP32) and P40 (TensorRT + Int8), nvCaffe GoogLeNet AlexNet batch size =128

NVIDIA TESLA P40 ACCELERATOR FEATURES AND BENEFITS

The Tesla P40 is purpose-built to deliver maximum throughput for deep learning workloads.



140X HIGHER THROUGHPUT TO KEEP UP WITH EXPLODING DATA

The Tesla P40 is powered by the new Pascal architecture and delivers over 47 TOPS of deep learning inference performance. A single server with 8 Tesla P40s can replace up to 140 CPU-only servers for deep learning workloads, resulting in substantially higher throughput with lower acquisition cost.



REAL-TIME INFERENCE

The Tesla P40 delivers up to 30X faster inference performance with INT8 operations for real-time responsiveness for even the most complex deep learning models.



SIMPLIFIED OPERATIONS WITH A SINGLE TRAINING AND INFERENCE PLATFORM

Today, deep learning models are trained on GPU servers but deployed in CPU servers for inference. The Tesla P40 offers a drastically simplified workflow, so organizations can use the same servers to iterate and deploy.



FASTER DEPLOYMENT WITH NVIDIA DEEP LEARNING SDK

TensorRT included with NVIDIA Deep Learning SDK and Deep Stream SDK help customers seamlessly leverage inference capabilities like the new INT8 operations and video trans-coding.

To learn more about the NVIDIA Tesla P40, visit www.nvidia.com/tesla.

© 2016 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, Tesla, NVIDIA GPU Boost, CUDA, and NVIDIA Pascal are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. OpenCL is a trademark of Apple Inc. used under license to the Khronos Group Inc. All other trademarks and copyrights are the property of their respective owners. SEP16

