NVIDIA® TESLA® P4 INFERENCING ACCELERATOR

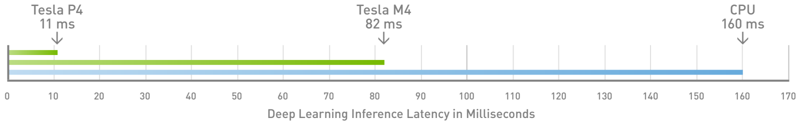
ULTRA-EFFICIENT DEEP LEARNING IN SCALE-OUT SERVERS

In the new era of AI and intelligent machines, deep learning is shaping our world like no other computing model in history. Interactive speech, visual search, and video recommendations are a few of many AI-based services that we use every day.

Accuracy and responsiveness are key to user adoption for these services. As deep learning models increase in accuracy and complexity, CPUs are no longer capable of delivering a responsive user experience.

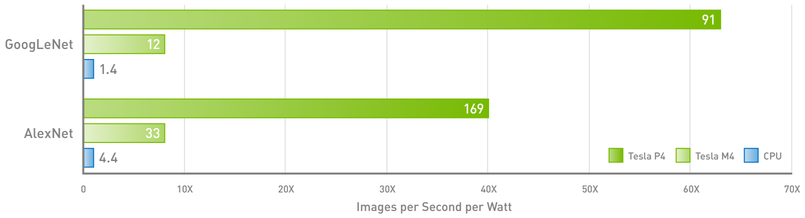
The NVIDIA Tesla P4 is powered by the revolutionary NVIDIA Pascal™ architecture and purpose-built to boost efficiency for scale-out servers running deep learning workloads, enabling smart responsive AI-based services. It slashes inference latency by 15X in any hyperscale infrastructure and provides an incredible 60X better energy efficiency than CPUs. This unlocks a new wave of AI services previous impossible due to latency limitations.

Reduce Application Latency by Over 15X



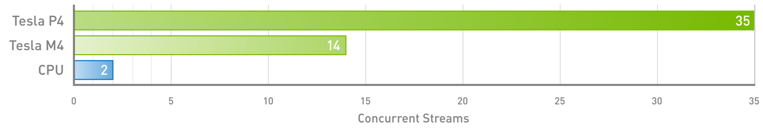
CPU: 22-Core Intel Xeon E5-2699V4, MKL2017 IntelCaffe+VGG19, Batch Size: 4 | GPU Tesla M4 (TensorRT + FP32) and P4 (TensorRT + Int8) , nvCaffe + VGG19, Batch Size: 4

Achieve Over 60X the Inference Efficiency

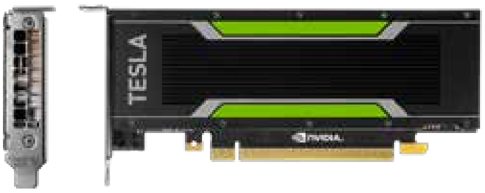


CPU: Intel Xeon E5-2690V4 MKL2017 IntelCaffe+GoogLeNet and AlexNet, Batch Size: 128 | GPU: Tesla M4 (TensorRT + FP32) and P4 (TensorRT + Int 8), nvCaffe GoogLeNet AlexNet, Batch Size: 128

Video Transcode and Inference on H.264 Streams



Note: Dual CPU Xeon E5-2650V4 | Tesla GPU M4 and P4 | Ubuntu 14.04. H.264 benchmark with FFMPEG slow preset | HD = 720p at 30 frames per second.



FEATURES

Small form-factor, 50/75-Watt design fits any scale-out server.

INT8 operations slash latency by 15X.

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

SPECIFICATIONS

GPU Architecture	**NVIDIA Pascal™**
Single-Precision Performance	**5.5 TeraFLOPS***
Integer Operations (INT8)	**22 TOPS* (Tera-Operations per Second)**
GPU Memory	**8 GB**
Memory Bandwidth	**192 GB/s**
System Interface	**Low-Profile PCI Express Form Factor**
Max Power	**50W/75W**
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

NVIDIA TESLA P4 ACCELERATOR FEATURES AND BENEFITS

The Tesla P4 is engineered to deliver real-time inference performance and enable smart user experiences in scale-out servers.

RESPONSIVE EXPERIENCE WITH REAL-TIME INFERENCE

Responsiveness is key to user engagement for services such as interactive speech, visual search, and video recommendations. As models increase in accuracy and complexity, CPUs are no longer capable of delivering a responsive user experience. The Tesla P4 delivers 22 TOPs of inference performance with INT8 operations to slash latency by 15X.

UNLOCK NEW AI-BASED VIDEO SERVICES WITH A DEDICATED DECODE ENGINE

Tesla P4 can transcode and infer up to 35 HD video streams in real-time, powered by a dedicated hardware-accelerated decode engine that works in parallel with the GPU doing inference. By integrating deep learning into the video pipeline, customers can offer smart, innovative video services to users which were previously impossible to do.

UNPRECEDENTED EFFICIENCY FOR LOW-POWER SCALE-OUT SERVERS

The Tesla P4's small form factor and 50W/75W power footprint design accelerates density-optimized, scale-out servers. It also provides an incredible 60X better energy efficiency than CPUs for deep learning inference workloads, letting hyperscale customers meet the exponential growth in demand for AI applications.

FASTER DEPLOYMENT WITH TensorRT AND DEEPSTREAM SDK

TensorRT is a library created for optimizing deep learning models for production deployment. It takes trained neural nets—usually in 32-bit or 16-bit data—and optimizes them for reduced precision INT8 operations. NVIDIA DeepStream SDK taps into the power of Pascal GPUs to simultaneously decode and analyze video streams.

To learn more about the NVIDIA Tesla P4, visit **www.nvidia.com/tesla.**