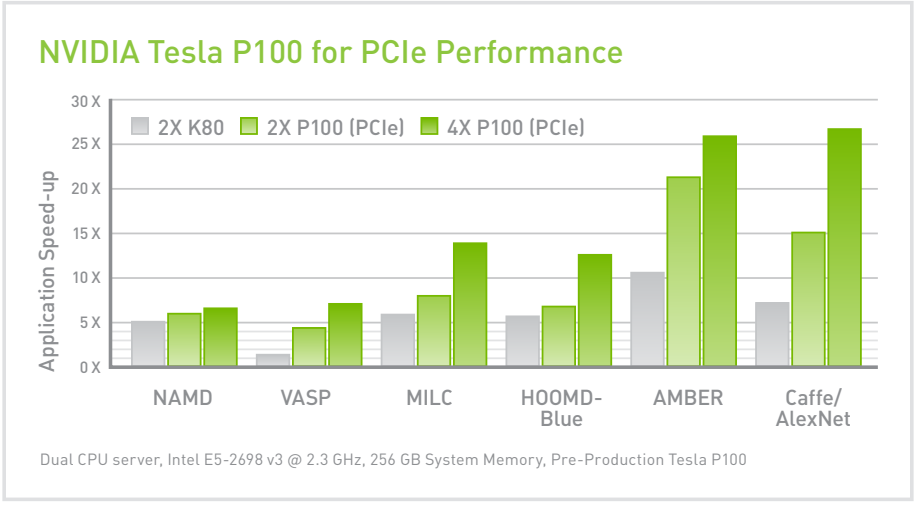
NVIDIA® TESLA® P100 GPU ACCELERATOR

World's most advanced data center accelerator for PCIe-based servers

HPC data centers need to support the ever-growing demands of scientists and researchers while staying within a tight budget. The old approach of deploying lots of commodity compute nodes requires huge interconnect overhead that substantially increases costs without proportionally increasing performance.

NVIDIA Tesla P100 GPU accelerators are the most advanced ever built, powered by the breakthrough NVIDIA Pascal™ architecture and designed to boost throughput and save money for HPC and hyperscale data centers. The newest addition to this family, Tesla P100 for PCIe enables a single node to replace half a rack of commodity CPU nodes by delivering lightning-fast performance in a broad range of HPC applications.

MASSIVE LEAP IN PERFORMANCE



Dual CPU server, Intel E5-2698 v3 @ 2.3 GHz, 256 GB System Memory, Pre-Production Tesla P100

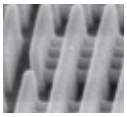
SPECIFICATIONS

GPU Architecture	**NVIDIA Pascal**
NVIDIA CUDA® Cores	**3584**
Double-Precision Performance	**4.7 TeraFLOPS**
Single-Precision Performance	**9.3 TeraFLOPS**
Half-Precision Performance	**18.7 TeraFLOPS**
GPU Memory	**16GB CoWoS HBM2 at 732 GB/s or 12GB CoWoS HBM2 at 549 GB/s**
System Interface	**PCIe Gen3**
Max Power Consumption	**250 W**
ECC	**Yes**
Thermal Solution	**Passive**
Form Factor	**PCIe Full Height/Length**
Compute APIs	**CUDA, DirectCompute, OpenCL™, OpenACC**

TeraFLOPS measurements with NVIDIA GPU Boost™ technology

A GIANT LEAP IN PERFORMANCE

Tesla P100 for PCIe is reimagined from silicon to software, crafted with innovation at every level. Each groundbreaking technology delivers a dramatic jump in performance to substantially boost the data center throughput.

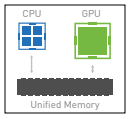


PASCAL ARCHITECTURE

More than 18.7 TeraFLOPS of FP16, 4.7 TeraFLOPS of double-precision, and 9.3 TeraFLOPS of single-precision performance powers new possibilities in deep learning and HPC workloads.

COWOS HBM2

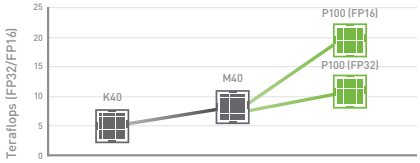
Compute and data are integrated on the same package using Chip-on-Wafer-on-Substrate with HBM2 technology for 3X memory performance over the previous-generation architecture.



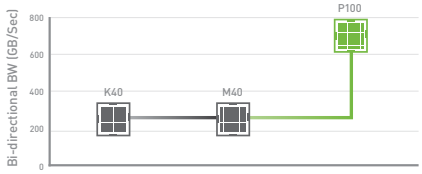
PAGE MIGRATION ENGINE

Simpler programming and computing performance tuning means that applications can now scale beyond the GPU's physical memory size to virtually limitless levels.

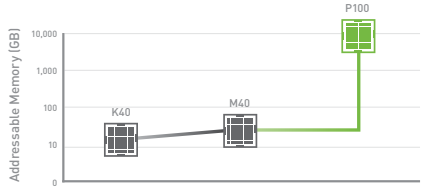
Exponential HPC and hyperscale performance



3X memory boost



Virtually limitless memory scalability



To learn more about the Tesla P100 for PCIe visit **www.nvidia.com/tesla**

NVIDIA

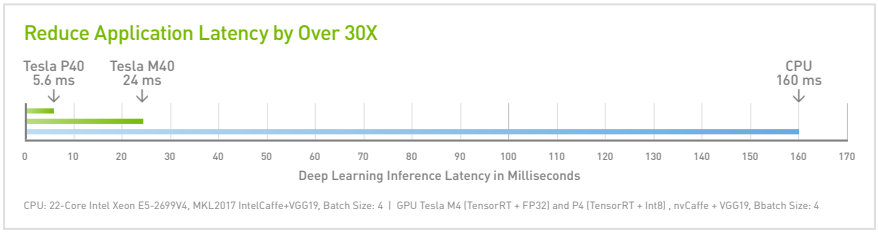
NVIDIA® TESLA® P40 INFERENCING ACCELERATOR

EXPERIENCE MAXIMUM INFERENCE THROUGHPUT

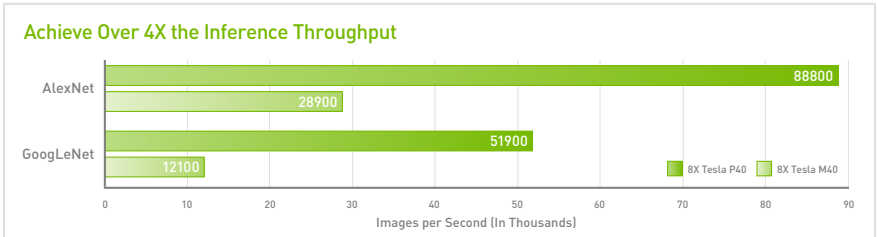
In the new era of AI and intelligent machines, deep learning is shaping our world like no other computing model in history. GPUs powered by the revolutionary NVIDIA Pascal™ architecture provide the computational engine for the new era of artificial intelligence, enabling amazing user experiences by accelerating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



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



Note: GPU: Tesla M40 (TensorRT + FP32) and P40 (TensorRT + Int 8), nvCaffe GoogLeNet AlexNet batch size =128

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

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.**

NVIDIA

NVIDIA® TESLA® P4 INFERENCING ACCELERATOR

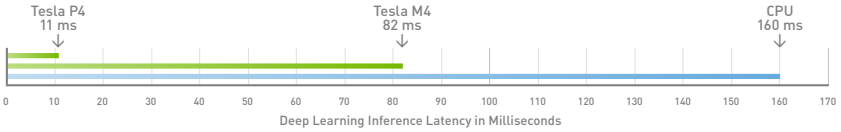
ULTRA-EFFICIENT DEEP LEARNING IN SCALE-OUT SERVERS

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.

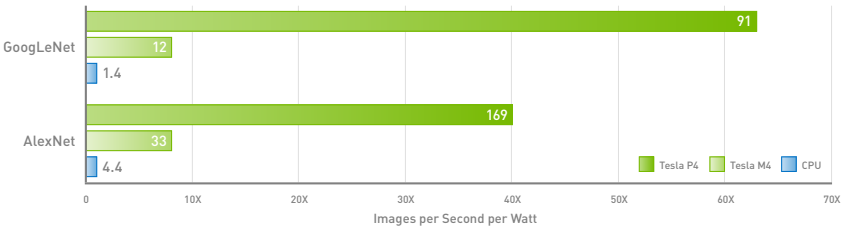
The NVIDIA Tesla P4 is powered by the revolutionary NVIDIA Pascal™ architecture and purpose-built to boost efficiency for scale-out servers running deep learning workloads, enabling smart responsive AI-based services. It slashes inference latency by 15X in any hyperscale infrastructure and provides an 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



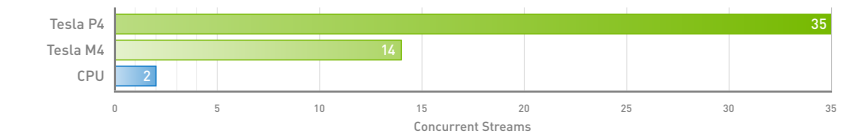
CPU: 22-Core Intel Xeon E5-2699V4, MKL2017 IntelCaffe+VGG19, Batch Size: 4 | GPU Tesla M4 (TensorRT + FP32) and P4 (TensorRT + Int8) , nvCaffe + VGG19, Batch Size: 4

Achieve Over 60X the Inference Efficiency

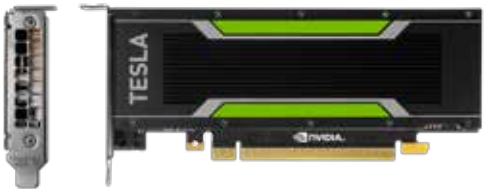


CPU: Intel Xeon E5-2690V4 MKL2017 IntelCaffe+GoogLeNet and AlexNet, Batch Size: 128 | GPU: Tesla M4 (TensorRT + FP32) and P4 (TensorRT + Int 8), nvCaffe GoogLeNet AlexNet, Batch Size: 128

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.**

NVIDIA