PART NUMBER:
TCSP40M-24GB-PB

NVIDIA TESLA P40 by PNY
INFERENCING ACCELERATOR

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.

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.

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.

SIMPLIFIED OPERATIONS WITH A SINGLE TRAINING 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.

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.

TESLA P40 - PRODUCT SPECIFICATION

MEMORY SIZE (PER BOARD)	24 GB GDDR5 (8 GB per board)	
MEMORY INTERFACE	384-bit	
MEMORY BANDWIDTH	346 Gb/s	
CUDA CORES	3840	
PEAK SINGLE PRECISION FLOATING POINT PERFORMANCE	~ 12 Tflops (GPU Boost Clocks)	
INTEGER OPERATIONS (INT8)	47 TOPS (Tera-Operations per Second, boost clocks)	
MEMORY INTERFACE	PCI Express 3.0 x16	
MAX POWER CONSUMPTION	250 W passiv	
THERMAL SOLUTION	passive heatsink	
FORM FACTOR	111,15 mm (H) x 267,7 mm (L) Dual Slot, Full Height	
DISPLAY CONECTORS	None	
POWER CONNECTORS	8-pin CPU power connector	
WEIGHT (W/O EXTENDER)	968g	
PACKAGE CONTENT	1x Power adapter (2 x PCIe 8-pit auf single CPU 8-pin)	
PART NUMBER UND EAN	TCSP40M-24GB-PB	3536403352875

PNY Technologies Europe
sales@pny.eu | T: +33 (0)5 56 13 75 75