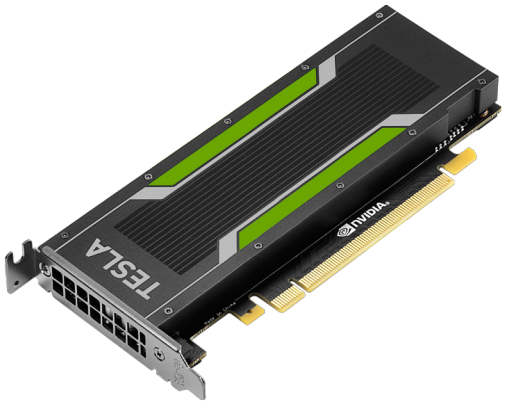
PART NUMBER:
TCSP4M-PB

NVIDIA TESLA P4 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. 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.

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

UNLOCK NEW AI-BASED VIDEO SERVICES

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.

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.

UNPRECEDENTED EFFICIENCY FOR LOWPOWER SCALE-OUT SERVERS

The Tesla P4's small form factor and 50W/75W power footprint design accelerates densityoptimized, 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.

TESLA P4 - PRODUCT SPECIFICATION

MEMORY SIZE (PER BOARD)	8 GB GDDR5 (8 GB per board)	
MEMORY INTERFACE	256-bit	
MEMORY BANDWIDTH	192 Gb/s	
CUDA CORES	2560	
PEAK SINGLE PRECISION FLOATING POINT PERFORMANCE	~ 5.5 Tflops (GPU Boost Clocks)	
MEMORY INTERFACE	PCI Express 3.0 x16	
MAX POWER CONSUMPTION	50W/75W	
THERMAL SOLUTION	passive heatsink	
FORM FACTOR	Low profile PCI Express Form Factor Single Slot	
DISPLAY CONECTORS	None	
POWER CONNECTORS	None	
WEIGHT (W/O EXTENDER)	240g	
PART NUMBER UND EAN	TCSP4M-PB	3536403352837

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