PNY

NVIDIA® TESLA® P4 by PNY

NVIDIA T4 by PNY

BREAKTHROUGH PERFORMANCE

T4 introduces the revolutionary Turing Tensor Core technology with multi-precision computing to handle diverse workloads. Powering breakthrough performance from FP32 to FP16 to INT8, as well as INT4 precisions, T4 delivers up to 40X higher performance than CPUs.

We're racing toward the future where every customer interaction, every product, and every service offering will be touched and improved by AI. Realizing that the future requires a computing platform that can accelerate the full diversity of modern AI, enabling businesses to create new customer experiences, reimagine how they meet—and exceed—customer demands, and cost-effectively scale their AI-based products and services.

The NVIDIA® T4 GPU accelerates diverse cloud workloads, including high-performance computing, deep learning training and inference, machine learning, data analytics, and graphics. Based on the new NVIDIA Turing™ architecture and packaged in an energy-efficient 70-watt, small PCIe form factor, T4 is optimized for scale-out computing environments and features multi-precision Turing Tensor Cores and new RT Cores. Combined with accelerated containerized software stacks from NGC, T4 delivers revolutionary performance at scale.

STATE-OF-THE-ART INFERENCE IN REAL-TIME

Responsiveness is key to user engagement for services such as conversational AI, recommender systems, and visual search. As models increase in accuracy and complexity, delivering the right answer right now requires exponentially larger compute capability. T4 delivers up to 40X times better low-latency throughput, so more requests can be served in real time.

VIDEO TRANSCODING PERFORMANCE

As the volume of online videos continues to grow exponentially, demand for solutions to efficiently search and gain insights from video continues to grow as well. T4 delivers breakthrough performance for AI video applications, with dedicated hardware transcoding engines that bring twice the decoding performance of prior-generation GPUs. T4 can decode up to 38 full-HD video streams, making it easy to integrate scalable deep learning into video pipelines to deliver innovative, smart video services.

TESLA T4 - PRODUCT SPECIFICATION

MEMORY SIZE (PER BOARD)	16 GB GDDR6	
MEMORY INTERFACE	256-bit	
MEMORY BANDWIDTH	320 Gb/s	
CUDA CORES	2560	
TURING TENSOR CORES	320	
SINGLE PRECISION FLOATING POINT PERFORMANCE	8,1Tflops (GPU Boost Clocks)	
MIXED PRECISION (FP16/FP32)	65 Tflops	
INT8-PRECISION	130 Tops	
INT4-PRECISION	260 Tops	
MEMORY INTERFACE	PCI Express 3.0 x16	
MAX POWER CONSUMPTION	70 W	
THERMAL SOLUTION	passive heatsink	
FORM FACTOR	Low profile PCI Express Form Factor Single Slot	
PART NUMBER UND EAN	TCST4M-PB TCST4MATX-PB	3536403367510 3536403367558

NVIDIA | PNY

Weitere Informationen: www.pny.eu/tesla

Follow us: @PNYproDE - @PNYproFR - @PNYproUK

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