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.

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

---

Achieve Over 4X the Inference Throughput

<table>
<thead>
<tr>
<th>Model</th>
<th>Images per Second (In Thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet</td>
<td>88800</td>
</tr>
<tr>
<td>GoogLeNet</td>
<td>51900</td>
</tr>
</tbody>
</table>

Note: GPU Tesla M60 (TensorRT + FP16) and Tesla M60 (TensorRT + INT8); NVIDIA Tesla P40 (TensorRT + INT8).
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.

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.

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.