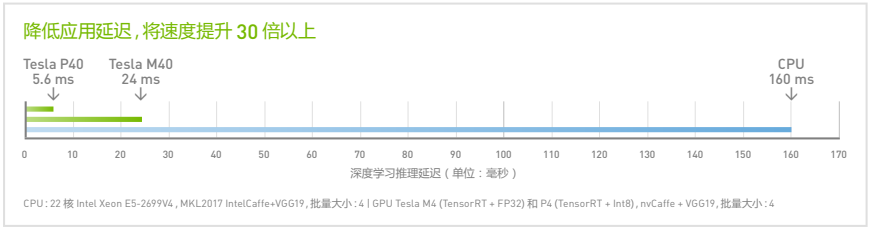
NVIDIA® TESLA® P40
推理加速器

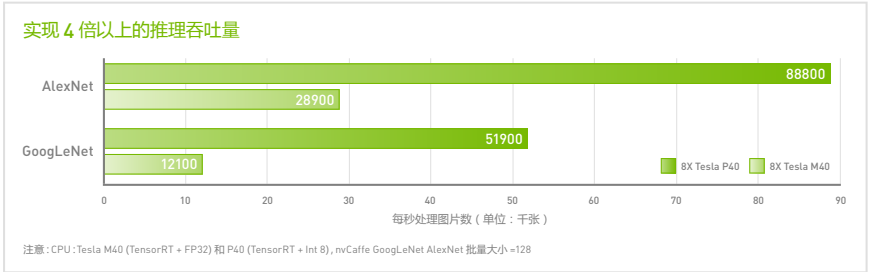
体验更大的推理吞吐量

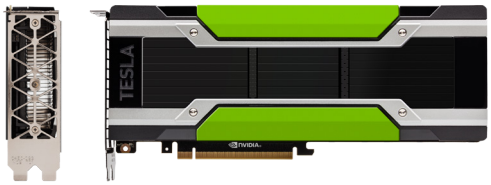
在人工智能和智能机器新时代，深度学习正以与历史上其他计算模型截然不同的方式改变着世界。采用革命性的 NVIDIA Pascal™ 架构的 GPU 是人工智能新时代的计算引擎，可加快大规模深度学习应用程序的速度，提供卓越的用户体验。

打造 NVIDIA Tesla P40 的主要目的是为深度学习部署提供更大的吞吐量。每个 GPU 可带来 47 TOPS（万亿次运算/秒）的推理性能和 INT8 运算能力，使得一台配备 8 个 Tesla P40 的服务器可提供相当于超过 140 台 CPU 服务器的性能。

随着模型的准确性和复杂性越来越高，CPU 已经无法再提供互动用户体验。Tesla P40 可在极其复杂的模型中实现实时响应，能够降低延迟，将性能提升为 CPU 的 30 倍以上。







功能

NVIDIA 家族迄今为止最快的推理工作负载处理器
47 TOPS 的 INT8 运算能力带来更大的推理吞吐量和响应速度
硬件解码引擎能对 35 路高清视频流进行实时转码和推理

规格

GPU 架构	**NVIDIA Pascal™**
单精度浮点运算能力	**12 TeraFLOPS***
整数运算能力 (INT8)	**47 TOPS*（万亿次运算/秒）**
GPU 显存	**24 GB**
显存带宽	**346 GB/s**
系统接口	**PCI Express 3.0 x16**
外形	**4.4”（高）x 10.5”（长），双插槽，全高**
最大功耗	**250 W**
已使用页面迁移引擎提升编程能力	**是**
ECC 保护	**是**
针对数据中心部署优化服务器	**是**
硬件加速视频引擎	**1 个解码引擎，2 个编码引擎**

* 启用了加速频率

NVIDIA TESLA P40 加速器的特性和利益点

打造 Tesla P40 的主要目的是为深度学习工作负载提供更大的吞吐量。

提供 140 倍的吞吐量以应对爆炸性数据的挑战

Tesla P40 配备新的 Pascal 架构，可带来超过 47 TOPS 的深度学习推理性能。在处理深度学习工作负载方面，一台使用 8 个 Tesla P40 的服务器最多可替代 140 台只使用 CPU 的服务器，因而可以大幅提升吞吐量并降低购买成本。

实时推理

Tesla P40 具备 INT8 运算能力，可在极其复杂的深度学习模型中实现实时响应，能将推理性能速度提升高达 30 倍。

通过单一的训练和推理平台简化了操作

目前，深度学习模型在 GPU 服务器上接受训练，但在 CPU 服务器上部署，以便进行推理。Tesla P40 提供极简工作流程，因此组织可以使用相同的服务器进行迭代和部署。

使用 NVIDIA 深度学习 SDK 加快了部署速度

通过 NVIDIA 深度学习 SDK 中所包含的 TensorRT 以及 Deep Stream SDK，客户可以轻松顺畅地利用新 INT8 运算能力和视频转码等推理功能。

如需详细了解 NVIDIA Tesla P40，请访问 **www.nvidia.cn/tesla**

NVIDIA