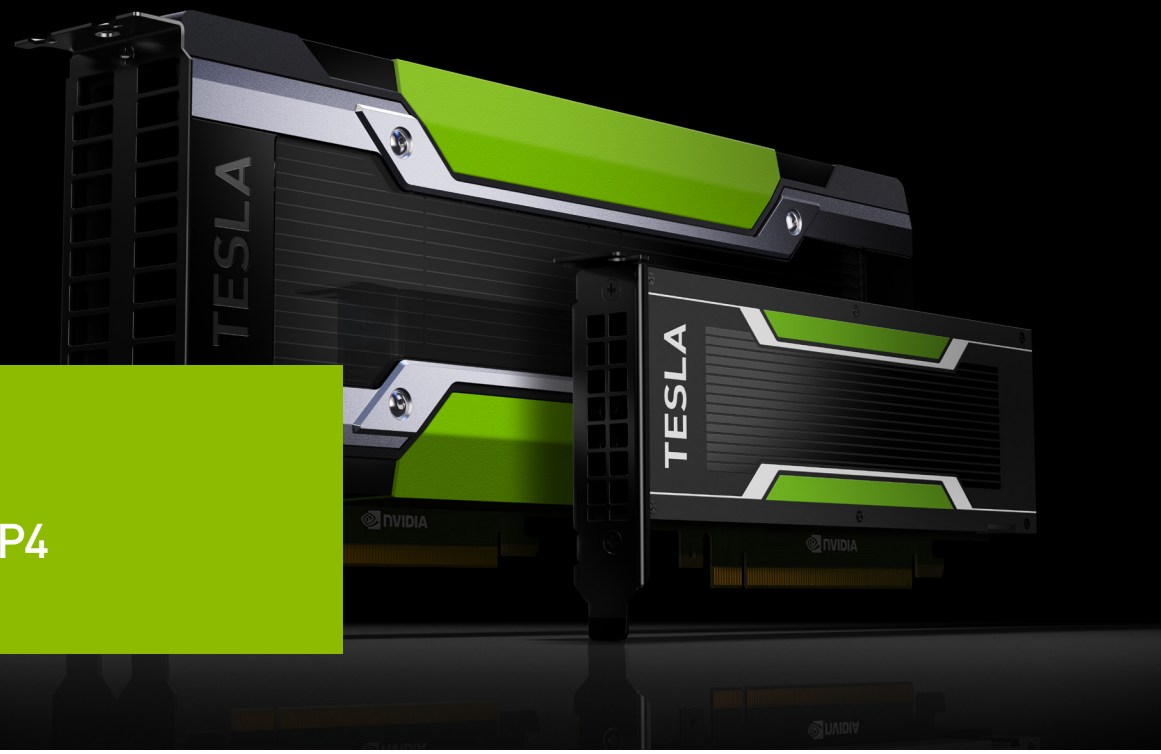


NVIDIA® TESLA® P4
推理加速器

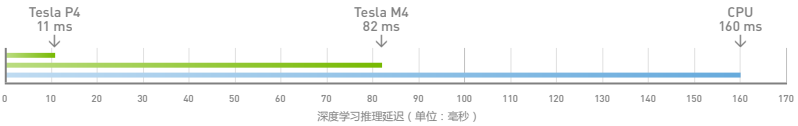
外扩型服务器中的超高效深度学习

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。互动语音、视觉搜索和视频推荐是我们日常使用的众多人工智能服务中的几项。

准确性和响应速度是决定用户是否采用这些服务的关键因素。随着深度学习模型的准确性和复杂性越来越高，CPU 已经无法再提供响应灵敏的用户体验。

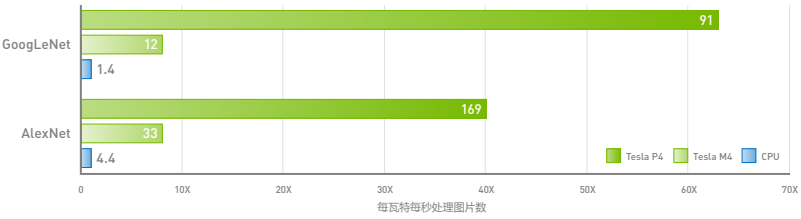
NVIDIA Tesla P4 采用革命性的 NVIDIA Pascal™ 架构，专为处理深度学习工作负载，启用智能响应人工智能服务的外扩型服务器而打造，可显著提升其运作效率。该显卡可将任何超大规模基础架构的延迟降低 15 倍，并可以提供比 CPU 高 60 倍的惊人能效。这使得我们开发了许多新的人工智能服务，这些服务在过去由于延迟限制而无法实现。

降低应用延迟，将速度提升 15 倍以上



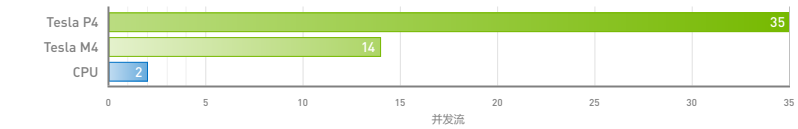
CPU：22 核 Intel Xeon E5-2699V4，MKL2017 IntelCaffe+VGG19，批量大小：4 | GPU Tesla M4 (TensorRT + FP32) 和 P4 (TensorRT + Int8)，nvCaffe + VGG19，批量大小：4

实现 60 倍以上的推理能效

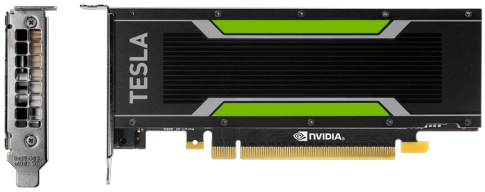


CPU：Intel Xeon E5-2690V4 MKL2017 IntelCaffe + GoogLeNet 和 AlexNet 批量大小：128 | GPU：Tesla M4 (TensorRT + FP32) 和 P4 (TensorRT + Int 8)，nvCaffe GoogLeNet AlexNet，批量大小：128

对 H.264 视频流进行转码和推理



注意：双 CPU Xeon E5-2650V4 | Tesla GPU M4 和 P4 | Ubuntu 14.04. H. 264 基准，FFMPEG 预设为 slow | HD = 720p，30 帧/秒。



功能

小巧的外形和 50/75W 的功耗，适用于任何外扩型服务器。

INT8 运算能力将延迟降低 15 倍。

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	**5.5 TeraFLOPS***
整数运算能力 (INT8)	**22 TOPS*（万亿次运算/秒）**
GPU 显存	**8 GB**
显存带宽	**192 GB/s**
系统接口	**PCI Express 半高外形**
最大功耗	**50W/75W**
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 P4 加速器的特性和利益点

打造 Tesla P4 的主要目的是在外扩型服务器中实现实时推理性能和智能用户体验。

通过实时推理提供快速响应的用户体验

响应速度是决定用户是否使用互动语音、视觉搜索和视频推荐等服务的关键因素。随着模型的准确性和复杂性越来越高，CPU 已经无法再提供响应灵敏的用户体验。Tesla P4 借助 INT8 运算能力提供高达 22 TOPS 的推理性能，可将延迟降低 15 倍。

为低功耗外扩型服务器带来更高效率

Tesla P4 凭借小巧的外形和 50W/75W 的功耗为经过密度优化的外扩型服务器加速。该显卡还可为深度学习推理工作负载提供比 CPU 高 60 倍的惊人能效，满足超大规模客户对人工智能应用程序飞速增长的需求。

借助专用解码引擎开发新的人工智能型视频服务

Tesla P4 配备硬件加速解码引擎，能对多达 35 路高清视频流进行实时转码和推理，该解码引擎可与进行推理的 GPU 并行运作。将深度学习集成到视频管线后，客户可以向用户提供之前无法实现的智能创新型视频服务。

使用 TENSORRT 和 DEEPSTREAM SDK 加快部署速度

TensorRT 是为优化部署到生产环境的深度学习模型而创建的库。它通常以 32 位或 16 位数据的形式获取接受训练的神经网络，并针对降低精度的 INT8 运算能力优化这些网络。NVIDIA DeepStream SDK 利用 Pascal GPU 的强大功能，可以同时解码和分析视频流。

如需详细了解 NVIDIA Tesla P4，请访问 **www.nvidia.cn/tesla**

NVIDIA