技术概述

NVIDIA 深度学习平台

从数据中心到网络终端，实现人工智能 (AI) 服务性能和效率的巨大飞跃

NVIDIA

简介

计算机科学家们为之奋斗半个多世纪的人工智能 (AI) 梦如今不再是科学幻想，它已在变革各行各业。AI 是指使用计算机来模拟人类智能。AI 可增强我们的认知能力——能够帮助我们解决极其复杂、缺失信息或细节易被忽略且需要专项训练的难题。

虽然机器学习领域已历经数十年进步，但深度学习 (DL) 的蓬勃发展是最近五年的事情。2012 年，多伦多大学的 Alex Krizhevsky 凭借使用 NVIDIA GPU 训练的深度神经网络在 ImageNet 图像识别大赛中一举夺魁——战胜了所有人类专家呕心沥血数十载研究得出的算法。同年，斯坦福大学的吴恩达认识到 “网络越大，认知越广” 后，与 NVIDIA 研发部合作开发出一种使用大型 GPU 计算系统训练网络的方法。这些开创性论文迅速点燃现代 AI 的爆发式发展，进而引发一系列 “超人” 般的成就。2015 年，Google 和 Microsoft 在 ImageNet 挑战赛中均超越了人类的最高得分。2016 年，DeepMind 的 AlphaGo 打破历史纪录，战胜了围棋冠军李世石，同时 Microsoft 的语音识别能力已达到人类水准。

GPU 已证明它们能够极有效地解决某些最复杂的深度学习问题，虽然 NVIDIA 深度学习平台是业界标准的训练解决方案，但其推理能力并非广为人知。从数据中心到终端，部分全球领先企业已使用 NVIDIA GPU 构建其推理解决方案。其中包括以下实例：

- **Twitter Periscope** 使用 GPU 运行推理，实时了解视频内容，实现更复杂的视频搜索和用户推荐功能。
- **Pinterest** 使用云端的 GPU 以最大程度缩短其 Related Pins 服务的用户等待时间（或延迟），从而根据用户的兴趣提供吸引人的推荐。
- **京东** 对 1,000 个高清视频频道的每帧视频实时运行基于推理的智能视频分析，并将其每台服务器的吞吐量提高 20 倍。
- **科大讯飞** 转而采用 Tesla GPU 在中国提供普通话语音识别服务，现在能够处理的并发请求数量是以前的 10 倍，而且其运营总体拥有成本 (TCO) 已降低 20%。
- **思科的 Spark Board 和 Spark Room Kit** 采用 Jetson GPU 重新塑造会议室，实现无线 4K 视频共享，运用深度学习提供语音和面部识别功能，同时加强资源规划。

NVIDIA 深度学习平台是一个覆盖数据中心到网络终端的平台。我们将在本文中介绍此平台如何实现性能和效率的巨大飞跃，显著降低数据中心的成本和网络终端的能耗。

深度学习工作流程

通过深度学习获得见解的两个主要过程是训练和推理。这两个过程虽然相似，但也有显著差异。在训练过程中，会提供诸如动物、交通标志等需要检测/识别的对象示例，让神经网络根据这些对象的内容作出预测。训练过程可强化正确的预测，更正错误的预测。经过训练后，所得神经网络的预测结果正确率最高可达 90-98%。“推理”即通过部署经过训练的网络来评估新对象，并按相似的预测精度作出预测。

图 1：依次显示训练、推理过程的高级别深度学习工作流程。

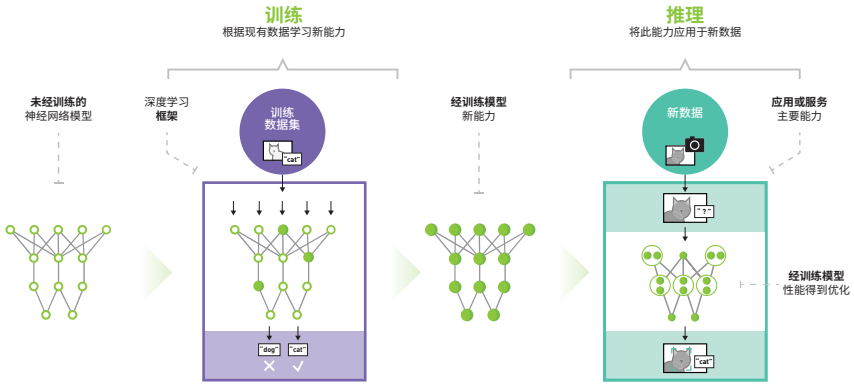


图 1

训练和推理都是先从前向传播计算开始，只是训练要完成更多步骤。训练时，在完成前向传播计算之后，还要将前向传播的结果与（已知的）正确答案进行比较，以计算误差值。后向传播阶段需将误差向后传播到网络的各层中，并使用梯度下降法更新各层的权重，以改善网络在尝试学习的任务中的表现。通常，在深度神经网络 (DNN) 训练过程中会将数百个训练输入（例如，图像分类网络中的图像或者用于语音识别的声谱图）分作一批并同时处理，旨在大量输入间摊销 GPU 显存的负载权重，从而提高计算效率。

推理过程也会批处理数百个样本，让数据中心彻夜运行的作业吞吐量达到最高水平，以便处理大量存储数据。这些作业的吞吐量往往比延迟更重要。但是，在实时使用情况下，批量大小较高也会增加延迟，对于这些使用情况，需要降低批量大小（最低一个样本），牺牲吞吐量以换取最低延迟。另外有种混合方法，有时被称为“自动批处理”，使用此方法要设置一个时间阈值（比如 10 毫秒），系统会在这 10 毫秒内批处理尽可能多的样本，然后发送这些样本进行推理。此方法在保持设定延迟量的同时可提供更高的吞吐量。

NVIDIA 深度学习平台

NVIDIA 平台旨在让世界各地的每一位开发者和数据科学家都能运用深度学习。所有主要的 DL 框架（包括 CAFFE、Caffe2、TensorFlow、Microsoft Cognitive Toolkit、PyTorch 和 MXNet）都可借助 NVIDIA 平台进行加速。此平台包含 NVIDIA DIGITS™ 等生产力工具，可让开发者快速设计最适合其数据的网络，同时无需编写任何代码。开发者可访问 NVIDIA 深度学习软件开发工具包（SDK）中的先进工具，为数据中心、自动驾驶汽车和网络终端设备开发应用程序。从数据中心到网络终端，皆可部署推理，NVIDIA 深度学习平台中用于优化神经网络以使其在部署中具有最佳表现的引擎即是 TensorRT。

基于 NVIDIA Volta 架构的 Tesla V100

NVIDIA® Tesla® V100 加速器采用强大的新型 NVIDIA Volta GPU 架构。Volta 不仅汲取了上一代产品 NVIDIA Pascal™ GPU 架构的精进之处，而且还显著提高了性能和可扩展性，并新增许多增强可编程性的功能。这些改进可有效提升 HPC、数据中心、超级计算机以及深度学习系统和应用程序的性能。

VOLTA 的主要特性

Tesla V100 的主要计算特性包括：

- **专为深度学习优化的全新流多处理器（SM）架构**：Volta GPU 中心配备有全新设计的 SM 处理器架构。专为深度学习设计的新 Tensor 核心在训练方面可提供高达 12 倍的 TFLOPS 峰值，而在推理方面则可提供 6 倍的 TFLOPS 峰值。
- **新一代 NVIDIA NVLink™：**新一代 NVIDIA NVLink 高速互联功能可增加带宽和链路，提高多 GPU 服务器配置的可扩展性。Volta GV100 最多支持六条 NVLink 链路，总带宽为 300 GB/s。
- **HBM2 显存：高速、高效，**Volta 拥有经重点调整的 16 GB HBM2 显存子系统，其显存带宽峰值达 900 GB/s。新一代 Samsung HBM2 显存与新一代 Volta 显存控制器结合，运行多工作负载时的显存带宽利用率高达 95%。
- **Volta 多进程服务：**Volta 多进程服务（MPS）是 Volta GV100 架构的新功能，可为 CUDA MPS 服务器的关键组件实现硬件加速，从而使共享 GPU 的多个计算应用程序提高性能、实现隔离并改进服务质量（QoS）。

- **统一内存寻址和地址转换服务质量提升：**
  V100 统一内存寻址技术包含新的存取计数器，可更准确地将内存分页迁移至对其读取最为频繁的处理器，同时提升处理器间共享显存范围的效率。
- **最大性能模式和最大效率模式：**
  最大性能模式下，Tesla V100 将以 300 W 的 TDP（热设计功耗）级别运行，提供极高的数据吞吐量。在最大效率模式下，数据中心管理员可调节 Tesla V100 加速器的功率利用率，使加速器以最佳性能功耗比运行。
- **协作组和新的协作启动 API：**协作组是 CUDA 9 中引入的新式编程模型，可用于组织线程通信群组。协作组允许开发者表示线程通信粒度，帮助他们表达更丰富、更高效的并行分解方法。
- **针对 Volta 优化的软件：**Caffe2、MXNet、Microsoft Cognitive Toolkit、PyTorch、TensorFlow 等深度学习框架新版本以及其他框架皆可发挥出 Volta 的强大性能，缩短训练时间并获得更高的多节点训练性能。

如需了解更多信息，请下载《Volta 架构白皮书》（链接：https://www.nvidia.com/zh-cn/data-center/volta-gpu-architecture/）

TensorRT - 可编程推理加速器

NVIDIA TensorRT™ 是一款高性能深度学习推理优化程序和运行时刻的环境，可为深度学习应用提供低延迟、高吞吐量的推理。TensorRT 可用于快速优化、验证和部署经训练的神经网络，从而在超大型数据中心、嵌入式平台或汽车产品平台上开展推理工作。

在训练神经网络后，TensorRT 可作为一种运行时刻的环境用于压缩、优化和部署该网络，同时不会产生框架开销。访问 TensorRT 的途径有三种：描述要运行的神经网络的 C++ API、可加载现有 Caffe 或 TensorFlow 模型的高级 Python 接口，或易用于 devops 环境的表现层状态转化（REST）API 接口。TensorRT 将网络层合并与模型压缩相结合，同时执行归一化和转换操作，以根据指定精度（FP32、FP16 或 INT8）转换为经优化的矩阵数学，从而减少延迟，提高吞吐量和效率。

推理计算可使用较低精度的张量运算，最大程度地减少精度损失。Tesla V100 和 P4 加速器分别对点积运算执行 16 位浮点（FP16）和 8 位整数（INT8）指令。这样可增加模型容量、显存利用率，缩短延迟，提高吞吐量以及效能。

经基准测试测量，Tesla P4 使用 INT8 后最多可提高 3 倍吞吐量，并能在缩短延迟的同时提高效能。

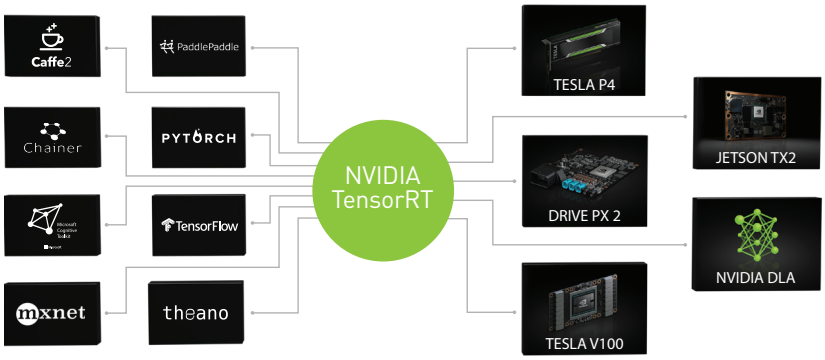


图 2

图 2：TensorRT 可从不同深度学习框架提取经训练的神经网络，然后在任何 NVIDIA 深度学习平台上为部署的推理优化这些神经网络。

NVIDIA TensorRT 是一款高性能深度学习推理优化程序和运行时刻的引擎，用于在生产环境中部署深度学习应用程序。您可在超大规模数据中心、嵌入式平台或汽车 GPU 平台上使用它快速优化、验证和部署经训练的神经网络进行推理。借助它，开发者可以发挥 NVIDIA Volta 架构 Tensor 核心的全部潜能，获得比上一代架构高三倍的性能。

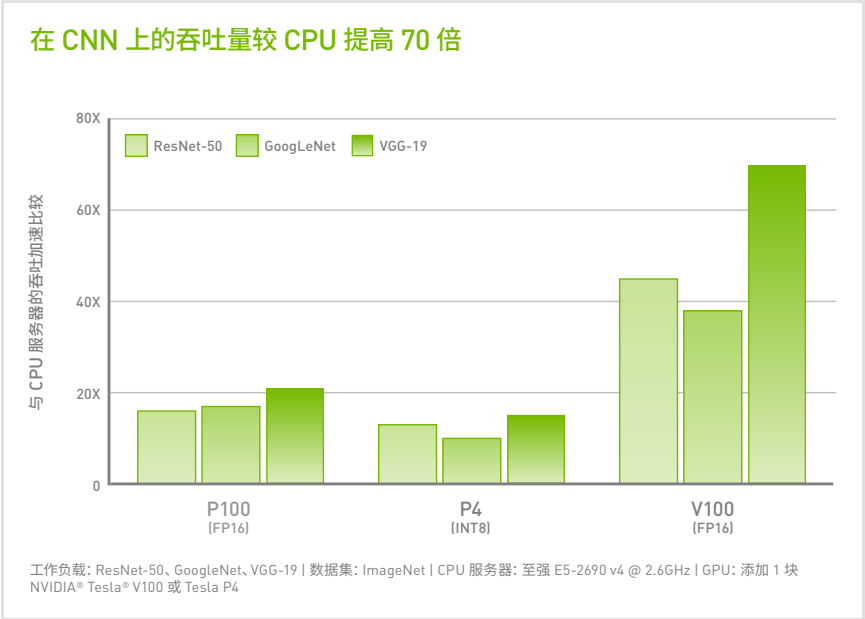
TensorRT 3 的主要特性包括：

- **支持 TensorFlow：**在 Tesla V100 上可以直接提取、优化和部署 TensorFlow 模型，与 TensorFlow 框架推理相比，性能最多提高 18 倍。
- **支持 Python API：**易用性更高，可让开发者使用 Python 脚本语言调用 TensorRT。
- **权重和激活精度优化：**将最大程度减少精度损失的同时将 FP32 全精度训练的模型量化为 FP16 和 INT8 精度，从而显著提高这些模型的推理性能。
- **网络层和张量融合（图形优化）：**将连续节点融合为单一节点以实现单内核执行，从而提高 GPU 利用率并优化内存和带宽。
- **内核自动调整：**为 Jetson、Tesla 或 DrivePX GPU 目标平台选择最佳的数据层和最佳的并行算法及内核，从而优化执行时间。
- **动态张量内存（内存优化）：**仅为每个张量分配其持续使用期间所需的内存，从而减少内存占用并改进内存重用
- **多流执行：**使用相同的模型和权重并行处理流，从而扩展为多个输入流图 2：TensorRT 可从不同深度学习框架提取经训练的神经网络，然后在任何 NVIDIA 深度学习平台上为所部署推理优化这些神经网络。

虽然在深度学习框架中也可执行推理运算，但 TensorRT 能轻松优化网络，显著提高性能。TensorRT 还可充分利用 Volta 架构，因此这一组合带来的吞吐量最多可比使用 CPU 的服务器高 70 倍。

下图显示了几种基于图像的卷积神经网络 (CNN) 的最高吞吐量，使用更大的批量大小 (128) 后，可提供最佳的吞吐量。

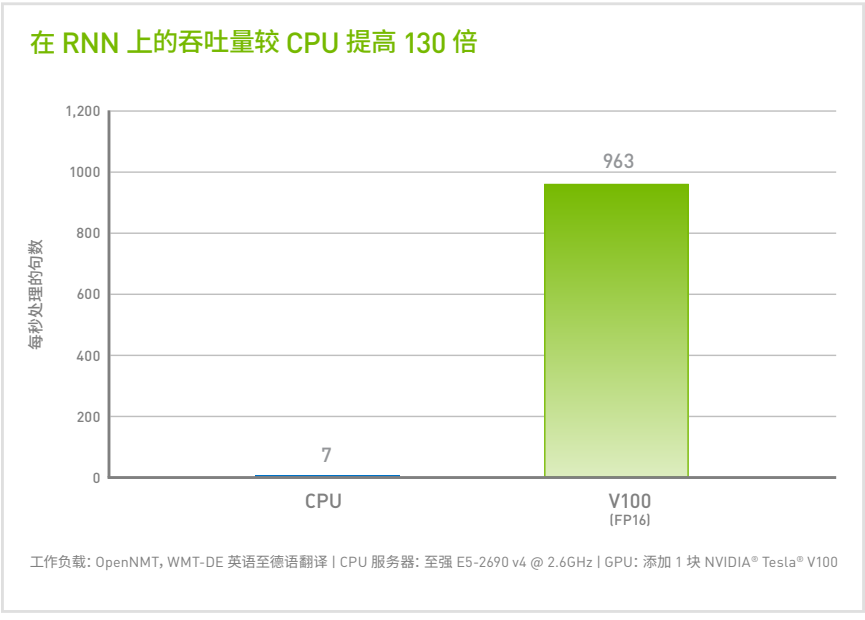
图表 1：TensorRT 支持 FP16 和 INT8 精度，且准确度损失几近于零。Tesla V100 与 TensorRT 相结合，其推理吞吐量可比使用 CPU 的服务器高 70 倍。



图表 1

随着深度学习使用案例研究的扩展，几乎每月都会有新型神经网络涌现，美国康奈尔大学 arXiv（链接：https://arxiv.org/）等网站发布的学术论文的数量即印证了这一点。语音识别、自然语言处理和翻译方面新兴的一类神经网络叫做递归神经网络 (RNN)。

图表 2：OpenNMT 是一个使用 Torch 数学工具包的全功能开源神经机器翻译系统。此处显示的结果来自名为 WMT-DE 的工作负载，此工作负载是将英语翻译为德语。在此处的推理测试中，对 CPU 测试使用英特尔深度学习 SDK 测试版 2，对 Tesla V100 测试使用 TensorRT 3。Tesla V100 提供的吞吐量高 130 多倍。



图表 2

GPU 推理：商业意义

Tesla V100 和 P4 可大幅提升性能和效能，但这对于购入预算和运营预算有何益处呢？简而言之：性能高，省得多。

下图显示了一台搭载八块 Tesla V100 的 HGX 服务器，其吞吐量等同于 120 台配备双插槽高端至强可扩展处理器 CPU 的服务器，后者占用三个服务器机架。这就代表着总体拥有成本 (TCO) 降低 13 倍。

数据中心 TCO 降低 13 倍

吞吐量为 48,000 张图像/秒，
1/12 成本 | 1/20 功耗 | 1 个机箱，3 个机架

图 1

图 1：一台搭载 8 块 Tesla V100 的 HGX 服务器 (左) 在 ResNet-50 提供的图像识别吞吐量约为 48,000 张图像/秒，其性能相当于含 120 台双路至强可扩展处理器 CPU 服务器的三个机架 (右) 的性能。为估计至强可扩展处理器的性能，我们使用已测得性能的至强 E5-2690v4，并应用 1.5 倍的性能扩展系数。

推理性能：概述

计算性能的衡量往往注重执行速度。但在深度学习推理性能中，速度只是四个发挥作用的关键因素之一。这里的四个因素是速度 (吞吐量)、延迟、效率和准确度。其中两个因素 (准确度和延迟) 是影响最终用户体验质量的关键因素，而另外两个 (吞吐量和效率) 则是数据中心效率的决定性因素。

推理性能剖析

图 3

图 3：最佳推理性能必须依靠四个方面的因素才能满足数据中心性能和用户体验要求。

数据中心效率

吞吐量：即给定时间段内的输出量。每台服务器的吞吐量通常使用每秒推理数或每秒样本数来衡量，它对经济高效的数据中心扩展至关重要，而 Tesla 加速器是业内一流的端到端数据中心训练和推理平台。

效率：即每单位功率提供的吞吐量，通常以性能功耗比来表示。效率是经济高效的数据中心扩展的另一个关键因素，因为服务器、服务器机架和整个数据中心的功率都必须限定在固定的功率预算内。

体验质量

延迟：即执行推理的时间，通常以毫秒为衡量单位。低延迟对快速提供日益增长且基于推理的实时服务至关重要。例如，为使基于语音的服务体现自然会话感，系统需要尽快为最终用户提供答案。即使滞后一秒也会让人觉得不自然。**Google 表示**[1] 7 毫秒是搜索等应用的最优延迟目标，因此在这 7 毫秒 内提供的吞吐量成为另一项重要的衡量标准。对于其他实时应用，最新的 **Google 演示**[2] 表明 200 毫秒是语音转文字或翻译应用的实际延迟目标。

准确度：即经训练的神经网络提供正确答案的能力。对于基于图像的应用，关键指标表现为排名前五或第一的百分比。这些“排名”指标表示的是分析样本最有可能出现的推理估值。排名前五或第一的百分比越高，推理答案的置信度就越高。对于语音和翻译服务，则以其他指标为准，比如双语评估研究 (BLEU) 评分。通常，神经网络训练需要的精度较高，而执行推理时通常会降低精度。降低精度后，可提高吞吐量、效率，缩短延迟，但高精度是提供一流用户体验的重要因素。TensorRT 可提供 FP32 和 FP16 浮点精度，以及 INT8 整数精度，且准确度损失几近于零。

Tesla V100 和 Tesla P4 在推理的四种关键推理性能指标值均比使用 CPU 的数据中心服务器高得多，且单个配备 Tesla GPU 的服务器节点最多可以替代 70 个使用 CPU 的节点。

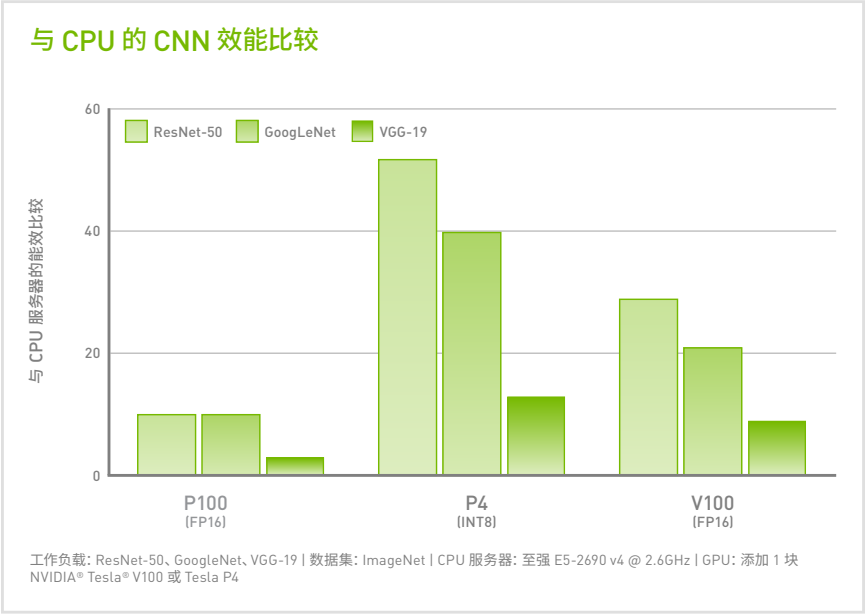
1. 参考网站：https://arxiv.org/ftp/arxiv/papers/1704/1704.04760.pdf

2. https://atscaleconference.com/videos/google-translate-breaking-language-barriers-in-emerging-markets/

效能

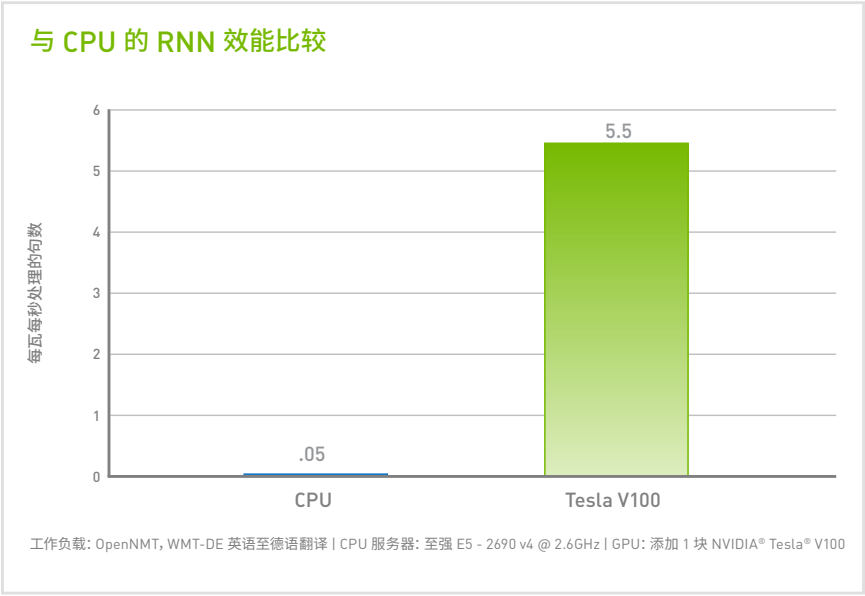
我们已达到最高吞吐量水平，尽管极高的吞吐量是深度学习工作负载的关键因素，但平台提供这种吞吐量的效率也是关键因素。

图表 3：Tesla V100 和 Tesla P4 均可大幅提升效能，效能按性能功率比计算得出。对于扩展推理和数据中心终端解决方案，Tesla P4 能够以 75W 的 TDP 提供卓越效率。Tesla V100 提供的吞吐量高得多，因为它使用的是 Volta 架构，且其 TDP 更高。



图表 3

图表 4：在 RNN 的语音推理过程中，配备 Tesla V100 GPU 服务器的效率比使用 CPU 服务器高 100 多倍。



图表 4

准确度

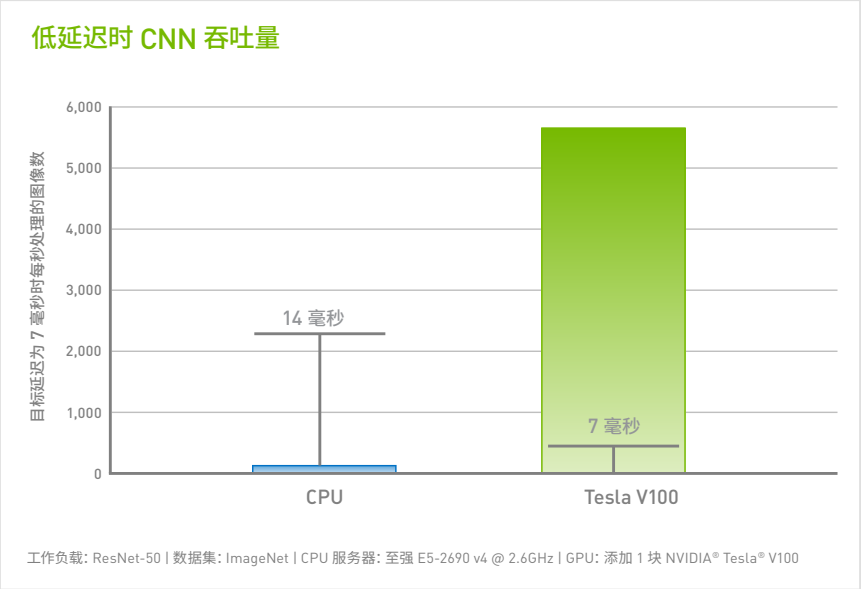
在研究推理性能的各个方面时，我们发现这些性能显然都休戚相关。一个顾此失彼的平台终将无法完成任务。Tesla V100 和 Tesla P4 与 TensorRT 3 结合堪称完美，因为它们的性能大幅提升，且准确度损失几近于零。

使用 TensorRT 后 FP16、INT8 的精度无损失

GoogLeNet (FP16)	FP32	FP16	差值
	72.23%	72.25%	+0.02%
GoogLeNet (INT8)	FP32	INT8	差值
	73.11%	72.54%	-0.57%

表 1

延迟



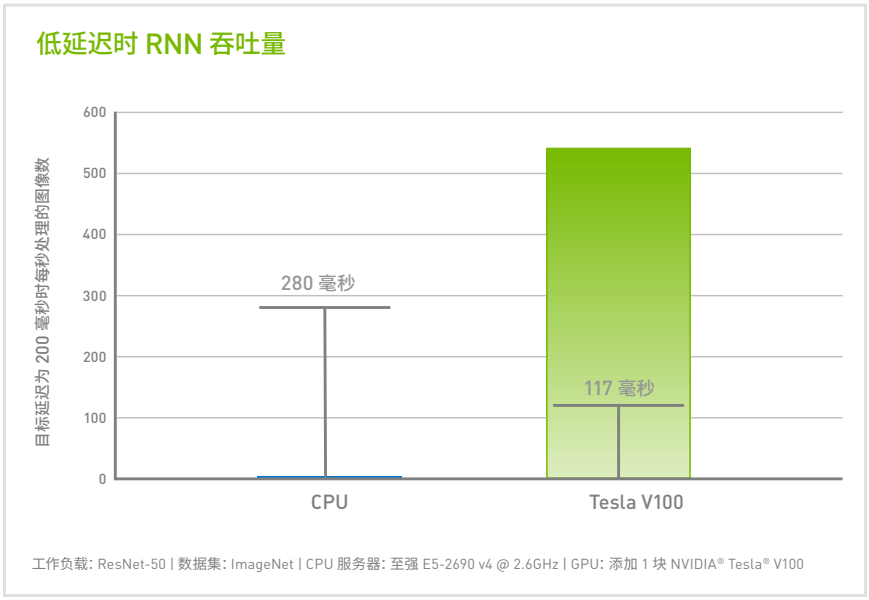
图表 5

图表 5：随着基于 AI 的实时服务涌现，延迟在推理性能中的重要性日益增加。对于优化最终用户体验，不仅高吞吐量至关重要，而且在指定延迟预算内提供高吞吐量也至关重要。**Google 表示**[3] 7 毫秒是最优延迟目标，此处应用了该延迟目标，上面的图表显示了在 7 毫秒延迟预算内 Tesla V100 的性能比使用 CPU 的服务器高 40 倍。相比之下，在指定延迟预算内，CPU 服务器的吞吐量无法达到 140 张图像/秒。

3. 参考网站：https://arxiv.org/ftp/arxiv/papers/1704/1704.04760.pdf

图表 6：对于语音应用，Google 表明[4] 在最新的演示中，语音应用的延迟目标约为 200 毫秒。此处 Tesla V100 的性能不仅比使用 CPU 的服务器高 150 多倍，而且其性能在指定的 200 毫秒 延迟预算内同样出色。相比之下，CPU 的吞吐量仅达到 4 句/秒左右，并且无法达到目标延迟界限。

4. 参考网站：https://atscaleconference.com/videos/google-translate-breaking-language-barriers-in-emerging-markets/



图表 6

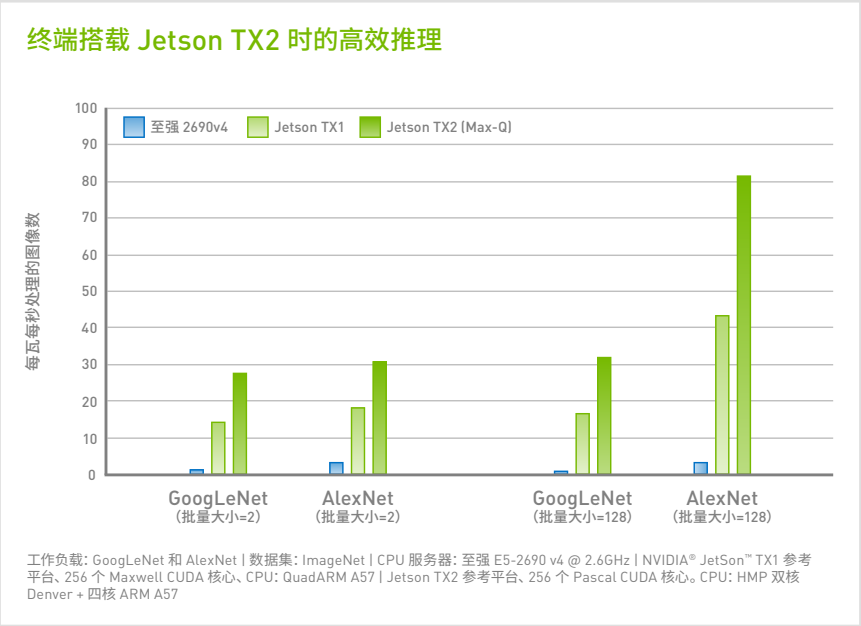
Jetson：终端推理

NVIDIA Jetson™ TX2 是一个信用卡大小的开放平台，可为终端赋予 AI 计算，打造高度智能的工厂机器人、商用无人机和智能摄像头，从而为我们开启通往 AI 城市的大门。基于 NVIDIA Pascal 架构的 Jetson TX2 提供两倍于其前代产品的性能，换言之，它的运行效能比前代产品高出两倍多，而功耗却不到 7.5 瓦。这让 Jetson TX2 能够在终端设备上运行更大、更深层次的神经网络，铸就更加智能、准确度更高且响应更迅速的设备，用于图像分类、导航和语音识别等任务。深度学习开发者在 Jetson 上使用的开发工具与他们在 CUDA、cuDNN 和 TensorRT 等 Tesla 平台上使用的极为相似。

Jetson TX2 经设计可在 7.5 瓦功耗的条件下达到处理效率峰值。这种性能水平被称为 Max-Q，在功率性能比曲线上表示最大性能和最大效能范围。此模块中包括电源在内的每个组件经过优化均可提供最高效率。GPU 的 Max-Q 频率为 854MHz，而 ARM A57 CPU 的频率为 1.2GHz。虽然动态电压频率调整 (DVFS) 允许 Jetson TX2 的 Tegra “Parker” SoC 根据用户负载和功耗调整运行时刻的频率，但 Max-Q 配置也可用于设置频率上限，以确保应用程序仅在最高效的范围内运行。

当不能连接到 AI 数据中心（例如遥感），或者实时应用（例如自主飞行无人机）的端到端延迟过高时，Jetson 可启用实时推理。虽然功率预算有限的大多数平台会因 Max-Q 状态而受益匪浅，但有些平台可能更偏好使用最高频率来实现吞吐峰值，纵使功耗增加且降低效率。DVFS 可经配置以其他频率范围（包括降频和超频）运行。Max-P 是另一种预设平台配置，允许平台在不到 15 瓦功耗的条件下达到最高系统性能。GPU 的 Max-P 频率为 1.12GHz，在启用 ARM A57 集群或 Denver 2 集群后，CPU 的频率为 2GHz，在同时启用这两种集群后，CPU 的频率为 1.4GHz。

图表 7：Jetson TX2 执行 GoogLeNet 推理的效率高达 33.2 张图像/秒/瓦，几乎是 Jetson TX1 的两倍。此外，Jetson TX2 的性能功率比基于至强* CPU 的服务器高 27 倍。



图表 7

对于许多网络终端应用程序而言，低延迟是必备条件。执行设备端推理远优于试着通过无线网络及在远程数据中心基于 CPU 的服务器内外发送此工作。除了设备端本地化功能以外，Jetson TX2 还可以通常不到 10 毫秒的超低延迟处理小批量工作负载。相比之下，基于 CPU 的服务器延迟约为 23 毫秒，再加上往返网络和数据中心的行程时间，该延迟数据会远超 100 毫秒。

加速计算的崛起

Google* 已宣布推出云张量处理器 (TPU)，适用于深度学习训练和推理。虽然 Google 和 NVIDIA 选择的开发途径不同，但这两种方法有几个共同点。具体而言，AI 需要加速计算。在摩尔定律逐渐趋缓的时代，为满足日益增长的深度学习需求，加速器可提供必要的数据处理能力。张量处理是提高深度学习训练和推理性能的核心技术。张量处理是各企业在构建现代数据中心时必须考虑的重要新工作负载。提高张量处理速度，可以显著降低现代数据中心构建成本。

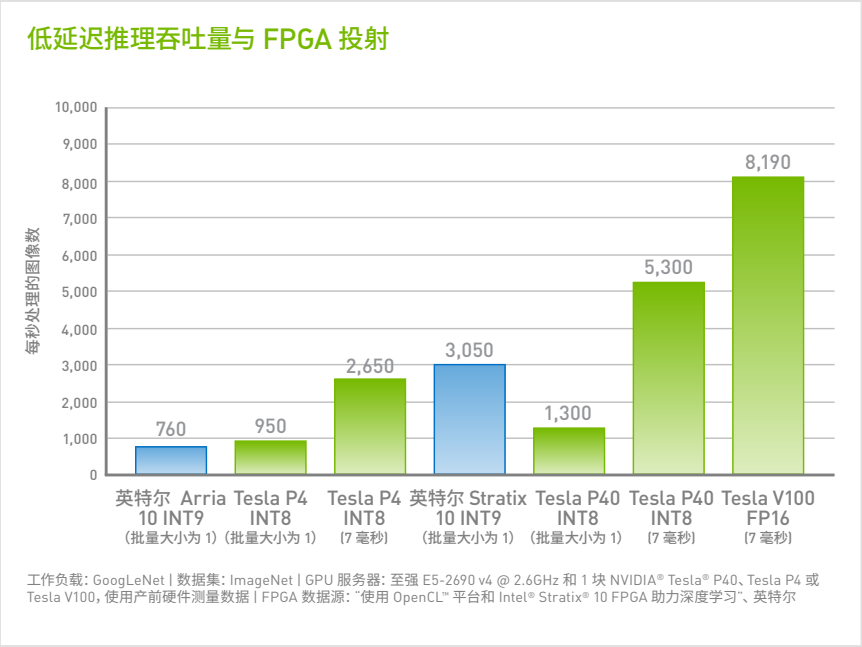
据 Google 称，Cloud TPU（亦称为“TPU2”）将于今年下半年面市，而单个 Cloud TPU 的计算能力可达 45 teraflops。NVIDIA Tesla V100 的深度学习训练和推理性能可达 125 teraflops。DGX-1 等 8 GPU 配置的深度学习计算能力现在可以达到 1 petaflop。

NVIDIA 的方法是面向每家公司、每个行业、每个计算平台普及 AI 计算，并提高从云到企业、汽车和网络终端中每个开发框架的速度。Google 和 NVIDIA 都是纯粹的领导者，我们可以密切合作，采用不同的方法开创 AI 世界。

FPGA 说明

随着深度学习领域持续发展，有人提议将其他类型的硬件用作潜在推理解决方案，比如现场可编程门阵列 (FPGA)。FPGA 已在网络交换机、4G 基站、汽车电机控制器和半导体测试设备等使用案例中用作特定功能。它具有大量通用可编程逻辑门，用途广泛，只需芯片事宜即可应用。由于是可编程门而非硬连接的 ASIC，因此 FPGA 本身的效率不高。

图表 8：比较在 GoogLeNet 网络上测得的 Tesla GPU 数据和英特尔指定的 Arria 10 和 Stratix 10 FPGA 投射的吞吐量。[5] 使用批量大小 1 后，会生成一个达到最低延迟的理论数值，最低延迟对依赖快速响应速度的推理型服务至关重要。但是，有一种改进的方法可设置延迟限制，然后获取该限制内的最高吞吐量，从而让开发者和最终用户两全其美，达到较高吞吐量和低延迟。Google 表示 7 毫秒 是基于推理的实时工作负载的卓越目标，使用此改进方法后，Tesla GPU 能够大幅提升吞吐性能和性能功率比效率。



图表 8

5. 使用 OpenCL™ 平台和英特尔产品提高深度学习速度：www.altera.com/content/dam/altera-www/global/en_US/pdfs/literature/wp/wp-01269-accelerating-deep-learning-with-opencl-and-intel-stratix-10-fpgas.pdf

关于可编程性和解决方案时间的注意事项

深度学习的快速创新刺激了可编程平台需求增长，而需求又促使开发者快速试验新型网络架构，随着新成果的面市这一格局周而复始地循环。

图 2：神经网络类型的多元性与复杂性仍在快速发展，一个能让开发者快速试验和更新换代的平台对助力深度学习创新至关重要。

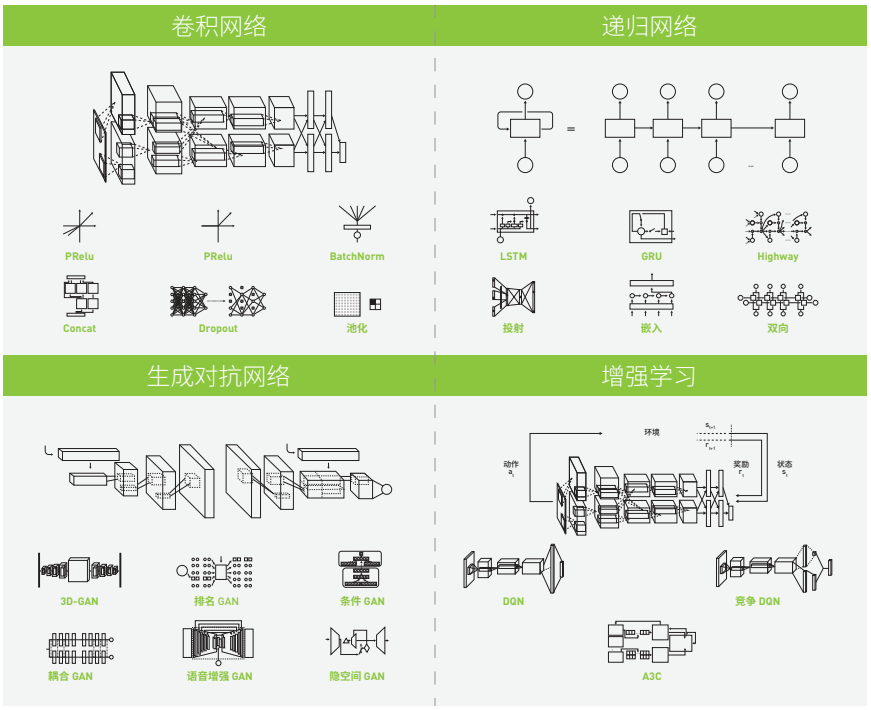


图 2

FPGA 在软件开发外仍面临的另一项挑战是，FGPA 必须重新进行硬件级配置，才能运行新一代的神经网络架构。复杂的硬件开发会让解决方案时间延迟数周，甚至数月，进而阻碍创新。而 GPU 仍然是可编程平台最理想的选择，得益于可靠的框架加速支持、Tesla V100 的 Tensor 核心等深度学习专用逻辑以及为部署推理优化经训练的网络的 TensorRT，它能够快速完成原型设计、测试和迭代前沿网络设计。

结束语

深度学习掀起了一场计算革命，为多个行业领域的企业带来了深远影响。NVIDIA 深度学习平台是训练的行业标准，各领先企业已纷纷为其推理工作负载部署 GPU 以利用其强大的优势。神经网络呈指数级迅速增长并不断复杂化，从而刺激了计算需求和成本激增。在一些情况下 AI 服务需要迅捷反应，而现代网络对于传统 CPU 而言计算任务过重。

推理性能的吞吐量、效率、延迟和准确度这四个方面对提高数据中心效率和用户体验至关重要。本文在“离线推理”使用案例中展示了 Tesla GPU 如何使数据中心 TCO 降低 13 倍。实际上，仅节省的能源成本就超过了配备 Tesla 的服务器成本。网络终端中的全新 Jetson TX2 可提供服务器级推理性能，而功耗还不到 10 瓦，并且可以启用设备本地推理，以显著缩短推理延迟时间。

高效的深度学习平台必须具备三种特质：必须具备专为深度学习定制的处理器；必须具备软件可编程特质；必须具备专门优化的行业框架，且框架由可在世界范围访问和采用的开发人员生态系统提供支持。NVIDIA 深度学习平台秉承这三种品质设计，是绝无仅有的端到端深度学习平台。从训练到推理，从数据中心到网络终端，都足见其特质。

如需了解有关 NVIDIA Tesla 产品的更多信息，请访问：
www.nvidia.com/zh-cn/data-center/tesla/

如需了解有关 JetsonTX2 的更多信息，请访问：
www.nvidia.cn/object/embedded-systems-cn

如需了解有关 TensorRT 和其他 NVIDIA 开发工具的更多信息，请访问：
developer.NVIDIA.com/tensorrt

如需了解目前已利用 GPU 加速的大量应用程序的列表，请访问：
www.NVIDIA.cn/object/GPU-applications-cn

性能数据表

CNN	TESLA V100（FP16/FP32 混合精度）				
网络	批量大小	性能（每秒处理的图像数）	主板总功率	性能/功率	延迟（毫秒）
GoogLeNet	1	876	98.6	8.88	1.14
	2	1,235	65.4	18.88	1.62
	4	2,194	80.4	27.29	1.82
	8	3,776	112.2	33.65	2.12
	64	8,630	209.2	41.25	7.42
	128	9,404	225.6	41.68	13.61
ResNet-50	1	504	94.2	5.35	1.99
	2	797	66.8	11.93	2.51
	4	1,450	83.7	17.32	2.76
	8	2,493	113.6	21.95	3.21
	64	5,572	196.4	28.37	11.49
	128	6,024	210.1	28.67	21.25
VGG-19	1	464	144	3	2
	2	718	138.7	5.18	2.79
	4	1,032	173.4	5.95	3.88
	8	1,334	203.4	6.56	6
	64	1,979	241	8.21	32.34
	128	2,030	238.4	8.52	63.04

CNN	TESLA P4（INT8 精度）				
网络	批量大小	性能（每秒处理的图像数）	主板总功率	性能/功率	延迟（毫秒）
GoogLeNet	1	837	42.4	19.74	1.19
	2	1,106	45.6	24.25	1.81
	4	1,489	49.1	30.33	2.69
	8	1,930	56.66	34.06	4.15
	64	2,531	64.25	39.39	25.29
	128	2,566	64.2	39.97	49.89
ResNet-50	1	600	32.9	18.24	1.67
	2	765	32.8	23.32	2.61
	4	1,019	33	30.88	3.93
	8	1,319	33.1	39.85	6.07
	64	1,715	33.2	51.66	37.32
	128	1,721	32.9	52.31	74.36
VGG-19	1	204	32.6	6	4.9
	2	273	32.9	8.30	7.33
	4	338	32.8	10.30	11.82
	8	380	32.66	11.64	21.04
	64	414	32.7	12.66	153.23
	128	438	32.8	13.35	292

RNN	TESLA V100 (FP16/FP32 混合精度)		
网络	批量大小	性能 (每秒处理的句数)	延迟 (毫秒)
OpenNMT	1	23	42
	2	46	43
	4	82	49
	8	156	51
	64	541	118
	128	725	176

JETSON TX2 (MAXQ 模式)							
网络	批量大小	性能 (每秒处理的图像数)	AP+DRAM 上行功率* (瓦)	AP+DRAM 性能/功率	GPU 下行功率* (瓦)	GPU 性能/功率	延迟 (毫秒)
AlexNet	1	119	6.6	18.0	2.3	52.4	8.4
	2	188	6.6	28.4	2.6	73.4	10.6
	4	264	6.7	39.3	2.9	92.6	15.2
	8	276	6.1	45.1	2.8	99.6	29.0
	64	400	6.4	62.6	3.2	125.7	160.0
	128	425	6.4	66.4	3.2	132.6	301.3
GoogLeNet	1	141	5.7	24.7	2.6	54.3	7.1
	2	156	5.9	26.2	2.7	57.6	12.8
	4	170	6.2	27.7	2.8	59.8	23.5
	8	180	6.4	28.2	3.0	60.6	44.5
	64	189	6.6	28.8	3.1	61.6	337.8
	128	191	6.6	28.9	3.1	61.6	671.8
ResNet-50	1	64.3	5.4	11.9	2.3	28.3	15.6
	2	76.5	5.3	14.4	2.3	33.7	26.2
	4	81.0	5.4	15.1	2.3	34.8	49.4
	8	83.4	5.4	15.4	2.4	35.4	95.9
	64	89.4	5.5	16.2	2.4	37.6	715.5
	128	89.9	5.5	16.2	2.4	37.7	1,424.3
VGG-19	1	18.8	7.2	2.6	2.9	6.4	53.1
	2	21.5	7.2	3.0	3.1	6.9	93.1
	4	22.6	7.3	3.1	3.1	7.2	176.8
	8	22.8	7.2	3.2	3.1	7.3	351.3
	64	22.9	7.2	3.2	3.2	7.1	2,792.4
	128	22.6	7.1	3.2	3.2	7.2	5,660.6

*Up = 上行功率 (高于电压调整器设定功率) , Down = 下行功率 (低于电压调整器设定功率)

JETSON TX2 (MAXP 模式)							
网络	批量大小	性能 (每秒处理的图像数)	AP+DRAM 上行功率* (瓦)	AP+DRAM 性能/功率	GPU 下行功率* (瓦)	GPU 性能/功率	延迟 (毫秒)
AlexNet	1	146	8.9	16.3	3.62	40.3	6.85
	2	231	9.2	25.2	4.00	57.7	8.66
	4	330	9.5	34.8	4.53	72.9	12.12
	8	349	8.8	39.8	4.42	79.0	22.90
	64	515	9.5	54.1	5.21	98.8	124.36
	128	546	9.6	56.9	5.28	103.5	234.32

GoogLeNet	1	179	8.2	21.8	4.14	43.2	5.6
	2	199	8.6	23.2	4.36	45.6	10.1
	4	218	9.0	24.2	4.61	47.2	18.4
	8	231	9.3	24.8	4.83	47.8	34.7
	64	243	9.7	25.1	5.03	48.3	263.6
	128	244	9.6	25.3	5.02	48.6	524.2
ResNet-50	1	82	7.4	11.1	3.49	23.5	12.2
	2	98	7.5	13.0	3.63	26.9	20.5
	4	104	7.6	13.6	3.71	27.9	38.6
	8	107	8.0	13.4	3.95	27.1	74.8
	64	115	7.9	14.6	3.81	30.1	558.9
	128	115	7.9	14.6	3.82	30.1	1,113.2
VGG-19	1	23.7	10	2.3	5	5.0	42.2
	2	26.8	10	2.6	4.93	5.4	74.7
	4	28.2	10	2.7	4.97	5.7	142.0
	8	28.3	10	2.8	4.96	5.7	282.7
	64	28.7	10	2.8	5.16	5.6	2,226.7
	128	28.4	10	2.8	5.09	5.6	4,514.0

*Up = 上行功率（高于电压调整器设定功率），Down = 下行功率（低于电压调整器设定功率）

JETSON TX1

网络	批量大小	性能（每秒处理的图像数）	AP+DRAM 上行功率*（瓦）	AP+DRAM 性能/功率	GPU 下行功率*（瓦）	GPU 性能/功率	延迟（毫秒）
AlexNet	1	95	9.2	10.3	5.1	18.6	10.5
	2	158	10.3	15.2	6.4	24.5	12.7
	4	244	11.3	21.7	7.6	32.0	16.4
	8	253	11.3	22.3	7.8	32.5	31.6
	64	418	12.5	33.5	9.4	44.5	153.2
	128	449	12.5		9.6	46.9	284.9
GoogLeNet	1	119	10.7	11.1	7.2	16.4	8.4
	2	133	11.2	12.0	7.7	17.4	15.0
	4	173	11.6	14.9	8.0	21.6	23.2
	8	185	12.3	15.1	9.0	20.6	43.2
	64	196	12.7	15.5	9.4	20.7	327.0
	128	196	12.7	15.5	9.5	20.7	651.7
ResNet-50	1	60.8	9.5	6.4	6.3	9.7	16.4
	2	67.8	9.8	6.9	6.5	10.5	29.5
	4	80.5	9.7	8.3	6.6	12.1	49.7
	8	84.2	10.2	8.3	7.0	12.0	95.0
	64	91.2	10.0	9.1	6.9	13.2	701.7
	128	91.5	10.4	8.8	7.3	12.6	1,399.3
VGG-19	1	13.3	11.3	1.2	7.6	1.7	75.0
	2	16.4	12.0	1.4	8.6	1.9	122.2
	4	19.2	12.2	1.6	8.9	2.2	207.8
	8	19.5	12.0	1.6	8.6	2.3	410.6
	64	20.3	12.2	1.7	9.1	2.2	3,149.6
	128	20.5	12.5	1.6	9.3	2.2	3,187.3

*Up = 上行功率（高于电压调整器设定功率），Down = 下行功率（低于电压调整器设定功率）

测试方法

我们的性能分析侧重于四种神经网络架构。AlexNet（2012 ImageNet 竞赛中胜出的架构）和较新的 GoogLeNet（2014 ImageNet 竞赛中胜出的架构），后者是比 AlexNet 更深层更复杂的神经网络，两者都是经典网络。VGG-19 和 ResNet-50 是最近的 ImageNet 竞赛的获胜架构。

为了涵盖一系列可能的推理情景，我们将考虑两种情况。第一种情况允许将许多个输入图像分作一批，以模拟某些使用案例，例如每秒有数千个用户提交图像的云环境中的推理。在这里，较大批量是能够接受的，因为等待分批并不会增加大量延迟时间。第二种情况涵盖了极其重视延迟的应用场合；在此情况下，某种程度的分批通常仍然可行，但对于我们的测试，考虑的是批量大小为 2 的小批量情况。

我们要比较五种不同的设备：NVIDIA Tegra X1 和 X2 客户端处理器，NVIDIA Tesla P4、V100 和英特尔*至强*数据中心处理器。为在 GPU 上运行这些神经网络，我们使用 TensorRT 2 EA，后者将在即将到来的 JetPack 更新（原定为 2017 年二季度发布）中推出。对于英特尔*至强* E5-2690 v4，我们运行英特尔*深度学习 SDK v2016.1.0.861 部署工具。

对于所有 GPU 结果，我们运行所有 TensorRT 版本附带的“giexec”二进制文件。这样可获取 prototxt 网络描述符和 caffe 模型文件，并可通过高斯分布使用随机图像和权重数据填充这些图像。对于 CPU 结果，我们运行“ModelOptimizer”二进制文件以及 prototxt 网络描述符和 caffe 模型文件，以生成执行与 MKL-DNN 关联的“classification_sample”二进制文件所必需的 .xml 模型文件。我们使用从 imagenet12 重新扩展、重新格式化到 RGB .bmp 文件的图像来运行英特尔*深度学习 SDK 推理引擎。TensorRT 和英特尔*深度学习 SDK 推理引擎均对 AlexNet 使用 227x227 图像，对 GoogLeNet、VGG-19 和 ResNet-50 使用 224x224 图像。在批量大小为 1 时运行英特尔*深度学习 SDK 推理引擎，我们测试的所有网络均会出现“bad_alloc”异常。而使用批量大小为 1 且与 MKL 2017.1.132 关联的 Intel Caffe 时，先使用 default_vgg_19 协议缓冲文件，然后使用 Caffe 的标准性能基准测试模式“caffe time”，图像与英特尔*深度学习 SDK 的相同。

我们比较 V100 上的 FP32 和 FP16 结果，以及 P4 上的 FP32 和 INT8 结果。所有 Tegra X1 和 X2 结果都使用 FP16。英特尔*深度学习 SDK 仅支持 FP32，因为至强* E5-2690 v4 不能对低精度的浮点数提供本地支持。

为在不同的系统之间比较功率，务必在配电网络中的某个一致点测量功率。电能是以高压形式配送的（调节前），之后调压器会将高电压转换为适合芯片上系统和 DRAM 的电压水平（调节后）。为进行分析，我们要比较整个应用处理器 (AP) 和 DRAM 组合的调节前功率。

在至强* E5-2690 v4 中，英特尔*深度学习 SDK 仅在一个插槽上运行。CPU 插槽和 DRAM 功率数据如同 pcm-power 实用程序报告的一样，我们认为这些数据是在相关调压器的输入端测得的。为测量 Tegra X1 和 X2 的调节前（上行）功率，我们使用皆由 9V 电源供电的 Jetson™ TX1 和 TX2 生产模块。TX1 的调压器输入端加装了主供电轨，TX2 有板载 INA 功率监控器。在 Tesla P4 和 V100 上，我们使用 NVSMI 实用程序报告生产显卡消耗的主板总功率。我们的 Tesla 测量结果中不包含系统 CPU 的功率，因为整个计算是在 GPU 上完成的；CPU 只是将工作提交给 GPU。

NVIDIA