技术概述

NVIDIA AI 推理平台

从数据中心到网络终端，实现 AI 服务性能和效率的巨大飞跃

NVIDIA

简介

人工智能革命如火如荼，为企业带来新的机遇，使他们能够另辟蹊径来解决客户面临的挑战。我们正在奔向一个 AI 遍地开花的未来，届时每次客户互动、每件产品和每项服务都将融入 AI 并借助 AI 实现改进。若要实现这一梦想，我们需要能够加速各种现代 AI 应用的计算平台，使企业能够创造新的客户体验，重新审视他们如何满足和超越客户需求，以及以经济高效的方式扩展其基于 AI 的产品和服务。

虽然机器学习领域已历经数十年进步，但深度学习 (DL) 在最近六年才开始蓬勃发展。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出的算法。同年，斯坦福大学的吴恩达在认识到“网络越大，认知越广”后，与 NVIDIA Research 团队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 已经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：

- **SAP 的品牌影响力服务**已实现 40 倍的增长，同时其成本降低到原来的 1/32。
- **Bing 视觉搜索**已将延迟时间缩短到原来的 1/60，并将自身成本降低到 1/10。
- **思科的 Spark Board 和 Spark Room Kit** 采用 NVIDIA® Jetson™ GPU，已实现无线 4K 视频共享，同时运用深度学习提供语音和面部识别功能。

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，你需要向神经网络提供诸如动物、交通标志等需要检测或识别的对象示例，让网络预测这些对象的内容。训练过程可强化正确的预测，并更正错误的预测。经过训练后，所得神经网络的预测结果正确率最高可达 90% 到 98%。“推理”是指通过部署经过训练的网络来评估新对象，并按相似的预测准确度作出预测。

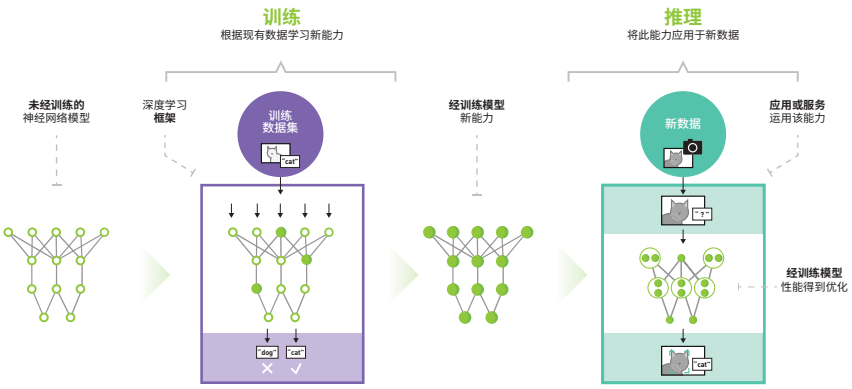


图 1

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，还要将结果与正确答案（“真值”）进行比较，以计算误差值。反向传播阶段需将误差发送到网络的各层中，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，以期在大量输入之间摊销 GPU 显存的负载权重，从而大幅提高计算效率。

推理过程也会批量处理数百个样本，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。另外还有一种混合方法，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 毫秒内批量处理尽可能多的样本，然后再发送这些样本以供推理。此方法在保持设定延迟量的同时可提供更高的吞吐量。

TensorRT 超大规模推理平台

NVIDIA TensorRT™ 超大规模推理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。该平台率先采用世界精尖的 AI 推理加速器：配备 NVIDIA Turing™ Tensor 核心的 NVIDA Tesla® T4 GPU。Tesla T4 依托 NVIDIA 的全新 Turing 架构，能够加速适用于图像、语音、翻译和推荐系统等各种领域的神经网络。Tesla T4 支持各种精度，并能加速各大 DL 框架，包括 TensorFlow、PyTorch、MXNet、Chainer 和 Caffe2。

强大的硬件需要精尖软件的加持，作为高性能深度学习推理平台，NVIDIA TensorRT 能为图像分类、分割、物体检测、机器语言翻译、语音和推荐引擎等应用程序提供低延迟、高吞吐量推理。它可以快速优化、验证和部署经过训练的神经网络，从而在超大型数据中心、嵌入式 GPU 或车用 GPU 平台上开展推理工作。TensorRT 优化程序和运行时支持 Turing GPU 在各类精度水平下发挥出色性能，从 FP32 到 INT8 无一不及。此外，TensorRT 还集成有 TensorFlow，能够支持各类采用 ONNX 格式的主要框架。

NVIDIA TensorRT 推理服务是 NVIDIA GPU Cloud 免费提供的即用型容器，也是一款适用于数据中心部署的生产就绪型深度学习推理服务。它能极大提高 GPU 服务器的利用率以降低成本，同时还可无缝集成到生产架构，从而节省时间。NVIDIA TensorRT 推理服务能够精简工作流程，同时还可简化向 GPU 加速推理架构的转换过程。

对于大规模、多节点部署，企业还可通过 NVIDIA GPU 上的 Kubernetes，将训练和推理部署无缝扩展到多云 GPU 集群。它能支持软件开发者及开发与运维 (DevOps) 工程师在节点集群之间自动部署、维护、调度和操作多个 GPU 加速应用容器。借助 NVIDIA GPU 上的 Kubernetes，开发者和工程师们能够大规模无缝构建 GPU 加速深度学习训练或推理应用程序，并将其部署到异构 GPU 集群。

基于 NVIDIA Turing 架构的 Tesla T4 Tensor 核心 GPU

NVIDIA Tesla T4 GPU 是全球顶级加速器，适用于所有 AI 推理工作负载。T4 搭载 NVIDIA Turing™ Tensor 核心，能够提供革命性的多精度推理性能以加速各种的现代 AI 应用程序。T4 是 NVIDIA AI 推理平台的组成部分，能够支持各类 AI 框架并提供全面的工具和集成功能，从而大幅简化高级 AI 的开发和部署工作。

Turing Tensor 核心专为加速 AI 推理而构建，并且 Turing GPU 还继承了 NVIDIA Volta™ 架构为 NVIDIA CUDA® 平台引入的所有增强功能，从而提升计算应用程序的能力、灵活度、效率和可移植性。Turing GPU 架构拥有诸多特性，包括独立线程调度、具有多应用程序地址空间隔离的硬件加速多进程服务 (MPS)、统一内存寻址和地址转换服务以及协作组等。

NVIDIA Turing 创新技术

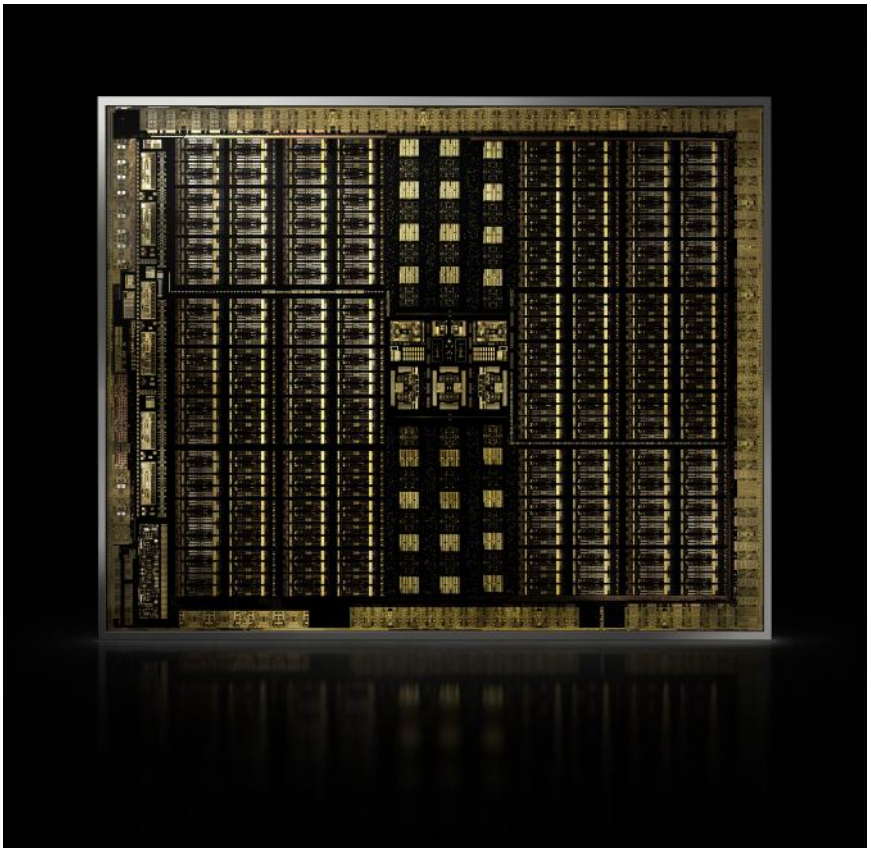


图 2：NVIDIA TURING TU102 GPU

Turing 的主要特性

> 具有 Turing Tensor 核心的新型流式多元处理器 (SM)

Turing SM 基于 Volta GV100 架构上经过重大改进的 SM 而构建，与上一代 NVIDIA Pascal™ GPU 相比，能够大幅提升性能和能效。Turing Tensor 核心不仅能像 Volta Tensor 核心一样提供 FP16 和 FP32 混合精度矩阵数学，而且新增了 INT8 和 INT4 精度模式，由此能大规模加速广泛的深度学习推理应用。

与 Volta 类似，Turing SM 也提供独立的浮点型和整型数据通路，能够通过混合计算和地址运算更有效地执行常见工作负载。此外，独立线程调度功能还可在线程之间实现细粒度同步与合作。最后，组合共享内存和 L1 缓存能够显著提高性能，同时简化编程。

> 用于推理的深度学习功能

Turing GPU 能够提供出色的推理性能、通用性和高效率。Turing Tensor 核心以及 TensorRT、CUDA 和 CuDNN 库的持续改进，共同助力 Turing GPU 为推理应用程序提供出色的性能。Turing 还包括实验特性，如支持 INT4 和 INT1 格式，能够进一步推动深度学习领域的研究和开发进程。

> GDDR6 高性能显存子系统

Turing 是首款利用 GDDR6 显存的 GPU 架构，该显存系统代表了高带宽 GDDR DRAM 显存设计的下一个重大进步，即最高可提供 320GB/ 秒的显存带宽。Turing GPU 中的 GDDR6 存储器接口电路经过全面重新设计，在速度、能效和降噪方面均实现了提升。与 Pascal GPU 中所用的 GDDR5X 显存相比，Turing 的 GDDR6 显存子系统分别在速度和能效方面实现了 40% 和 20% 的提升。

> 将视频解码性能提升一倍

视频持续呈爆炸式增长，已占据互联网全部流量的三分之二以上。借助 AI 进行的精确视频解释正在助力实现最相关的内容推荐，挖掘体育赛事中品牌植入的影响，向自动驾驶汽车提供感知能力，同时还可扩展至更多其他用途。Tesla T4 凭借专用的硬件转码引擎将解码性能提升至上一代 GPU 的两倍，从而为 AI 视频应用程序实现了性能突破。T4 可以解码多达 38 路全高清视频流，能够轻松将可扩展深度学习集成到视频流水线中，从而提供创新的智能视频服务。T4 具有性能和效率模式，能够在不损失视频画质的前提下实现快速编码或最低比特率编码。

TensorRT 5 特性

NVIDIA TensorRT 超大规模推理平台是一款完整的推理解决方案，包括前沿的 Tesla T4 推理加速器、TensorRT 5 高性能深度学习推理优化器和运行时以及 TensorRT 推理服务。此款强大的三合一解决方案能够为深度学习推理应用程序提供低延迟和高吞吐量，并能支持它们进行快速部署。该平台还可利用 Kubernetes 等工具，在多个主机上快速扩展容器化应用程序。借助 TensorRT 5，我们能够优化且精确校准低精度神经网络模型的准确度，并最终将模型部署到超大规模数据中心、嵌入式或汽车产品平台。在对各大框架中训练的模型进行推理时，GPU 上基于 TensorRT 的应用程序推理性能最高可达 CPU 的 50 倍。

图 3

TensorRT 针对多种深度学习推理应用程序的生产部署提供了 INT8 和 FP16 优化，例如视频流式传输、语音识别、推荐和自然语言处理。降精度的推理可以显著减少应用程序延迟，同时还可维持模型的准确度，恰巧满足了许多实时服务以及自动和嵌入式应用程序的要求。

TensorRT 和 TensorFlow 现已紧密集成，能够让开发者同时尽享 TensorFlow 的灵活性和 TensorRT 的超强优化性能。 MATLAB 已通过 GPU 编码器实现与 TensorRT 的集成，这能协助工程师和科学家在使用 MATLAB 时为 Jetson、NVIDIA DRIVE™ 和 Tesla 平台自动生成高性能推理引擎。

TensorRT 能够加速各种各样的应用程序，包括图像、视频、语音识别、神经网络机器翻译和推荐系统。

虽然深度学习框架也支持开展推理操作，但 TensorRT 不仅能轻松优化网络以提供更出色的性能，还能为多层感知器 (MLP) 和时间递归神经网络 (RNN) 添加新层。此外，TensorRT 还可充分利用 Turing 架构。后文将提供相关数据，用以说明二者的结合如何提供高达 CPU 服务器 45 倍的吞吐量。

Tesla GPU 与 TensorRT 推理优化器组合后，能够为卷积神经网络 (CNN)（常用于基于图像的网络）以及 RNN（常用于语音和翻译应用程序）带来巨大的性能提升。

推理性能：概述

在衡量计算性能时，我们往往注重执行速度。 但在深度学习推理性能中，速度只是七个发挥作用的关键因素之一。首字母缩略语 PLASTER 概括了这七个因素，如下所示：

Programmability	可编程性
Latency	延迟
Accuracy	准确度
Size of Model	网络大小
Throughput	吞吐量
Energy Efficiency	能效
Rate of Learning	学习率

图 4

深度学习是一项复杂的任务，因此我们要选择正确的深度学习平台。任何一种决策分析都应考虑这七个因素，而且这些因素中许多都是相互关联的。下面我们来了解一下这七个因素及其各自的作用。

- **可编程性：**机器学习正在经历爆炸式发展，这不仅体现在模型的大小和复杂性上，还体现在迅速涌现的多种神经网络架构上。为解决训练和推理难题，NVIDIA 开发出两个关键工具 CUDA 和 TensorRT，后者是 NVIDIA 的可编程推理加速器。此外，NVIDIA 的深度学习平台还能加快所有深度学习框架的训练和推理速度。

- **延迟：**人类和机器都需要对输入作出响应后才能制定决策并采取行动。延迟是指从提出请求到收到响应所经历的时间。尽管 AI 仍在迅猛发展，但实时服务的延迟目标却始终不变。例如，消费者和客户服务应用程序均对数字助理有着广泛的需求。但是，在人类尝试与数字助理交互时，即使是短短几秒的延迟也会开始让人感到不自然。

- **准确度：**准确度在各行各业都很重要，但医疗保健行业对准确度的需求尤为突出。过去数十年，医学成像技术取得了长足发展；在该技术的使用率不断增加的同时，我们也需要开展更多分析来识别医疗问题。医学成像技术的发展和应用也意味着，我们需要将大量数据从医疗设备传输给医疗专家进行分析。深度学习的一个优点是高精度训练和低精度实施。

- **网络大小：**深度学习模型的大小和处理器之间的物理网络容量会对性能造成影响，尤其是在 PLASTER 的延迟和吞吐量方面。深度学习网络模型的数量正在激增。此类模型的大小和复杂性也在增长，这不仅有助开展更详细的分析，还能推动对更强大训练系统的需求。

- **吞吐量：**开发者正在指定的延迟阈值内逐渐优化推理性能。延迟限定能够确保良好的客户体验，但在该限值内最大化吞吐量对于最大程度提高数据中心效率和营收至关重要。一直以来，业界都倾向于将吞吐量用作唯一的性能指标，原因是每秒计算次数越高，其他方面的性能通常也越好。如果未能在吞吐量和延迟之间取得适当的平衡，可能会导致客户服务水平低下、无法达到服务水平协议 (SLA) 的要求，同时还有可能致使服务失败。

- **能效：**随着 DL 加速器性能的提高，功耗也相应增加。要想让深度学习解决方案带来投资回报 (ROI)，我们不应仅关注系统的推理性能。功耗可能会迅速增加服务提供成本，因此我们有必要关注设备和系统的能效。因此，业界开始使用每瓦特推理次数（越高越好）来衡量运营成效。超大规模数据中心正设法最大程度地提高能效，也即在固定的功耗预算下提供尽可能多的推理次数。

- **学习率：**“AI” 由两个词组成，其中一个是智能，因此，用户希望神经网络能够在合理的期限内学习和适应。要使复杂的 DL 系统获得商业界的青睐，软件工具开发者必须支持 DevOps 行动。由于推理服务会收集新的数据，并且会不断发展和变化，因此必须定期重新训练 DL 模型。有鉴于此，IT 组织和软件开发者必须提升模型接收新数据和重新训练的频率。

吞吐量

基于图像的网络常用于图像和视频搜索、视频分析、物体分类和检测，以及许多其他用途。通过观察基于图像的数据集 (ImageNet) 在三个不同网络上的运行情况，我们可以得出：单个 Tesla P4 GPU 的运行速度是 CPU 服务器的 12 倍，而 Tesla V100 Tensor 核心 GPU 则是同一 CPU 服务器的 45 倍。

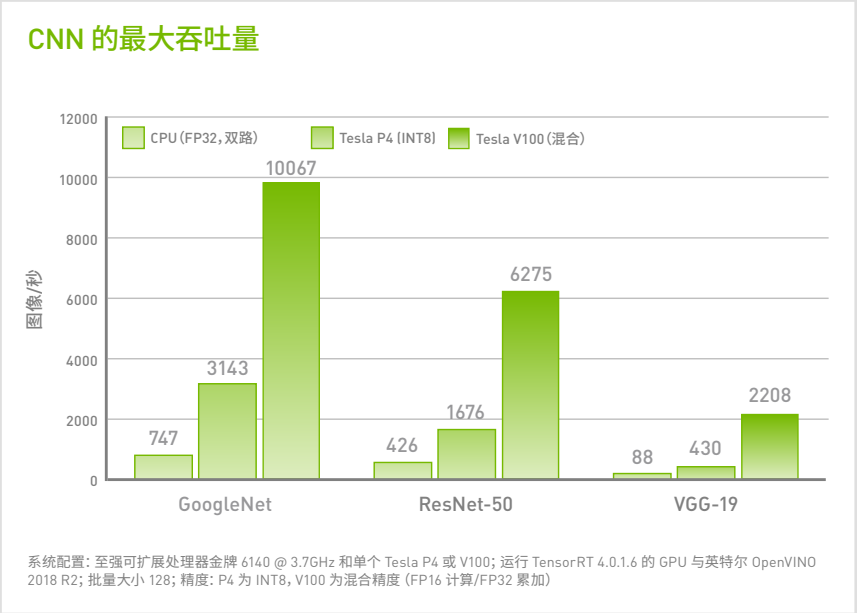


图 1

RNN 适用于时间序列或序列数据，并且常用作翻译、语音识别、自然语言处理乃至语音合成等应用的解决方案。此处所示数据源自 OpenNMT（神经网络机器翻译）网络，具体情境为将一个数据集从德语翻译成英语。相比 CPU 服务器，Tesla P4 能够提供 81 倍的吞吐量，而 Tesla V100 Tensor 核心 GPU 更为惊人，可提供高达 352 倍的吞吐量。

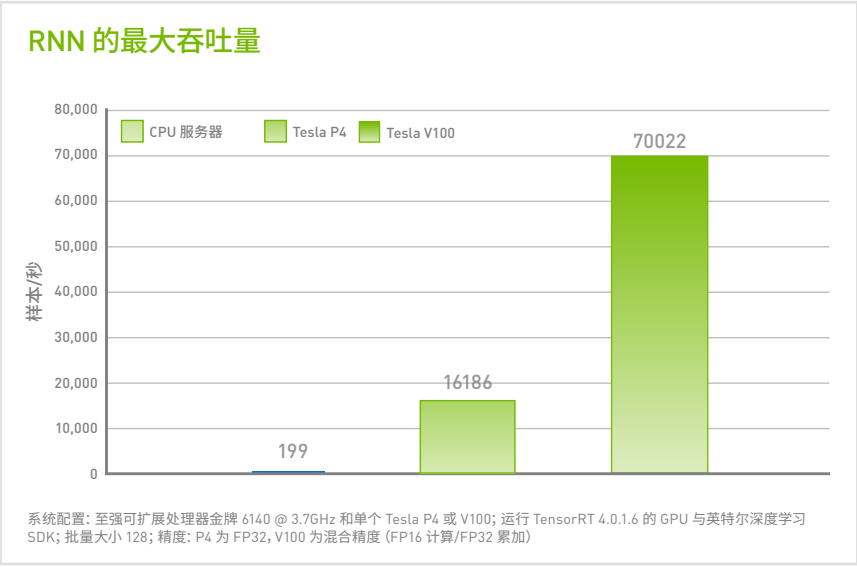


图 2

低延迟吞吐量

实现高吞吐量对于推理性能至关重要，而延迟亦是如此。许多实时用例实际都涉及对一个问题进行多个推理。例如，口语问题将涉及自动语音识别 (ASR)、语音转文字、自然语言处理、推荐系统、文字转语音，然后是语音合成。每个步骤都是不同的推理操作。虽然部分工作可以通过流水线完成，但是每个推理操作的延迟最终都会导致整体体验出现延迟。此处展示了 CNN 和 RNN 的低延迟吞吐量。 开发者通常采用两种方法来处理低延迟推理：1) 在没有批处理（即批量大小为 1）的情况下立即处理请求；或 2) 使用“自动批处理”技术，即首先设置一个延迟限值（例如 7 毫秒），之后对样本进行批处理，直至达到该延迟限值或某个批量大小的值（例如，批量大小为 8），然后再通过网络发送所处理的样本以进行推理。前一种方法更易于实现，而后一种方法则能在维持规定延迟限值的同时提供更多的推理吞吐量。为此，我们打算将延迟时间设为 7 毫秒来得出 CNN 结果，而使用批量大小 1 得出 RNN 结果。

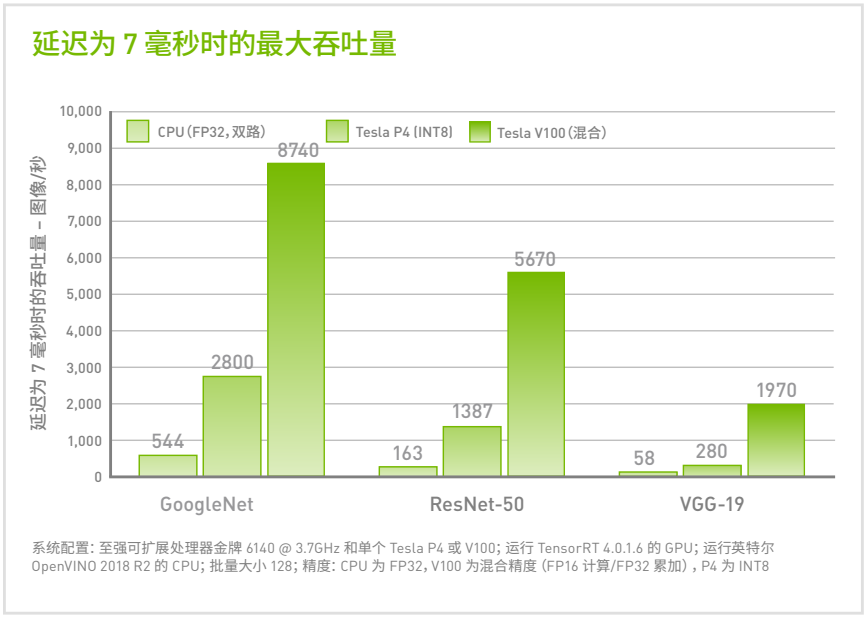


图 3

一直以来，人们都存有一些误解，认为 GPU 无法在批量大小为 1 的情况下实现极低延迟。但如下图所示，当批量大小为 1 时，Tesla P4 和 Tesla V100 的延迟时间分别为 1.8 毫秒和 1.1 毫秒，而 CPU 服务器则为 6 毫秒。此外，CPU 服务器每秒仅能传输 163 张图像，而 Tesla P4 每秒可传输 562 张图像，Tesla V100 每秒则可传输 870 张图像。

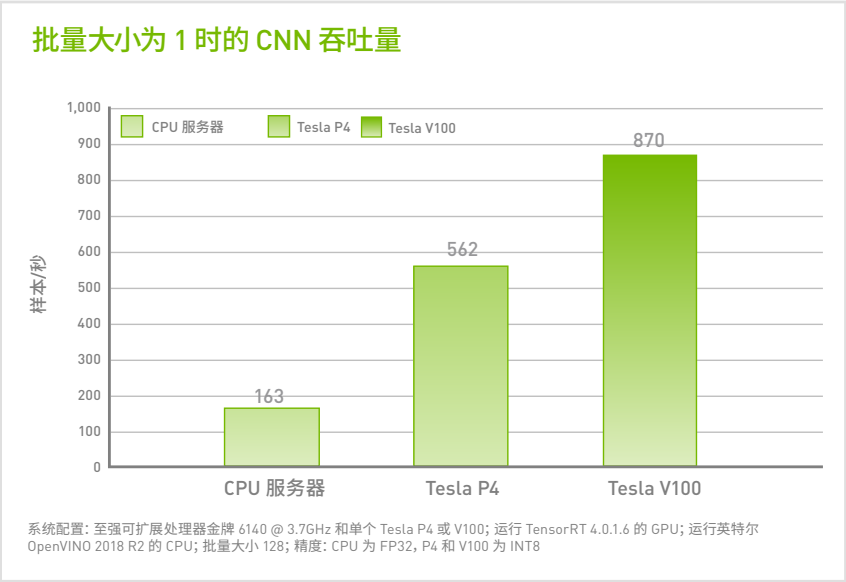


图 4

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。

我们首先来看看基于 Turing 的 Tesla T4，其效率远远超过 Tesla P4 或 Tesla V100。Tesla T4 凭借小巧的外形规格和 70 瓦的功耗设计，已成为全球顶尖的通用推理加速器。T4 搭载 Turing 架构 Tensor 内核，能够提供革命性的多精度推理性能，有效加速各式各样的现代 AI 应用程序。此外，研究人员还致力将 Tesla T4 的效率提升至其前一代 Tesla P4 的两倍以上。

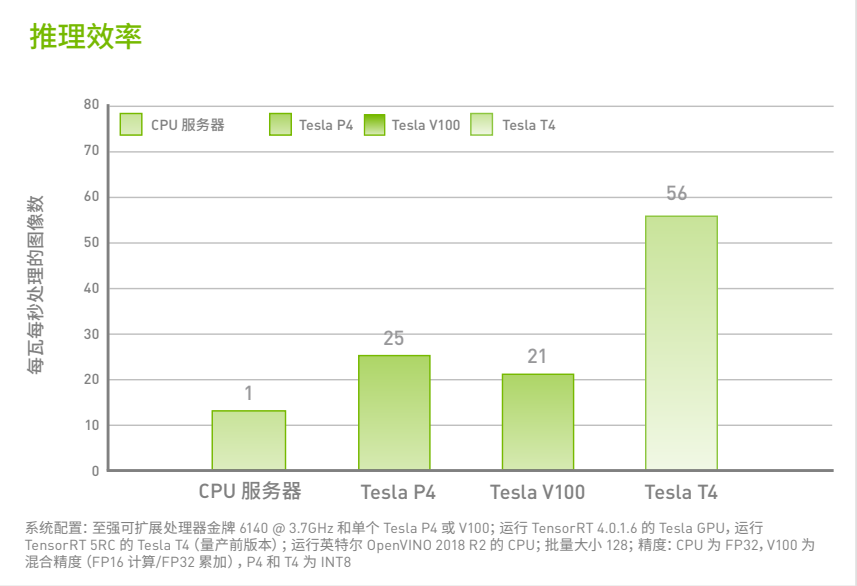


图 5

GPU 推理：商业意义

Tesla V100 和 P4 可大幅提升性能和能效，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 16 块 Tesla T4 GPU 的单一服务器，该服务器支持语音、NLP 和视频应用，吞吐量等同于 200 台占用四个服务器机架并需要 60 千瓦功耗的 CPU 服务器。成果如何？这台配备 Tesla T4 的服务器可将功耗降低至原来的 1/30，并能将服务器数量减少至原来的 1/200。

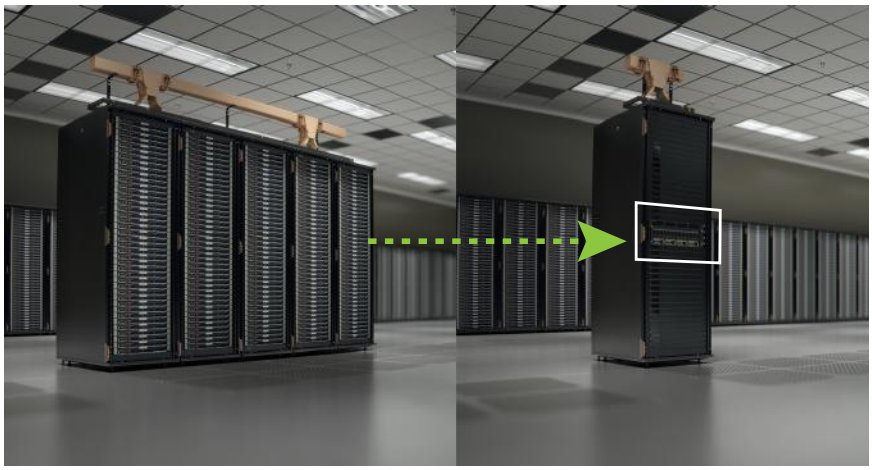


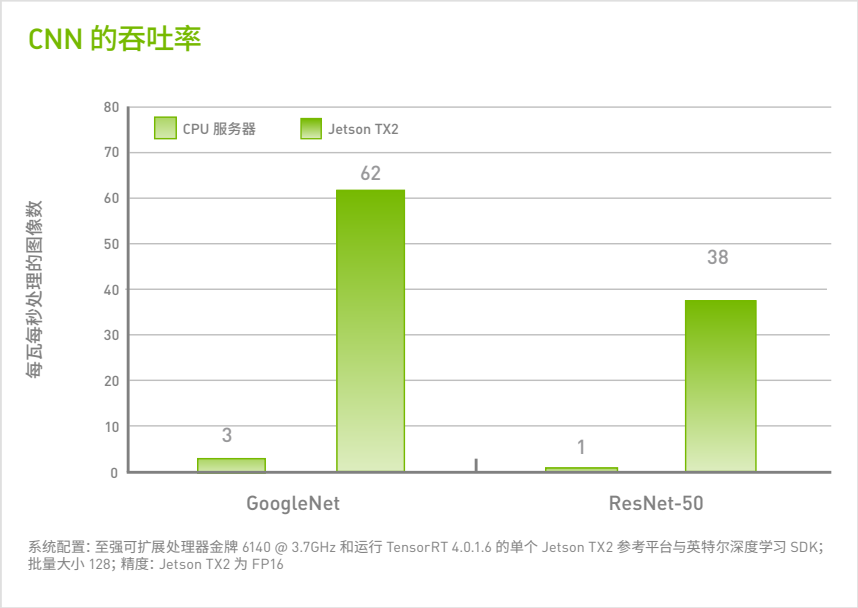
图 5

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。Jetson TX2 依托 NVIDIA Pascal 架构，可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，用于处理图像分类、导航和语音识别等各项任务。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 平台上使用的工具极为相似。

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能效范围。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，ARM A57 CPU 则为 1.2GHz。虽然动态电压频率调整 (DVFS) 允许 Jetson TX2 所依托的 NVIDIA Tegra® “Parker” 片上系统 (SoC) 根据用户负载和功耗调整运行时的时钟频率，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（例如遥感情境），或者实时应用的端到端延迟过高时（例如自主飞行无人机情境），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时钟频率来实现吞吐峰值，纵使这会增加功耗并降低效率。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 6

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 还能以通常低于 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)，该款处理器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，但这两种方法也存在一些共同点。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。张量处理是各企业在构建现代数据中心时必须考虑的新的主要工作负载，也是提高深度学习训练和推理性能的核心所在。提高张量处理速度，可以显著降低现代数据中心的构建成本。

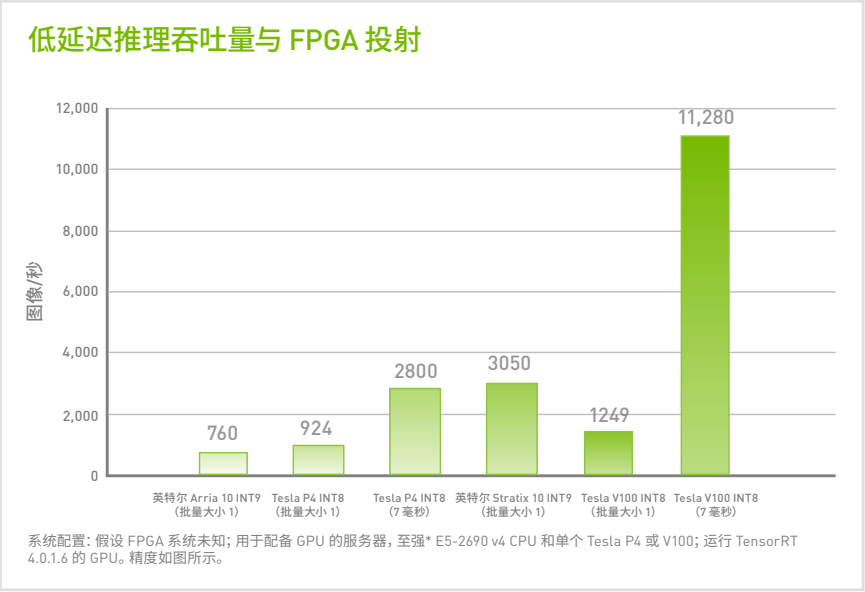
据 Google 表示，TPUv2（也称为“TPU 2.0”）可作为“Cloud TPU”提供，其由四个 TPUv2 芯片构成。但在比较芯片时，我们会发现单个 TPU 芯片可提供每芯片 45TFLOPS 的计算性能。NVIDIA Tesla V100 的深度学习训练和推理性能可达 125 TFLOPS。NVIDIA DGX-1™ 等 8 GPU 配置的深度学习计算能力现在可以达到 1 petaflop (PFLOP)。

NVIDIA 的方法是面向每家公司、每个行业以及每个计算平台普及 AI 计算，并为从云端到企业、汽车乃至网络终端中的每个开发框架实现加速。Google 和 NVIDIA 都是公认的领导者，我们可以密切合作，同时采用不同的方法开创 AI 世界。

FPGA 说明

深度学习领域依旧保持迅猛的发展势头，有人提议将其他类型的硬件用作潜在的推理解决方案，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，专门设计用于模拟专用集成电路 (ASIC)，用途广泛，只需芯片适宜即可应用。由于是可编程门而非硬连接的 ASIC，因此 FPGA 本身的效率并不高。

在最近的 Build 大会上，微软声称其基于 FPGA 的 Project BrainWave 推理平台可以在 ResNet-50 图像网络上每秒约传输 500 张图像。不过，我们要正确地看待这一点，因为单个 Tesla P4 GPU 可以提供超过 3 倍的吞吐量，或在 75 瓦解决方案中每秒传输 1676 张图像。为进一步比较，下方展示了英特尔最近一篇白皮书中关于 Altera 和 Stratix FPGA 所作的预测。请注意这些结果均在 GoogLeNet 网络上运行得出。



图表 7

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对可编程平台需求的增长，而不断增加的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。回顾一下，可编程性是指 PLASTER 框架中的“P”。最近几年，我们经历了新型网络架构的寒武纪大爆发，而且这一创新速度丝毫没有减缓的迹象。

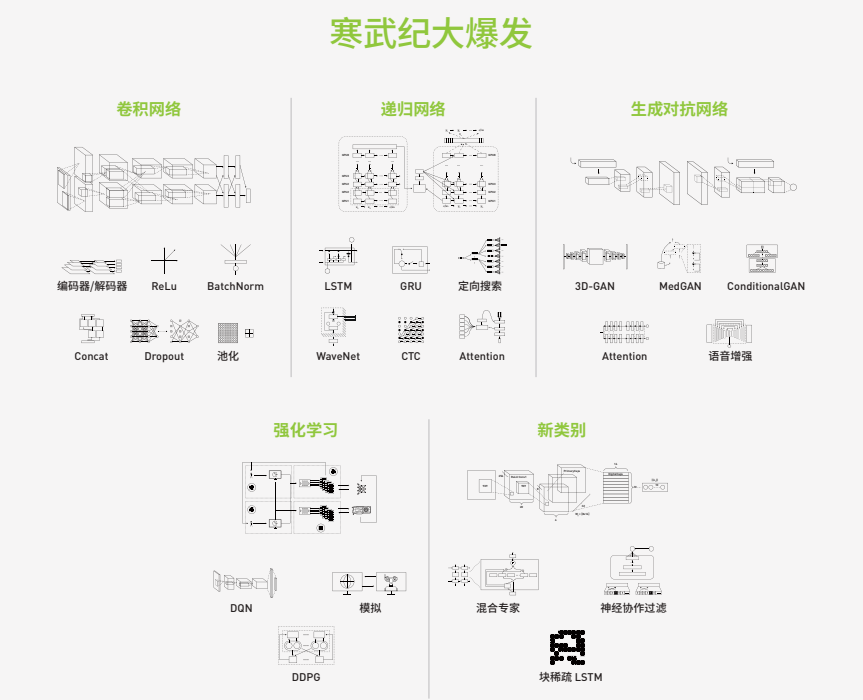


图 6

除软件开发外，FPGA 提出的另一项挑战是，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。这种复杂的硬件开发减缓了提供解决方案的时间，由此也将创新速度减慢了数周乃至数月。另一方面，GPU 仍然是可编程平台的理想选择，得益于可靠的框架加速支持、Tesla V100 的 Tensor 内核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作业的行业标准，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。

PLASTER 有效概括了推理性能的七个方面：可编程性、延迟、准确度、网络大小、吞吐量、效率和学习率。这七个方面对于提升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 如何使数据中心所需的服务器数量最多降至原来的 1/200。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 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。这些巨大的改进将能使顶尖的 AI 端到端应用于实时服务中，包括语音识别、语音转文字、推荐系统、文字转语音和语音合成。

高效的深度学习平台必须具备三种特质：1) 必须具备专为深度学习定制的处理器；2) 必须具备软件可编程特质；3) 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者生态系统提供支持。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.cn/tesla

如要深入了解 Jetson TX2，请访问：
www.nvidia.com/zh-cn/autonomous-machines/embedded-systems

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.com/GPU-applications

性能数据表

CNN		TESLA P4 (INT8)			
网络	批量大小	性能 (每秒处理的图像数)	主板平均功率	每瓦性能	延迟(毫秒)
GoogLeNet	1	923	37	24.9	1.1
	2	1215	40	30.4	1.6
	4	1631	42	38.8	2.5
	8	2197	46	47.8	3.7
	64	3118	63	49.5	20
	128	3191	64	49.1	40
ResNet-50	1	569	44	12.9	1.8
	2	736	44	16.7	2.7
	4	974	49	19.9	4.1
	8	1291	57	22.6	6.2
	64	1677	63	26.6	38
	128	1676	62	27	76
VGG-19	1	206	55	3.7	4.9
	2	280	53	5.3	7.1
	4	346	60	5.8	12
	8	398	65	6.1	20
	64	429	63	6.8	149
	128	430	62	6.9	298

CNN		TESLA V100(混合精度)			
网络	批量	性能 (每秒处理的图像数)	主板平均功率	每瓦性能	延迟(毫秒)
GoogLeNet	1	1027	131	8	0.97
	2	1553	104	15	1.3
	4	2684	118	23	1.5
	8	4502	152	29	1.8
	64	9421	284	33	6.8
	128	10067	290	35	13
ResNet-50	1	476	120	4	2.1
	2	880	109	8.1	2.3
	4	1631	132	12.4	2.5
	8	2685	153	17.5	3
	64	5877	274	21.4	11
	128	6275	285	22	20

VGG-19	1	497	151	3.3	2
	2	793	194	4.1	2.5
	4	1194	220	5.4	3.4
	8	1488	254	5.9	5.4
	64	2161	290	7.5	30
	128	2208	291	7.6	58

RNN		TESLA P4 (FP32)			
网络	批量大小	性能（每秒处理的令牌数）	主板平均功率	每瓦性能	延迟（毫秒）
OpenNMT	1	894	103	8.7	1.1
	2	1260	126	10	1.6
	4	1746	129	13.5	2.3
	8	2901	168	17.3	2.8
	64	5903	289	20.4	11
	128	6259	294	21.3	20

RNN		TESLA V100（混合精度）			
网络	批量大小	性能（每秒处理的令牌数）	主板平均功率	每瓦性能	延迟（毫秒）
OpenNMT	1	3457	96	36	15
	2	4791	100	47.9	21
	4	8076	105	76.9	25
	8	13475	108	124.8	30
	64	50758	74	685.9	64
	128	70022	84	833.6	93

		JETSON TX2（MAX-Q 模式）					
网络	批量	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	8.4
	2	188	6.6	28.4	2.6	72	10.6
	4	264	6.7	39.3	2.9	91	15.2
	8	276	6.1	45.1	2.8	99	29.0
	64	400	6.4	62.6	3.2	125	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	5.4	11.9	2.3	28.3	15.6
	2	77	5.3	14.4	2.3	33	26.2
	4	81	5.4	15.1	2.3	34.8	49.4
	8	83	5.4	15.4	2.4	35.4	95.9
	64	89	5.5	16.2	2.4	37	715.5
	128	90	5.5	16.2	2.4	37.7	1,424.3

VGG-19	1	19	7.2	2.6	3	7	53.1
	2	22	7.2	3.0	3.1	6.9	93.1
	4	23	7.3	3.1	3.1	7.2	176.8
	8	23	7.2	3.2	3.1	7.3	351.3
	64	23	7.2	3.2	3.2	7.1	2,792.4
	128	23	7.1	3.2	3.2	7.2	5,660.6

*Up = 上行功率（高于调压器设定功率），Down = 下行功率（低于调压器设定功率）

JETSON TX2（MAX-P 模式）

网络	批量	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	41	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	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	49	263.6
	128	244	9.6	25.3	5.02	48.6	52
ResNet-50	1	82	7.4	11.1	3.49	23	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（高于调压器设定功率），Down = 下行功率（低于调压器设定功率）

JETSON TX1

网络	批量	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	32.0	31.6
	64	418	12.5	33	9.4	44.0	153.2
	128	449	12.5	36	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	15.0	9.4	20.7	327.0
	128	196	12.7	15.0	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	10.0	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（高于调压器设定功率），Down = 下行功率（低于调压器设定功率）

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 层次更深、更复杂的神经网络，两者都是经典网络。VGG-19 和 ResNet-50 是在最近的 ImageNet 竞赛中获胜的架构。

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。我们此时可以使用较大批量，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 和 V100 以及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 JetPack 更新（原定于 2017 年第二季度发布）中推出。对于英特尔至强可扩展处理器金牌 6140，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 使用尺寸为 227 x 227 的图像，对 GoogLeNet、VGG-19 和 ResNet-50 使用 224 x 224 大小的图像。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 FP16 混合精度结果，以及 P4 上的 INT8 结果。所有 Tegra X1 和 X2 结果都使用 FP16。Intel 深度学习 SDK 仅支持 FP32。

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。电能以高压形式配送（调节前），之后调压器会将高电压转换为适合片上系统和 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。

在英特尔至强可扩展处理器金牌 6140 上，英特尔 OpenVINO 库运行在单插槽上，用于 CNN 测试（GoogLeNet、ResNet-50 和 VGG19）。对于 RNN 测试，我们则使用英特尔深度学习 SDK，因为 OpenVINO 仅支持 CNN。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 量产版模块。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 实用程序报告量产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。

法律声明和商标

一般条款与条件。NVIDIA 产品并非针对医学、军事、航空、航天或生命保障设备而设计，并未授权用于也不保证适合用于上述设备，亦不得用于 NVIDIA 产品之失效或故障合理预计会造成人身伤亡或财产或环境破坏的应用场合。客户如果在此类设备或应用场合中融入和 / 或使用 NVIDIA 产品，NVIDIA 不承担任何相关责任，风险由客户自行承担。在未经进一步测试或改动的情况下，NVIDIA 并不表示，也不担保基于这些规范的产品适合任何具体用途。每款产品所有参数的测试不一定由 NVIDIA 进行。确保产品适合客户所计划的应用场合并针对该应用场合进行必要的测试以避免应用场合出现问题或产品失灵，是客户单方面的责任。客户产品设计中的缺点可能会影响 NVIDIA 产品的质量和可靠性，并且可能会导致超出本规范以外的额外或不同的条件及要求。NVIDIA 不承担因下列情况造成失灵、损坏、成本或问题相关的任何责任：(i) 以违反本规范的方式使用 NVIDIA 产品或 (ii) 客户产品设计。本规范对 NVIDIA 专利权、版权或其他 NVIDIA 知识产权并未作出任何明示或暗示的许可。NVIDIA 所发布的有关第三方产品或服务的信息并不构成 NVIDIA 对于使用该产品或服务的许可，亦不构成担保或支持。使用此类信息可能需要获得第三方的专利权或其他知识产权的许可，或者需要获得 NVIDIA 的专利权或其他知识产权的许可。只有在获得 NVIDIA 书面批准的情况下才可以复制本规范中的信息，而且必须毫无改动地复制并附带所有相应的条件、限制条款和通知。

NVIDIA