WP504 (v1.0) 2018 年 10 月 2 日

使用赛灵思 Alveo 加速器卡加速 DNN

采用赛灵思 Alveo 数据中心加速器卡的赛灵思 xDNN 处理引擎是一款高性能高能效 DNN 加速器，在原始性能和功率效率（用于实时推断工作负载）方面优于当今众多常见的 CPU 和 GPU 平台。xDNN 处理引擎可通过 ML Suite 在众多云环境（例如 AWS EC2 或 Nimbix NX5）中使用。

摘要

Xilinx® 深度神经网络 (xDNN) 引擎使用 Xilinx® Alveo™ 数据中心加速器卡提供高性能、低时延、高能效的 DNN 加速。通过保持较低能源成本以及最大限度地减少实现过程中所需的特定加速器的数量，可以显著降低总体拥有成本 (TCO)。

赛灵思 Alveo 加速器卡在高性能、高能效、灵活的机器学习 (ML) 推断方面表现出色。xDNN 处理引擎已被开发用于一般执行卷积神经网络 (CNN)，如 ResNet50、GoogLeNet v1、Inception v4 - 甚至包含自定义层的 CNN。

本白皮书概述了 xDNN 硬件架构和软件堆栈，以及支持“业界一流”高能效推断声明的基准数据。

为读者提供指导，帮助其实现 Alveo 数据中心加速器卡上的结果再创造。

数据中心应用中的深度学习关联性

在过去几年中，深度学习方法在各种应用领域均取得了巨大成功。应用机器学习 (ML) 的一个领域是视觉和视频处理。互联网上的视频内容在过去几年中也迅速增长，对图像整理、分类和识别方法的需求也相应增长。

卷积神经网络 (CNN) 是 ML 神经网络的一种，其一直是处理图像数据的有效方法，尤其是在数据中心部署的背景下。CNN图像网络可用于对云中的图像进行分类和分析。在许多情况下，图像处理的时延对最终应用至关重要，例如标记流视频中的非法内容。

本白皮书介绍了赛灵思深度神经网络 (xDNN) 引擎，这是一款可编程推断处理器，其能够在赛灵思 Alveo 加速器卡上运行低时延、高能效的推断。xDNN 推断处理器通用 CNN 引擎，支持各种标准的 CNN 网络。xDNN 引擎通过赛灵思 xfDNN 软件堆栈集成到诸如 Caffe、MxNet 和 TensorFlow 等广受欢迎的 ML 框架中。在 Alveo 加速器卡上运行的 xDNN 处理引擎能够以每秒 GoogLeNet v1 吞吐量 4,000 或更多图像的速度进行处理，意味着在 Batch = 1 时超过 70% 的计算效率。

正如该计算效率所述，运行在赛灵思 Alveo 加速器卡上的 xDNN，在实现低时延推断方面优于 GPU 等加速平台。众所周知，GPU 平台能够通过同时批处理多个图像来提高其性能；然而，虽然批处理可以提高性能并减少所需的 GPU 存储器带宽，但批处理的副作用是时延显著增加。

相比之下，xDNN 处理引擎不依赖于批处理来实现最大化的吞吐量性能。而是每个引擎独立运行，并且不共享权重存储器。每个引擎在 Batch = 1 下运行，并且可以在单个 Alveo 加速器卡上实现多个引擎。因此，增加器件中 xDNN 引擎的数量仅增加了聚合器件 Batch = 1 的吞吐量。

xDNN 架构概览

xDNN 硬件架构如图 1 所示。每个 xDNN 引擎由一个脉动阵列、指令存储器、执行控制器和元素级处理单元组成。引擎通过指令队列从在主处理器上执行的命令软件接收张量指令。仅当目标网络改变时，CNN 网络的指令（张量和存储器操作）才改变。重复执行相同的网络会重复使用先前加载的驻留在指令缓冲器中的指令。

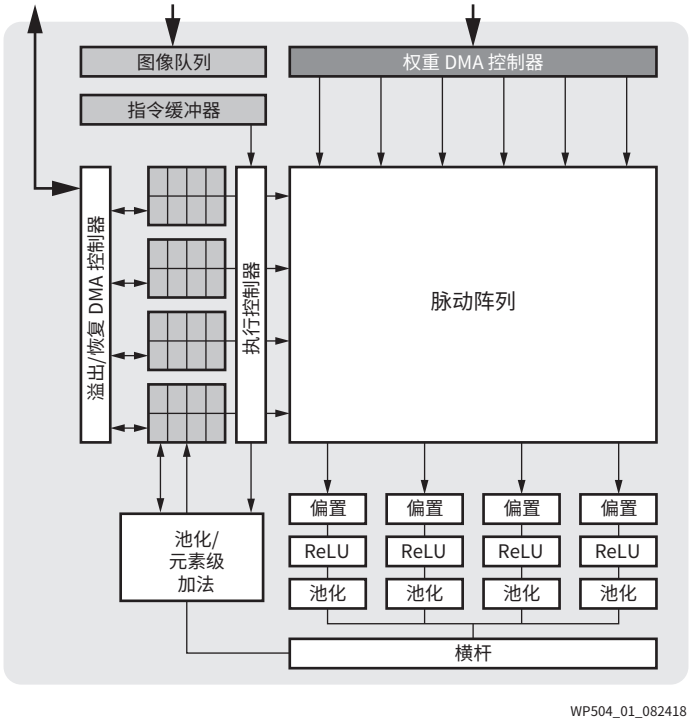


图 1： **xDNN 硬件架构**

xDNN 处理引擎架构亮点

- 双模式：吞吐量优化或时延优化
- 命令级并行执行
- 硬件辅助图像分块
- 自定义层支持（异构执行）
- 脉动阵列架构

吞吐量和时延优化模式

xDNN 处理引擎的架构特性之一是包括两种操作模式，一种用于吞吐量优化，另一种用于时延优化。在吞吐量优化模式中，通过创建优化的处理引擎 (PE) 来利用数据流并行性，以处理低效映射到一般脉动阵列的特定层。

例如，GoogLeNet v1 的第一层是 RGB 层，占整体计算开销的近 10%，不能有效地映射到有效计算网络其余部分的脉动阵列。在此吞吐量优化模式中，xDNNv3 包括为三个输入通道定制的其他脉动阵列。这种变化的网络效应是更高的整体计算效率，因为可以在先前图像卷积和 FC 层完成其各自的处理的同时，计算下一图像的第一层。

对于需要最低单图像时延的应用，用户可以选择部署时延优化版本的引擎。对于这些应用，可以调整 xDNN PE 流水线以减少时延。

命令级并行执行

xDNN 处理引擎为每种类型的命令（下载、Conv、池化、元素级和上传）提供专用的执行路径。如果网络图允许，则允许卷积命令与其他命令并行运行。某些网络图具有不同指令类型的并行分支，有时允许并行处理。例如，在 GoogLeNet v1 inception 模块中，3 x 3 最大池化层是一个层的主要示例，该层可以使用 xDNN 处理引擎与其他 1x1/3x3/5 x 5 卷积并行运行。图 2 显示了 GoogLeNet v1 网络的 inception 模块。

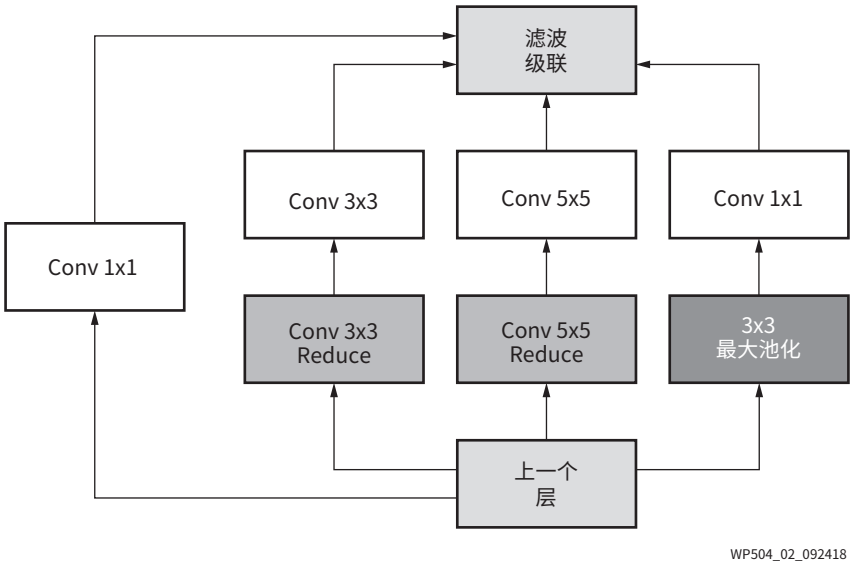


图 2： **GoogLeNet v1 中的 Inception 层**

如图 3 所示，软件可以与第二个分支的 3 x 3 卷积并行地调度 3 x 3 最大池化。

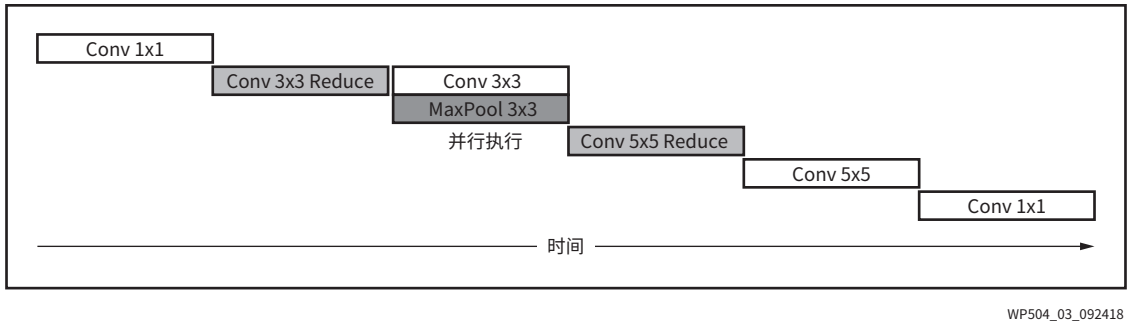


图 3： **GoogLeNet v1 中 Inception 层的 xDNN 调度**

硬件辅助图像分块

xDNN 处理引擎具有内置的硬件辅助图像分块功能，用来支持具有大图像/激活大小的网络。xDNN 处理引擎允许跨宽度和高度的输入特性映射分块。如图 4 所示。

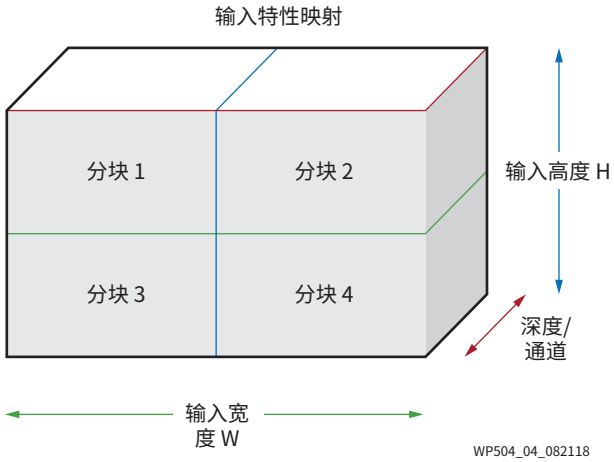


*图 4：**硬件辅助图像分块功能***

硬件辅助图像分块采用单个非数据移动指令 (Conv、Pool、EW) 并生成正确的微操作序列（下载、操作、上传）。通过将激活存储器逻辑分区为两个区域（如双缓冲区），微操作在硬件中完全实现流水线化。

通过异构执行自定义网络支持

尽管 xDNN 处理引擎支持广泛的 CNN 操作，但新的自定义网络仍在不断开发中 - 有时，FPGA 中的引擎可能不支持选择层/指令。由 xfDNN 编译器来识别 xDNN 处理引擎中不受支持的网络层，并且可以在 CPU 上执行。这些不受支持的层可以位于网络的任何部分 - 开始、中间、结束或分支中。

图 5 显示了编译器如何将处理划分到 xDNN 处理引擎甚至 CPU 中的各种 PE 上。

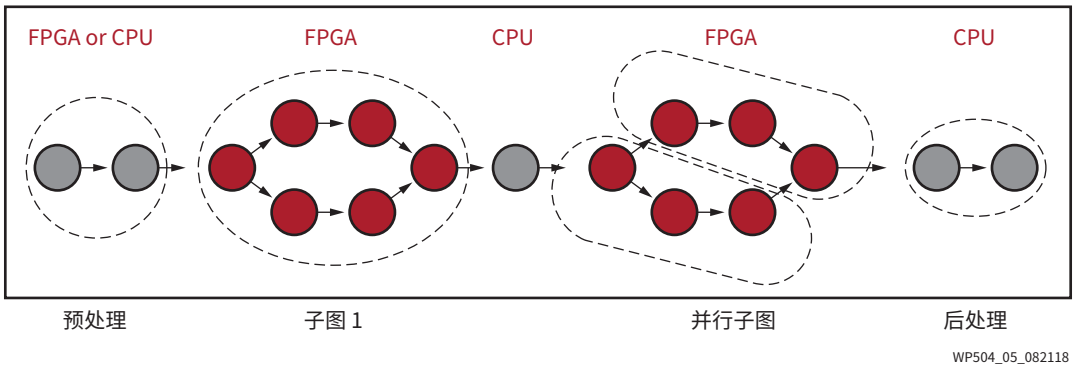


*图 5：**由编译器分区的处理***

脉动阵列架构

xDNN 处理引擎利用诸如“SuperTile”论文 [1] 中描述的技术来实现高工作频率。这个 SuperTile DSP 宏提供了一个关系放置的宏，其可以进行分块以构建更大的计算阵列，例如矩阵乘法和卷积，这是 CNN 计算最为密集的操作。

图 6 显示了映射到 FPGA 中的 DSP48 和 CLB-M (LUTRAM) 分块的逻辑处理元件的示例。该宏单元是 xDNN 脉动阵列中的基本处理单元。

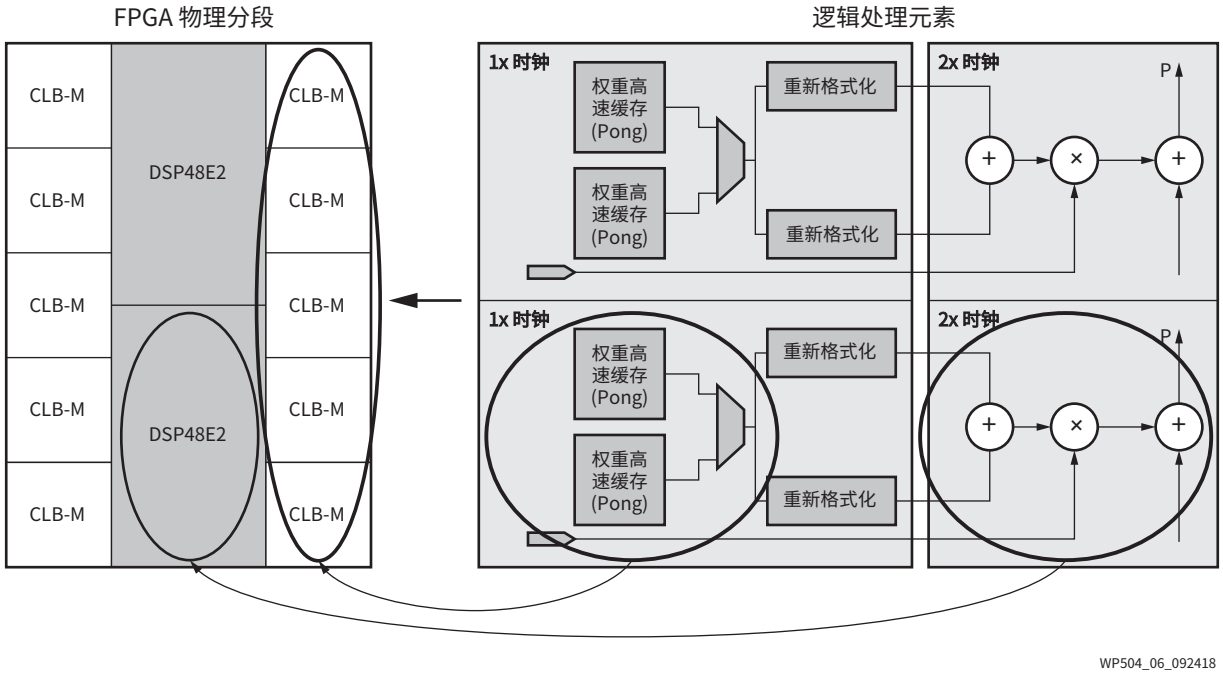


图 6： **DSP 宏示例中的 MAC 和权重包装**

1. E. Wu et al., Xilinx Inc. IEEEXplore Digital Library, Sept. 2017, *A High-Throughput Reconfigurable Processing Array for Neural Networks.*

xfDNN 软件堆栈概览

xfDNN 软件堆栈是软件工具和 API 的组合，其可通过各种常见的 ML 框架实现 xDNN 处理引擎的无缝集成和控制。图 7 中的流程图详细说明了如何准备网络和模型，以便通过 Caffe、TensorFlow 或 MxNet 在 xDNN 上进行部署。在 CPU 上运行不受支持的层的同时，xfDNN 编译器还支持 xDNN 层。在编译和量化网络/模型之后 - 该流程通常需要不到一分钟 -用户可以通过选择简单易用的 Python 或 C++ API 与 xDNN 处理引擎进行接口连接。

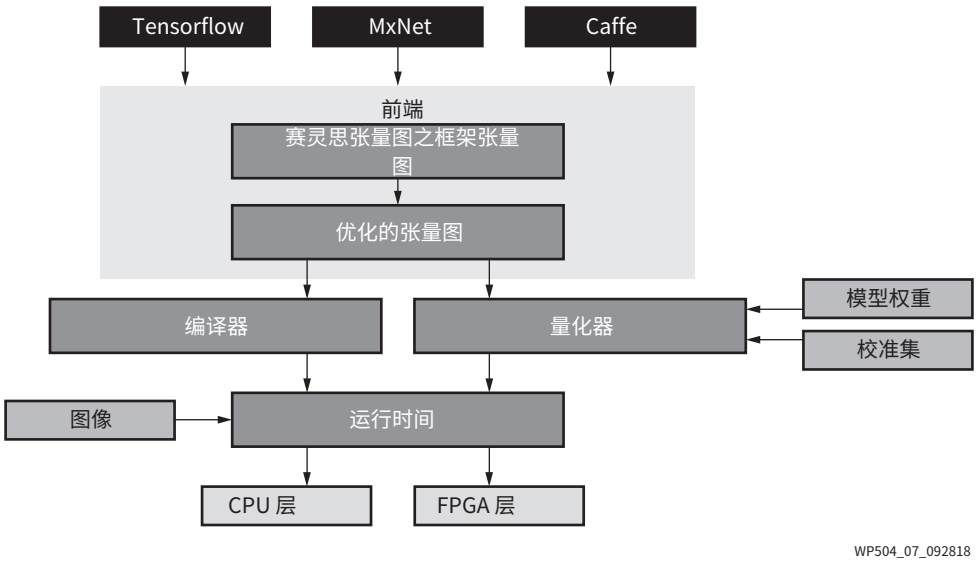


图 7： **xfDNN 流程图**

赛灵思 xfDNN 软件堆栈包括：

1. 网络编译器和优化器

编译器产生在 xDNN 引擎上执行的指令序列，其提供张量级控制和数据流管理，以实现给定的网络。

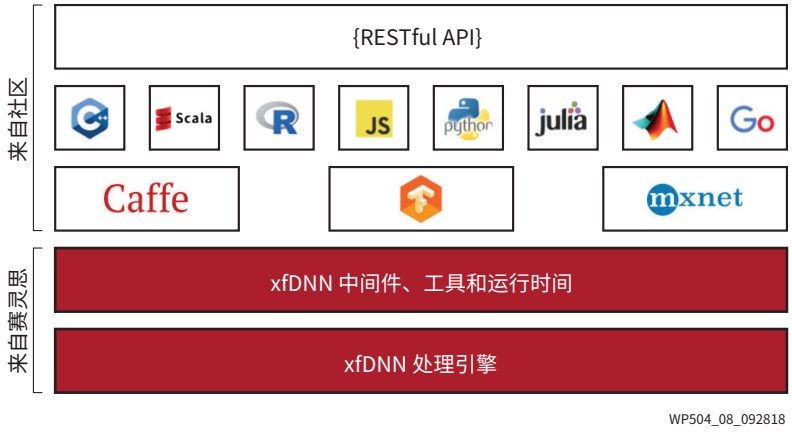
2. 型号量化器

量化器从经训练的 CNN 网络模型产生目标量化（INT8 或 INT16），而无需数小时的再训练或标记的数据集。

3. 运行时间和调度器

xfDNN 简化了 xDNN 处理引擎的通信和编程，并利用了符合 SDx 的运行时间和平台。

图 8 显示了 xfDNN 库的流程图，其将深度学习框架与在赛灵思 FPGA 上运行的 xDNN IP 相连接。



8： **xfDNN 软件堆栈**

有关 xfDNN 编译器的更多信息

现代 CNN 是数百个单独操作的图表，即卷积、Maxpool、Relu、偏置、批处理规范、元素级加法等。编译器的主要工作是分析 CNN 网络并生成在 xDNN 上执行的优化指令集。

xfDNN 编译器不仅提供简单的 Python API 来连接到高级 ML 框架，而且还通过融合层、优化网络中的内存相关性以及预调度整个网络来提供网络优化工具。这消除了 CPU 主机控制瓶颈。请参见图 9 作为示例。

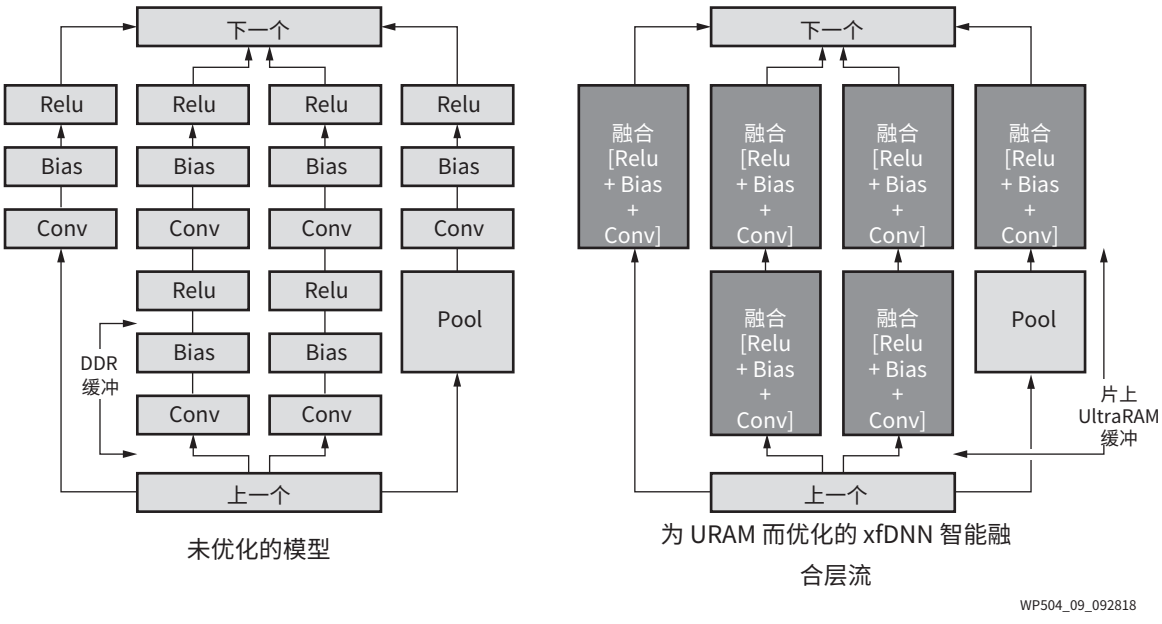
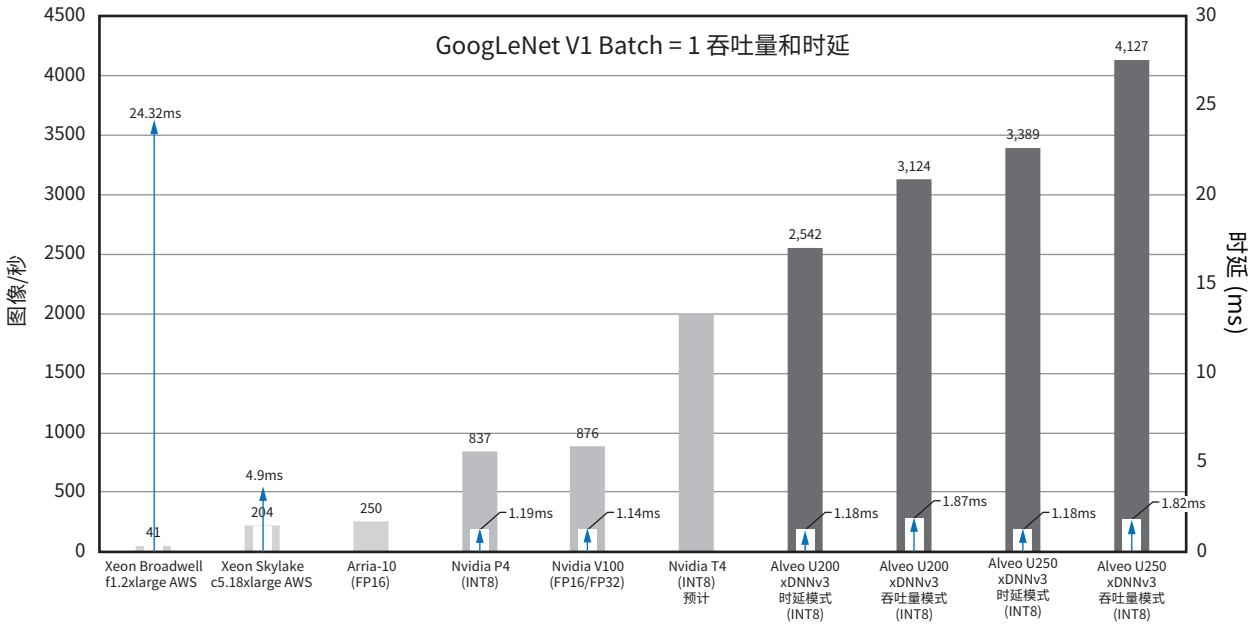


图 9： **xfDNN 编译器优化**

性能基准测试结果

随着实时 AI 服务的日益增多，时延成为整体 AI 服务性能的重要方面。GPU 在时延和吞吐量之间存在显著的权衡，与此不同的是，xDNNv3 DNN 引擎可以提供低时延和高吞吐量。此外，xDNNv3 内核提供简单的 Batch = 1 接口，无需任何排队软件来自动批量输入数据便可实现最大吞吐量，从而降低了接口软件的复杂性。

图 10 和图 11 显示了 Alveo 加速器卡和广受欢迎的 GPU 和 FPGA 平台上的CNN、时延和吞吐量基准。图 10 显示了沿 左 Y 轴以每秒图像数量来测量的 GoogLeNet V1 Batch = 1 吞吐量。吞吐量上方显示的数字是以毫秒为单位的测量/报告时延。



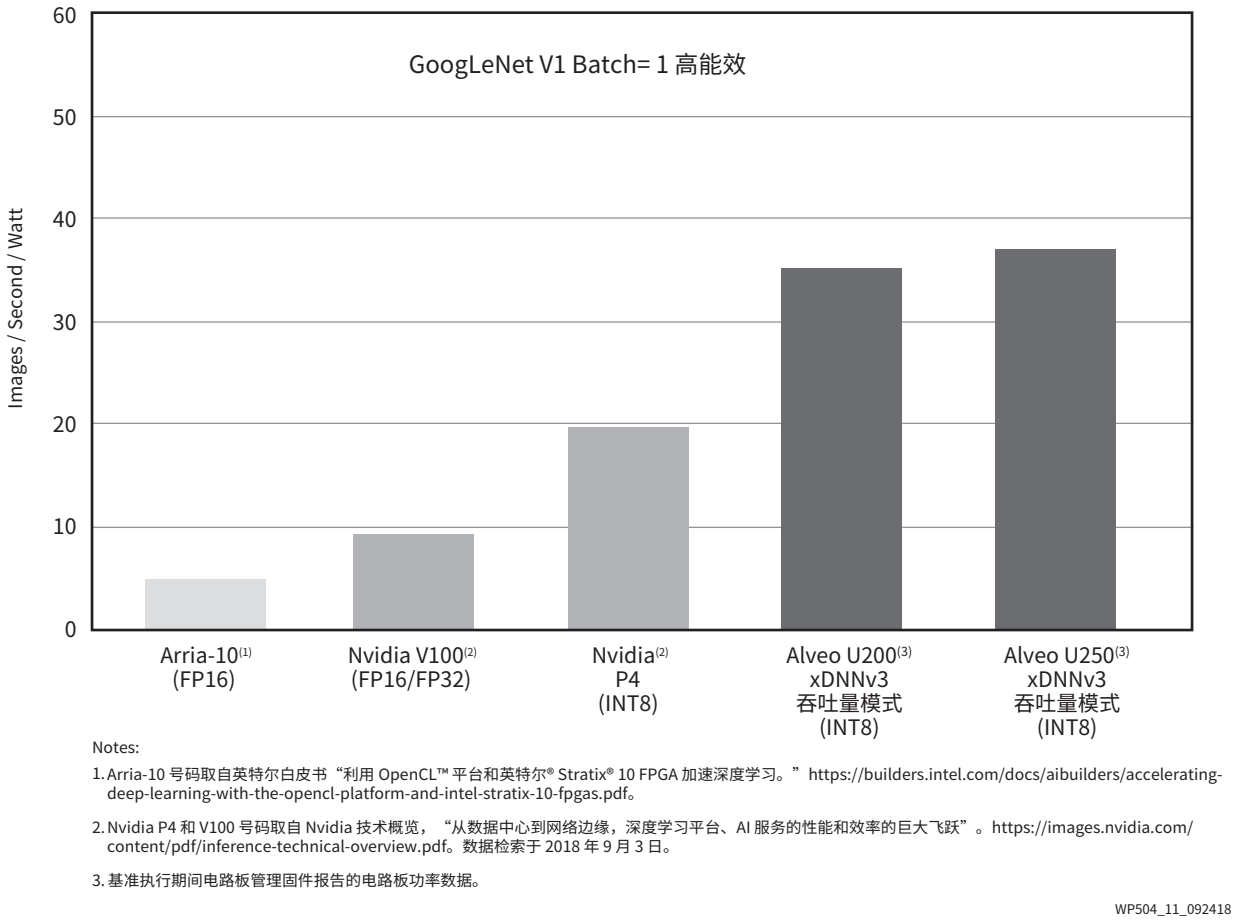
注释：

1. Xeon E5-2696 v4 f1.2xlarge AWS 实例，Ubuntu 16.04LTS，amd64 xenial 映像建于 2018 年 08 月 14 日，Intel Caffe (https://github.com/intel/caffe)，Git 版本：a3d5b02，run_benchmark.py w/Batch = 1 修改。
2. Xeon Platinum 8124 Skylake，c5.18xlarge AWS 实例，Ubuntu 16.04LTS，amd64 xenial 映像建于 2018 年 08 月 14 日，Intel Caffe，Git 版本：a3d5b02，run_benchmark.py w/Batch = 1 修改。
3. Arria-10 号码取自英特尔白皮书“利用 OpenCL™ 平台和英特尔 Stratix 10 FPGA 加速深度学习。”https://builders.intel.com/docs/aibuilders/accelerating-deep-learning-with-the-opencl-platform-and-intel-stratix-10-fpgas.pdf。Arria 时延数据尚未公布。
4. Nvidia P4 和 V100 号码取自 Nvidia 技术概览，“从数据中心到网络边缘，深度学习平台、AI 服务的性能和效率的巨大飞跃”。https://images.nvidia.com/content/pdf/inference-technical-overview.pdf。数据检索于 2018 年 9 月 3 日。
5. 基于当前可用的已公布基准的 Nvidia T4 投影。根据早期的功率效率基准，GoogLeNet Batch = 1 性能范围在 1700-2000 个图像/秒之间。
6. Alveo U200 数字测量 Intel Xeon CPU E5-2650v4 2.2GHz、2400MHz DDR4、Ubuntu 16.04.2 LTS 实例在 OpenStack Pike，Centos 7.4 上运行，预发布版本 MLSuite，streaming_classify.py，合成数据，MLSuite DSA Thin Shell，FC 和在 Xeon 主机上运行的 SoftMax 层和不包含在计算总计中的操作（占总计算的 0.06%）。
7. Alveo U250 数字测量 Intel Xeon Silver 4110 CPU @ 2.10GHz，CentOS Linux 发布 7.4.1708，预发布版本 MLSuite，streaming_classify.py，合成数据，DSA：Thin Shell，FC 和在 Xeon 主机上运行的 SoftMax 层和不包含在计算总计中的操作

WP504_10_092418

图 10： **GoogLeNet v1 Batch = 1 吞吐量**

图 11 显示了沿 Y 轴以每秒每瓦特图像数量来测量的 GoogLeNet V1 吞吐量。



Notes:

1. Arria-10 号码取自英特尔白皮书“利用 OpenCL™ 平台和英特尔® Stratix® 10 FPGA 加速深度学习。” https://builders.intel.com/docs/aibuilders/accelerating-deep-learning-with-the-opencl-platform-and-intel-stratix-10-fpgas.pdf。
2. Nvidia P4 和 V100 号码取自 Nvidia 技术概览，“从数据中心到网络边缘，深度学习平台、AI 服务的性能和效率的巨大飞跃”。https://images.nvidia.com/content/pdf/inference-technical-overview.pdf。数据检索于 2018 年 9 月 3 日。
3. 基准执行期间电路板管理固件报告的电路板功率数据。

WP504_11_092418

图 11： **GoogLeNet v1Batch= 1 能源效率**

虽然 GoogLeNet v1 性能可用于基准测试，但 xDNN 支持广泛的 CNN 网络。如需了解有关运行其他 CNN 网络的更多信息，请参阅 ML Suite 文档 (https://github.com/Xilinx/ml-suite)。

结论与行动呼吁

如共享性能结果所示，xDNN 处理引擎是一种高性能、高能效的 DNN 加速器，在实时推断工作负载方面优于当今众多常见的 CPU/GPU 平台。xDNN 处理引擎可通过 ML Suite 在众多云环境（例如 Amazon AWS/EC2 或 Nimbix NX5）中使用。其通过赛灵思的新 Alveo 加速器卡无缝扩展到本地部署。

赛灵思的可重配置 FPGA 芯片允许用户通过 xDNN 更新继续接收新的改进和功能。这使用户能够跟上不断变化的需求和不断演进发展的网络。

如需了解有关入门的更多信息，请访问：https://github.com/Xilinx/ml-suite 或 https://www.xilinx.com/applications/megatrends/machine-learning.html

修订历史

下表列出了本文档的修订历史。

日期	版本	修订描述
10/02/2018	1.0	赛灵思初始版本。

免责声明

关于与汽车相关用途的免责声明