白皮书

物联网
采用英特尔® Arria® 10 FPGA 的
英特尔® 视觉加速器设计

借助采用英特尔® Arria® 10 FPGA 的英特尔® 视觉加速器设计，提升视频分析的性能

借助集成的深度学习推理构建高性能计算机视觉应用

概述

人工智能（AI）正在推动计算领域的新一轮浪潮，不仅引领着业务运营的深刻变革，还将彻底改变人们生活的方方面面。机器学习和深度学习是人工智能的重要组成部分，可使用数据训练模型和构建推理引擎。这些引擎应用训练后模型进行数据分类、识别和检测。低延迟解决方案支持推理引擎更快速处理数据，提高系统的总体响应速度以进行实时处理。

计算机视觉正快速成为关键人工智能数据的重要来源，催生着物联网（IoT）在智慧城市、零售等领域的新应用和用例。为满足实时分析和高成本效益处理的需求，企业必须在“边缘”完成许多数据处理工作，包括设备、内部服务器、云端和数据中心等。英特尔® 视觉加速器设计为特定的推理加速器卡提供了全新系列的“蓝图”，以支持从边缘到云端的人工智能计算机视觉解决方案和深度神经网络推理。

英特尔正在推动制定一项开放软件标准，以便为开发人工智能计算机视觉应用和识别适当的芯片技术提供一种通用框架。

采用英特尔® Arria® 10 FPGA 的英特尔视觉加速器设计为开发人员和解决方案提供商带来了出色的加速性能和灵活性，可帮助他们更快地实现成效和产品上市。该产品适用于创建支持边缘视频分析和许多物联网用例的应用。

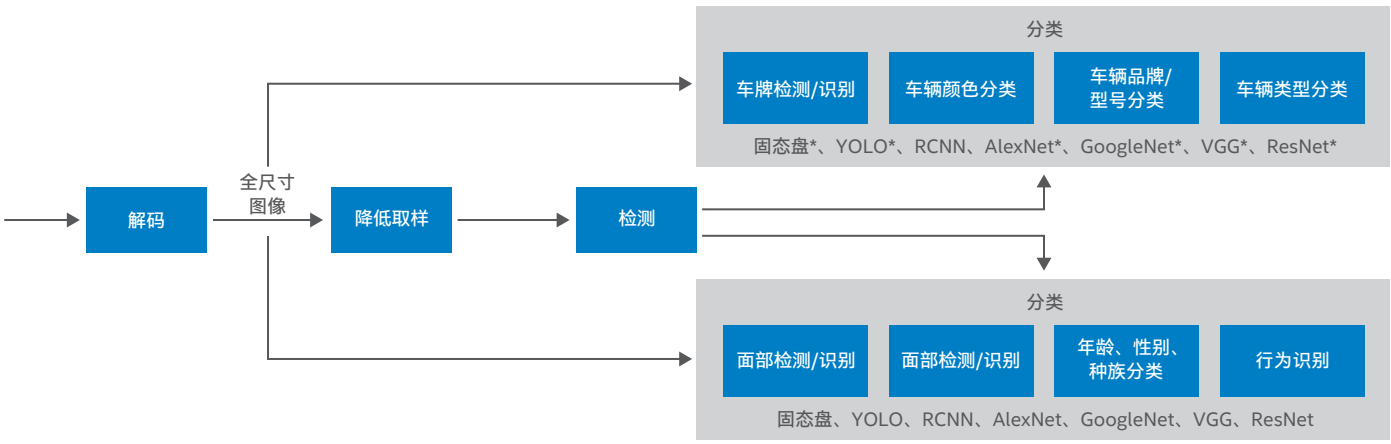


图 1. 采用英特尔® Arria® 10 FPGA 的英特尔® 视觉加速器设计支持深度神经网络推理用于人工智能用例

边缘人工智能计算机视觉
（边缘设备或内部服务器）

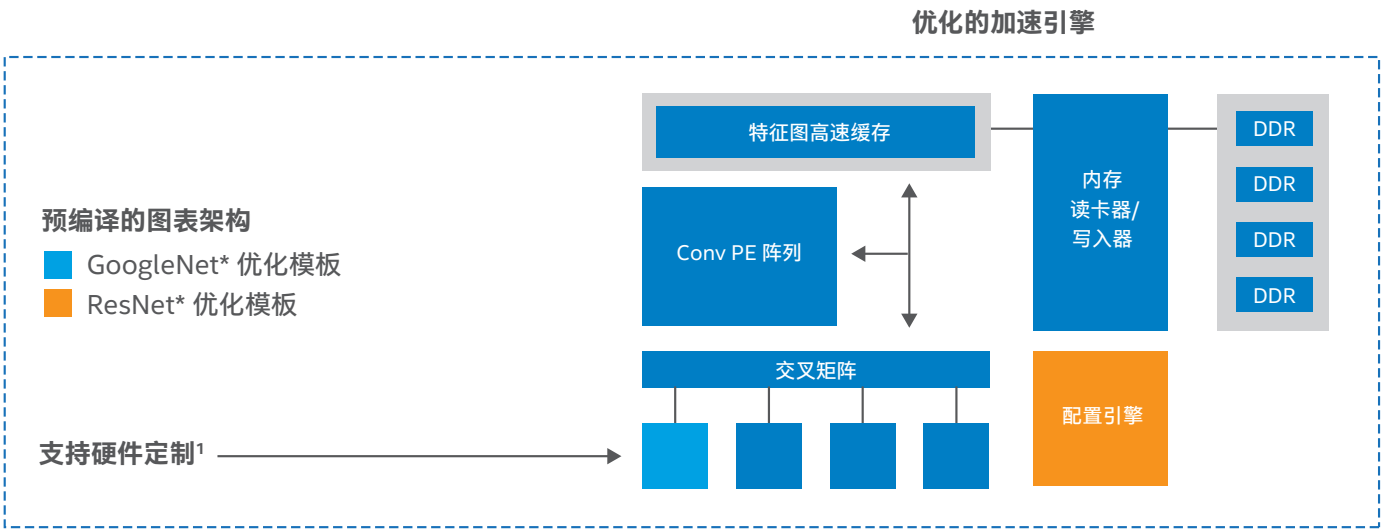
- 大容量就绪型 PCIe* 卡具备强大的 FPGA 加速能力，提供了简单的模块化应用开发方法。
- 系统级优化 — 可定制数据路径和推理精度可带来节能型数据流。
- 为低延迟系统而优化。
- 精细的并行化，支持低批量工作负载实现高吞吐量。
- 提高工作效率，支持跨英特尔凌动® 处理器和英特尔® 酷睿™ 处理器的快速代码复用。

云端人工智能计算机视觉

- API 和驱动程序可帮助软件开发人员在更高的抽象化层面为采用 FPGA 平台的英特尔® 至强® 可扩展处理器进行编程。
- 英特尔® FPGA 与英特尔® 至强® 可扩展处理器相结合，提供占用空间小的低延迟实施。
- 提高工作效率，支持跨英特尔® 至强® 可扩展处理器的快速代码复用。

采用英特尔 Arria 10 FPGA 的英特尔视觉加速器设计的优势包括：

- **优化边缘视觉分析：**在内部和本地装配视觉分析系统以支持各种最终用途。
- **高效的视频加速：**分析多路边缘视频流，同时保持出色的成本效益（较高的单位功耗成本帧速），边缘环境并非总能实现温度控制。
- **降低总体拥有成本：**开发功耗更低、能在许多散热环境中可靠运行更长时间的解决方案。
- **交付多功能解决方案：**FPGA 采用了多功能设计。借助英特尔视觉加速器设计，开发人员可简化上市路径，即时支持强大功能，获得功能编码和定制的选项。
- **加速视频解码和加密：**借助多路摄像头数据流的快速同步处理，加速实时解码和编码。借助英特尔® FPGA，软 IP 可提供实时性能，专用解码器的数量可视情况增加或减少以提高效率。



需要加密的 DLA 源代码许可；单独销售

图 2. 对更多拓扑和算法的支持为开发人员提供了更多选择

采用英特尔 Arria 10 FPGA 的英特尔视觉加速器设计

可编程的软件定义英特尔 FPGA 可为深度学习和计算机视觉解决方案提供卓越的性能和灵活性，以及较低的功耗，无论是在边缘设备、内部服务器还是云端。英特尔 FPGA 可帮助确保连续的性能优化，能够充分利用位流更新，可以避免硬件升级。该产品采用了长生命周期设计，可应对恶劣和/或室外的环境。

该解决方案可与 OpenVINO™ 工具套件无缝配合，支持开发人员轻松运行可为异构硬件平台优化的定制拓扑。英特尔 FPGA 专为高效处理数据而设计，能够摆脱操作系统和应用所带来的烦扰。精细的并行化和高吞吐量功能内置到硬件架构中，可帮助众多工作负载实现极低批量延迟。极高的精细片上内存带宽能够更高效地化解内存挑战。

FPGA 的软件 IP 支持实时解码和加密（软件在 GPU 或 CPU 上运行无法提供该支持），能够以更低功耗实现出色的每秒图像性能，并提供动态灵活性、一致功耗、面向定制或新工作负载的未来支持及低延迟。

不同于固定功能设备，英特尔 FPGA 的功能可随时更改或修改，以提升或强化智能性。因此，英特尔 FPGA 可通过相应设计帮助解决特定的问题。例如，当某个问题正在消耗带宽并损害性能时，该产品可加速整个系统。FPGA 支持广泛的视觉用例和应用。英特尔已将 FPGA 定制为卷积神经网络（CNN）工作负载的高性能加速器，减轻了开发人员的负担。

英特尔 FPGA 卡可支持超过 40 个视频通道以及丰富视觉用例，如面部和车辆检测、车牌识别（LPR）、车辆前灯位置估计、面部视频分析及人员和汽车视频分析。支持的网络拓扑包括 GoogleNet*、ResNet*、SqueezeNet*、VGG-16* 和 MobileNet*。可定制的数据路径和推理精度为系统级优化创建了节能型数据流。

采用英特尔 Arria 10 FPGA 的英特尔视觉加速器设计适用于复杂或较大的工作负载和定制应用；您希望添加自有基元或子层配置的用例；以及英特尔® Movidius™ Myriad™ X VPU 并不支持的工作负载。

FPGA 本身具有出色的适应性，能够率先支持新解决方案和不断变化的网络拓扑。凭借出色的灵活性和加速算法处理的能力，采用英特尔 Arria 10 FPGA 的英特尔视觉加速器设计在复杂的大规模深度学习分析和视觉智能方面大有作为。

采用英特尔 Arria 10 FPGA 的英特尔视觉加速器设计，支持越来越多的通用拓扑和算法用于人工智能计算机视觉解决方案。借助该产品，开发人员可在完成试验后轻松添加算法和拓扑。

拓扑
GoogleNet*
ResNet–18*
ResNet-50*
ResNet-101*
SqueezeNet*
SqueezeNext*
VGG-16*
DenseNet*
MobileNet*
Tiny YOLO*
SSD300*
SSD512*

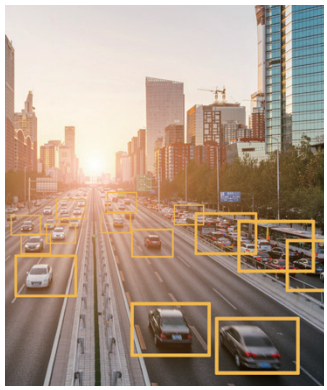
算法	
面部检测	面部选择
面部识别	手部追踪
面部特征分类	立体匹配
人员检测	摄像头姿态
人员跟踪	3D 重建
人员特征	人员再识别
年龄识别	可视化即时定位和地图创建（SLAM）
性别识别	变更检测
对象检测	多摄像头跟踪
对象追踪	传感器融合
物体识别	光学字符/单词识别
多目标跟踪	行为检测
身体检测	手势识别
身体识别	活动识别
	遗弃对象识别

示例用例

零售

面部和对象识别用于识别顾客的身份等信息，如年龄和性别、在店内停留的时间、回头客和 VIP 顾客等。顾客行为数据可用于调整商品组合及优化热图，检测商品的存放和库存情况，帮助预防缺货，以及改进顾客的逛店体验。

交通管理



车牌识别（LPR），人员身份、性别和年龄识别及行为检测被用于从交通模式分析到执法的广泛领域。

智能城市

城市纷纷使用基于多路摄像头数据流的视频分析，以实时检测和跟踪相关人员，改善公共安全状况。异常检测可发现公民是否遇到麻烦，及是否需要医疗救助或紧急帮助。

工业 4.0

制造工厂开始使用多颗摄像头对车间运营和/或资源进行目视检查。

实现最大性能，最大限度缩短开发时间。

采用英特尔 Arria 10 FPGA 的英特尔视觉加速器设计可与 OpenVINO 工具套件相结合，加速开发高性能计算机视觉和深度学习应用。OpenVINO 工具套件能够提升计算机视觉解决方案的性能，并缩短其开发时间。该套件可帮助用户更轻松地获得各种英特尔硬件选项的优势，从而提升性能、降低功耗并显著提高硬件利用率，让您实现事半功倍之效，开辟新的设计可能性。

OpenVINO 工具套件

OpenVINO 工具套件可支持硬件加速器上的深度学习，及多种英特尔® 平台上的简化异构执行。它包括采用模型优化器和推理引擎的英特尔® 深度学习部署工具套件，以及面向 OpenCV* 和 OpenVX* 的优化的计算机视觉库和函数。这款全面的工具套件可支持广泛的视觉解决方案，可加快计算机视觉工作负载的处理速度，简化深度学习部署并在从设备到云的英特尔平台上实现轻松的异构执行。

使用 OpenVINO 工具套件和英特尔® 架构，将公共模型上的深度学习工作负载性能提升高达 19.9 倍。[1]

- **提升性能：** 充分利用英特尔计算机视觉加速器，增强代码性能。支持异构处理和异步执行。
- **整合深度学习：** 使用通用 API 和超过 40 个预训练模型，发掘基于卷积神经网络（CNN）的深度学习推理潜力。
- **加速开发：** 借助优化的 OpenCV 和 OpenVX 功能与预制示例库，缩短开发时间，并充分利用通用算法。
- **编写一次：** 仅需开发一次，即可部署到当前和未来的英特尔架构设备。
- **创新和定制：** 使用 OpenCV 中不断扩展的资源库添加您自己的独有代码。

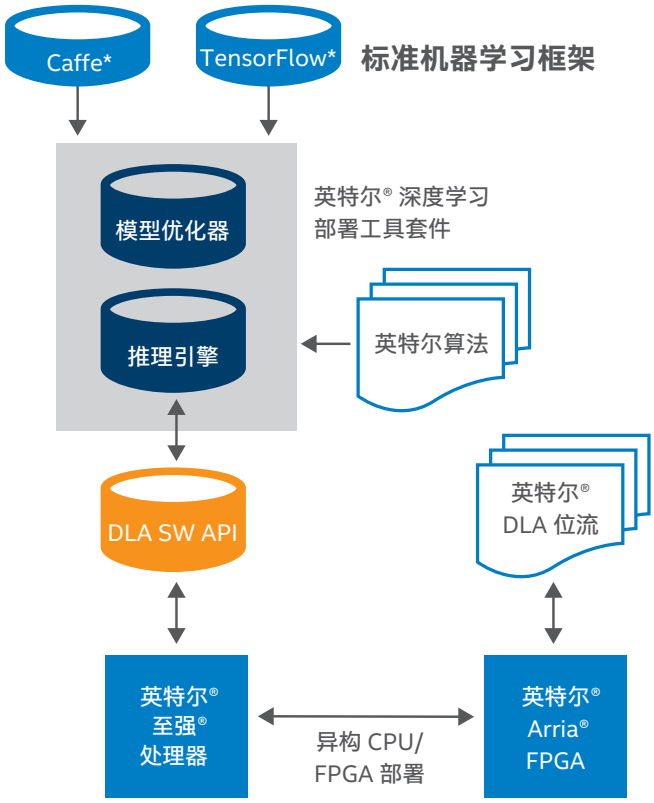


图 3. OpenVINO™ 工具套件可帮助开发人员进一步优化深度学习推理的性能

英特尔® 深度学习部署工具套件

OpenVINO 工具套件包括英特尔深度学习部署工具套件。通过输入来自标准网络的训练后模型，如 Caffe*、TensorFlow* 和 MXNet*，模型优化器可将其转换为统一的中间代码（IR）文件，然后在推理引擎上运行该文件。使用标准层或用户提供的自定义层时，无需加载 Caffe 或其他框架。推理引擎的 API 提取硬件，通用于各类硬件。这有助于跨不同加速器进行测试，无需重新编码。此外，该推理引擎还支持从 FPGA 上的自定义层恢复至 CPU，可实现异构性。

基本上，模型优化器可针对性能和空间进行优化，支持传统拓扑变换。该推理引擎接口用作每个硬件类型的动态加载插件，可为相应类型提供最佳性能，无需实施和保持多个代码路径。

提升深度学习推理性能

- 模型优化器可将 Caffe、TensorFlow 和 MXNet 转换为 IR 文件
- 具有 CPU、GPU、FPGA 和 VPU 插件的推理引擎

OpenCV

- 具有英特尔® CPU 优化的预编译 OpenCV 3.3
- 英特尔® 图像视觉库，具有面部检测/识别、眨眼检测和微笑检测功能

OpenVX

- 针对简短的传统计算机视觉操作和 CNN 基元列表进行基于图形的实施
- Khronos OpenVX* 神经网络扩展 1.2
- 视觉算法设计器（VAD）
- Eclipse* 插件支持集成的 OpenVX 应用开发

OpenCL™

- 包括驱动程序、运行时和媒体驱动程序，可简化英特尔® 媒体 SDK 及面向 OpenCL™ 的英特尔® FPGA SDK 应用在计算机视觉领域的使用

模型

- 除了可下载公共模型的模型下载器，该封装包包括 40 多个预训练模型和示例

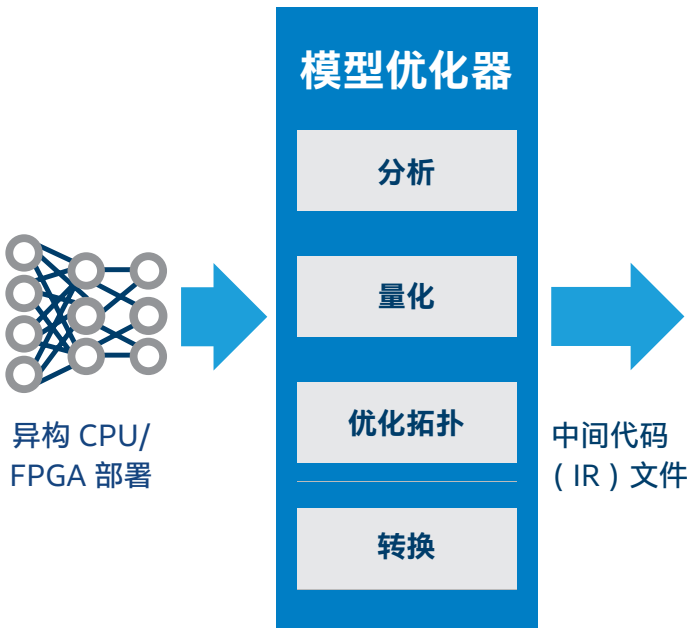


图 4. 模型优化器可将 Caffe*、TensorFlow* 和 MXNet* 等深度学习框架转化为中间代码（IR）文件以支持高效处理

主要规格

面向深度学习的英特尔 Arria 10 FPGA

- 英特尔 Arria 10 1150GX PCIe 插卡，带有 OpenVINO™ 工具套件支持和位流更新工具
- 提供各种推理精度选项，帮助使用相同硬件支持不同的性能目标
- 定期位流更新可不断增强性能

硬浮点，任意精度 DSP 模块

- 英特尔 Arria 10 FPGA 可实现高达 1.5 TFLOPS，英特尔® Stratix® 10 FPGA 可实现高达 10 TFLOPS
- 任意精度数据类型（FP16 => FP9）在英特尔 Arria 10 上可实现 2-20 TOPS（最高 26 TOPS）

分布式精细 DSP、内存和逻辑

- 片上带宽比 GPU 高出多个数量级
- 更少的数据移动有助于提升功效
- 确定性低延迟

内核与 I/O 可编程性有助于提升灵活性

- 任意深度学习架构
- I/O 配置和使用模式的多功能性
- 多功能加速

生态系统解决方案

越来越多的生态系统合作伙伴利用英特尔® 可编程加速卡（英特尔® PAC），及采用英特尔 Arria 10 FPGA 的英特尔视觉加速器设计，提供应用定制解决方案。

IEI 视频分析加速器卡*

- 专为高性能、低延迟应用而设计
- 支持多种网络拓扑（GoogleNet、ResNet 等）
- 面向加速器定制的 OpenCL BSP
- 英特尔 Arria 10 GX 1150KLE FPGA
- 两个内存条（DDR4，每个 4 GB）
- 被动或主动冷却
- ½ 长、½ 高 PCIe（Gen3 x8，x8 机械）

图 5. IEI 和 QNAP 借助采用英特尔® Arria® 10 FPGA 的英特尔® 视觉加速器设计提供解决方案

英特尔视觉加速器设计产品组合

借助全新系列的硬件加速主板，英特尔可为边缘深度学习推理提供更多硬件加速选择。英特尔视觉加速器设计还采用英特尔 Movidius Myriad X VPU。

您可根据自己的应用和创新需求，单独或同时使用 Movidius 或 FPGA 解决方案。借助英特尔视觉加速器设计产品，英特尔正帮助消除从芯片到软件堆栈的障碍。

	采用英特尔® MOVIDIUS VPU 的英特尔® 视觉加速器设计	采用英特尔® ARRIA® 10 FPGA 的英特尔® 视觉加速器设计
	以超低功耗实现高性能机器视觉和感知的专用处理器。	*旨在支持高级显示、视频和图像处理工作负载的灵活可定制处理器。*
按需选择	• 出色的计算能力和效率、低功耗及卓越的计算机视觉和深度学习性能功耗成本比。	• 基于单个芯片的动态灵活性和高性能，一致的功耗，未来对定制或新工作负载的支持以及低延迟。 • 需要更多深度学习优化特性，如计算密集型网络（VGG*、ResNet-101*），和全新网络拓扑。
用例	• 在具有功耗、大小和/或成本限制的摄像头和视频设备用例，以及可优化到 ASIC 中的主流拓扑中表现出色。	• 在需要软 IP 及近即时的功能修改与高性能的视频设备和服务器用例中表现出色。
配置	• 面向摄像头和其他功耗/大小受限的系统。	• 面向内部视频设备和视频分析服务器。
流支持	• 通常支持每 VPU 2 路视频流。	• 每设备聚合 3-32 路视频流。
批次大小	• 批次大小 = 1+。	• 批次大小 = 1+。
功耗	• 每 VPU 功耗通常为 2W 到 3W。（设计范围在大约 4W 到 25W 之间）。	• 更一致的功耗（英特尔® Arria® 10 1150 设备通常约为 42W）。
效率	• 出色的效率（每秒每瓦的推理吞吐量）。	• 良好的效率（每秒每瓦的推理吞吐量）（对比 GPGPU）。
网络内存	• 小于 2.5 亿个参数。	• 大于 2.5 亿个参数。
精度	• 支持 fp16 精度网络。	• 支持较低精度的网络（如 FP16/11/9）。
定制	• 针对通用案例进行了硬件优化。	• 面向特定拓扑的定制硬件架构。
分区	• 对于大型网络（如 YOLO* v1），需要跨多个设备分区。	• 无需跨多个硬件设备分区。

结论

采用英特尔 Arria 10 FPGA 的英特尔视觉加速器设计，为开发支持人工智能计算机视觉解决方案的软件应用提供了一种高效且成本效益出色的平台。该产品拥有英特尔 FPGA 的卓越性能和灵活性，可与 OpenVINO 工具套件协同优化基于 CNN 的深度学习推理。采用英特尔 Arria 10 FPGA 的英特尔视觉加速器设计是加速和改善广泛边缘分析应用的英特尔解决方案之一。轻松采用支持关键芯片技术的开放软件环境，其强大性能可完美满足您的需求。

英特尔可帮助开发人员和解决方案提供商降低总体拥有成本，实现更出色的价值，在基于人工智能的视频分析领域发掘无限商机。

了解更多信息

探索英特尔® 视觉产品：intel.com/visionproducts

了解英特尔 FPGA：intel.com/fpga

详细了解英特尔人工智能创新：https://www.intel.cn/content/www/cn/zh/analytics/artificial-intelligence/overview.html

下载免费的 OpenVINO 工具套件。

联系 IEI 下单，获取支持请发送邮件至：sales@usa.ieiworld.com

开发建议

对于网关、边缘或端点的分布式推理，您可以先从英特尔® CPU 入手，然后添加目标 FPGA 加速功能，以实现更高的吞吐量和/或每瓦吞吐率。集成的英特尔® 处理器显卡是进一步提升吞吐量的常用工具。FPGA 支持在很多方面进行平台定制，包括 I/O、多流聚合、内嵌处理以及深度学习和传统传感器处理加速的组合等。

1. 与英特尔® 深度学习部署工具套件中的某些标准框架模型和英特尔优化模型相比，性能有大幅提升。

在特定系统中对组件性能进行特定测试。硬件、软件或配置的任何差异都可能影响实际性能。当您考虑采购时，请查阅其他信息来源评估性能。如欲了解有关性能及性能指标评测结果的更完整信息，请访问：https://www.intel.cn/content/www/cn/zh/benchmarks/benchmark.html

随着更多测试的开展，报告的估计结果可能会进行修改。结果取决于测试中使用的具体平台配置及工作负载，可能不适用于任何特定用户的组件、计算机系统或工作负载。结果可能无法代表其他性能指标评测，其他性能指标评测可能会显示更大或更小的影响。

在性能检测过程中涉及的软件及其性能只有在英特尔微处理器的架构下方能得到优化。SYSmark* 和 MobileMark* 等性能测试采用特定的计算机系统、组件、软件、操作和功能进行测量。上述任何要素的变动都有可能导致测试结果的变化。您应当参考其它信息和性能测试以帮助您完整评估您的采购决策，包括该产品与其它产品一同使用时的性能。如需了解关于性能指标评测和性能测试结果的更多信息，请访问：http://www.intel.cn/content/www/cn/zh/benchmarks/benchmark.html

英特尔技术的特性和优势取决于系统配置，可能需要支持的硬件、软件或服务激活。性能会因系统配置的不同而有差异。没有任何计算机系统能保证绝对安全。请联系您的系统制造商或零售商，或访问：intel.com/visionproducts

描述的成本降低方案旨在作为举例，说明指定的英特尔架构产品在特定环境和配置下，可能如何影响未来的成本和提供成本节省。环境将有所不同。英特尔不保证任何成本和成本的节约。

1018/BH/CMD/PDF 338179-001CN