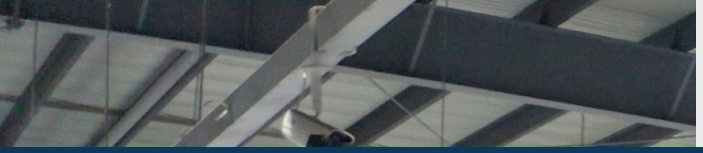
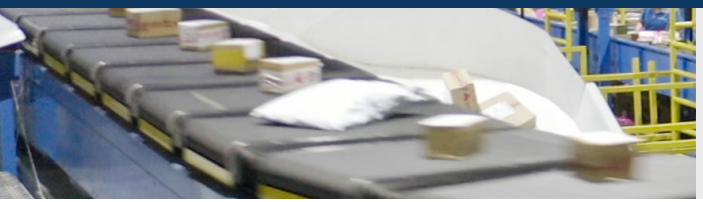


案例简介
物流
边缘视觉 AI 推理
英特尔® 数据中心 GPU Flex 系列
英特尔® 分发版 OpenVINO™ 工具套件

中通快递采用英特尔® 数据中心 GPU 和 OpenVINO™ 以更高性价比扩展边缘视觉 AI 应用



需求与挑战

追求更先进、易用的技术方案来优化物流园区作业、运输和管理，是中通快递（下文简称为“中通”）积极进步的宗旨。中通在早期部署的边缘视觉 AI 方案，就能有效监测园区内是否存在攀爬传送带等危险作业，踩踏、暴力分拣等违规作业，以及未戴安全帽等着装安全问题。但随着业务的快速发展以及技术应用的不断深入，中通对边缘视觉 AI 方案提出了更高的要求：

- **满足业务端的更多需求：**在中通，越来越多的业务场景开始利用 AI 技术。例如：在场区的分拣方面，作业流程需借助 AI 技术来识别和分析分拣过程中可能存在的小件堵包、流水线拥堵和挂包等情况，以做到“实时发现、实时告警”，进而降低错分率并减少因错分造成的错派、人工核对成本增加和时效性低等问题；而在装卸车管理方面，则需要利用 AI 技术来检测车辆到达与发车的准时性、装载率以及装卸车作业规范等，以保证时效和降低运输成本。对此，AI 方案需要更积极地做出快速响应，从而满足这些新需求。
- **模型开发与维护需要更简洁、更高效：**目前市场上的既有边缘视觉 AI 方案大多基于 ARM 架构，但中通可利用的服务器多为 x86 架构。这就要求开发人员不仅要解决将基于前者的方案适配到 x86 架构过程中面临的编解码问题，后期还要同时维护两个生态，耗时耗力又复杂。因此，实现统一且成体系的模型开发是当前技术层级的关注重点之一。
- **降低成本，在实际场景中实现更高性价比：**边缘视觉 AI 应用涉及大量的 AI 推理工作，如果通过添置更多 GPU 来确保推理性能，成本会过于高昂。因此，在看到方案的切实成果前，需要控制技术和实施成本，以较高的投产比来满足业务需求，也是当前亟需解决的问题之一。

解决方案与成果

基于异构基础设施，利用 XPU 实现 AI 推理加速

中通众多中心或网点都配备了 x86 服务器，同时也部署了新的英特尔® 数据中心 GPU Flex 系列。在英特尔工程师的协助下，中通只需在相同模型上进行开发，即可基于 XPU 实现 AI 推理加速，从而实现对各种资源的充分利用。例如，同一模型，在对轻量级 AI 业务场景时，可以直接使用 CPU，而在对实时性要求较高或者多并发的场景时，则使用英特尔® 数据中心 GPU Flex 系列，从而减少针对不同硬件开发不同模型的负担，降低全网部署的难度。

借英特尔® 分发版 OpenVINO™ 工具套件简化开发与运维

英特尔® 分发版 OpenVINO™ 工具套件是一个旨在优化和部署 AI 推理的开源工具套件。中通利用其中的模型优化器可将基于其他深度学习框架的模型转换为统一且性能经过优化的 OpenVINO™ IR 模型，有效降低了模型优化与运维的复杂程度。

其次，此工具套件中的 Open Model Zoo 提供了大量的免费且预训练好的深度学习模型及演示应用。中通在此次项目中也选用了其中的一些模型，有效地降低了模型开发难度并缩短了应用开发时间。

同时，中通还利用了工具套件中的英特尔® Deep Learning Streamer（DL Streamer）[1] 并结合自身应用场景的特点创建了用于视频解码、编码和媒体智能分析的业务流，实现了在边缘对音视频进行智能分析和对英特尔® 硬件平台的充分利用。

以英特尔® 数据中心 GPU Flex 系列应对更严苛的需求

为满足像视频流计算这样对算力和实时性要求较高的应用需求，中通按需导入了英特尔® GPU Flex 系列 170，对部分服务器进行了升级改造并针对其 ZTO Yolo v4 推理业务流进行了测试。测试结果（如图 1 所示）显示，这一产品性能出色，可很好地满足中通相关应用场景的需求。

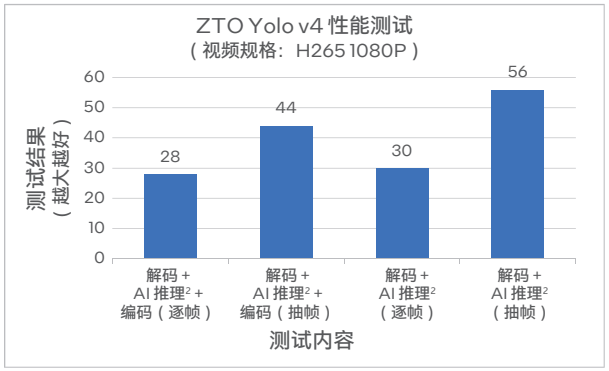


图 1. 基于英特尔® 数据中心 GPU Flex 系列 170 的 ZTO Yolo v4 性能测试结果[3]

为什么选择英特尔

一个模型，多设备部署：提升效率，节约成本

与英特尔合作，可以充分利用其架构下的各种基础设施，一个模型可以部署到多种设备，不仅提升了开发效率还节约了部署成本。中通估算，本次方案可帮他们实现约 34.8% 的成本节约[4]。

完备的软硬件产品组合为全开发链路护航

英特尔拥有完备的软硬件产品，可支持从模型训练、推理到应用开发和运维的整个开发链路。

硬件层面，除内置 AI 加速技术的 CPU（例如英特尔® 至强® 可扩展处理器）外，英特尔还提供数据中心 GPU 和 FPGA 等产品。中通本次使用的英特尔® 数据中心 Flex 系列 170 运算速度高达每秒 150 万亿次 (150 TOPS)[5]；并且配备了英特尔首款基于硬件加速的 AV1 编码器，能够在不牺牲画面质量的前提下将比特率提升 30%[5]，能以更低的功耗提供更出色的解码性能[6]。

软件层面，英特尔提供包括英特尔® 分发版 OpenVINO™ 工具套件和英特尔® oneAPI 工具套件等来帮助用户简化 AI 应用开发并实现应用跨 XPU 的无缝切换。

强大的生态系统和可靠的专业支持

英特尔强大的生态系统和专业的技术支持团队可为用户在项目前、项目中和项目后提供参考方案和专业支持，可显著提升企业 IT 团队解决问题和完成应用开发的效率。

更多信息

- 有关中通快递的更多信息，请访问：https://www.zto.com。
- 有关英特尔® 数据中心 GPU Flex 系列的更多信息，请访问：https://www.intel.cn/content/www/cn/zh/products/details/discrete-gpus/data-center-gpu/flex-series.html。
- 有关英特尔® 分发版 OpenVINO™ 工具套件的更多信息，请访问：https://docs.openvino.ai/cn/latest/index.html。
- 有关英特尔® 数据中心 GPU Flex 系列的其他案例研究，请访问：https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/arcvideo-tech-efficient-ai-system-old-film-repair.html。

intel®

[1] 有关英特尔® Deep Learning Streamer (DL Streamer) 的更多信息，请访问 https://dlstreamer.github.io/。

[2] 测试中的 AI 推理业务流包括 yolov4 目标检测、追踪和分类。

[3] 性能测试结果基于中通快递于 2022 年 10 月进行的测试。配置详情：单节点，双路英特尔® 至强® 金牌 6348 处理器（28 核/路，56 线程/路），启用超线程，启用睿频；GPU：英特尔® 数据中心 GPU Flex 系列 170；内存总容量：256 GB (16 x 16 GB, DDR 2933)；操作系统：Ubuntu 20.04；内核版本：5.10.54；工作负载：dlstreamer；编译器：gcc；库：英特尔® oneAPI 工具套件；其他软件：英特尔® OpenVino™ 工具套件 2022.2 版。

[4] 数据援引自中通快递内部估算结果，如需了解详情，请与中通快递联系。

[5] 《英特尔公布代号 Arctic-Sound M 数据中心 GPU 的更多细节》，https://www.intel.cn/content/www/cn/zh/newsroom/news/intel-announced-details-data-center-arctic-sound-m.html?wapkw#gs.bhps4i。

[6] 《英特尔® 数据中心 GPU Flex 系列》产品简介：https://www.intel.com/content/www/us/en/products/docs/discrete-gpus/data-center-gpu/flex-series/product-brief.html。