解决方案简介

英特尔® Edge AI Server 参考架构

简化高性能边缘原生视频分析的开发工作

英特尔® Edge AI Server 参考架构为 AI 推理、结构化、聚类和特征匹配等云边协同视频分析服务的构建提供经过验证的软硬件设计模式。这一解决方案利用英特尔® 高级矩阵扩展（Intel® Advanced Matrix Extensions，英特尔® AMX）加速 AI 推理特征匹配和聚类，进一步优化了英特尔® 至强® 处理器的性能。

将云原生架构扩展到云边协同模型需要依靠云的敏捷性和可扩展性，以及边缘的低时延、高可靠性、更高的隐私性和更低的传输成本等优势。边缘产生的海量视频数据正在推动 AI 增强型边缘分析需求的增长，以将数据转换为有价值的信息和洞察。

边缘原生开发方法就是利用边缘作为云的自然延伸这一理念，将基础设施扩展为云边协同连续体。这样一来，基于这种架构的边缘视频服务器就可以利用类云基础设施和边缘计算的优势：

- **支持高密度计算**，提供云原生灵活性和可扩展性。
- **专为视频工作负载设计和优化**，实现单位成本下更高的能效和性能。
- **在本地处理海量视觉数据**，以降低时延和传输成本并提高数据安全性。

英特尔® Edge AI Server 参考架构借助经过预先验证的软硬件构建模块，助力实现和加速这一转换。解决方案提供商可以简化边缘原生架构的落地，用于交通流量管理、制造业中的缺陷检测以及零售店客流量模式分析等视频分析用途，并降低开发成本，加快上市速度。这些解决方案自然承袭了第五代英特尔® 至强® 可扩展处理器的多项优化，其中包括众多内置加速器、更高的能效和更低的总体拥有成本（TCO），这些均有助于解锁边缘新机遇。

基于英特尔® 高级矩阵扩展（英特尔® AMX）的内置加速器扩展了 CPU 的 AI 推理能力。

加速边缘视频创新

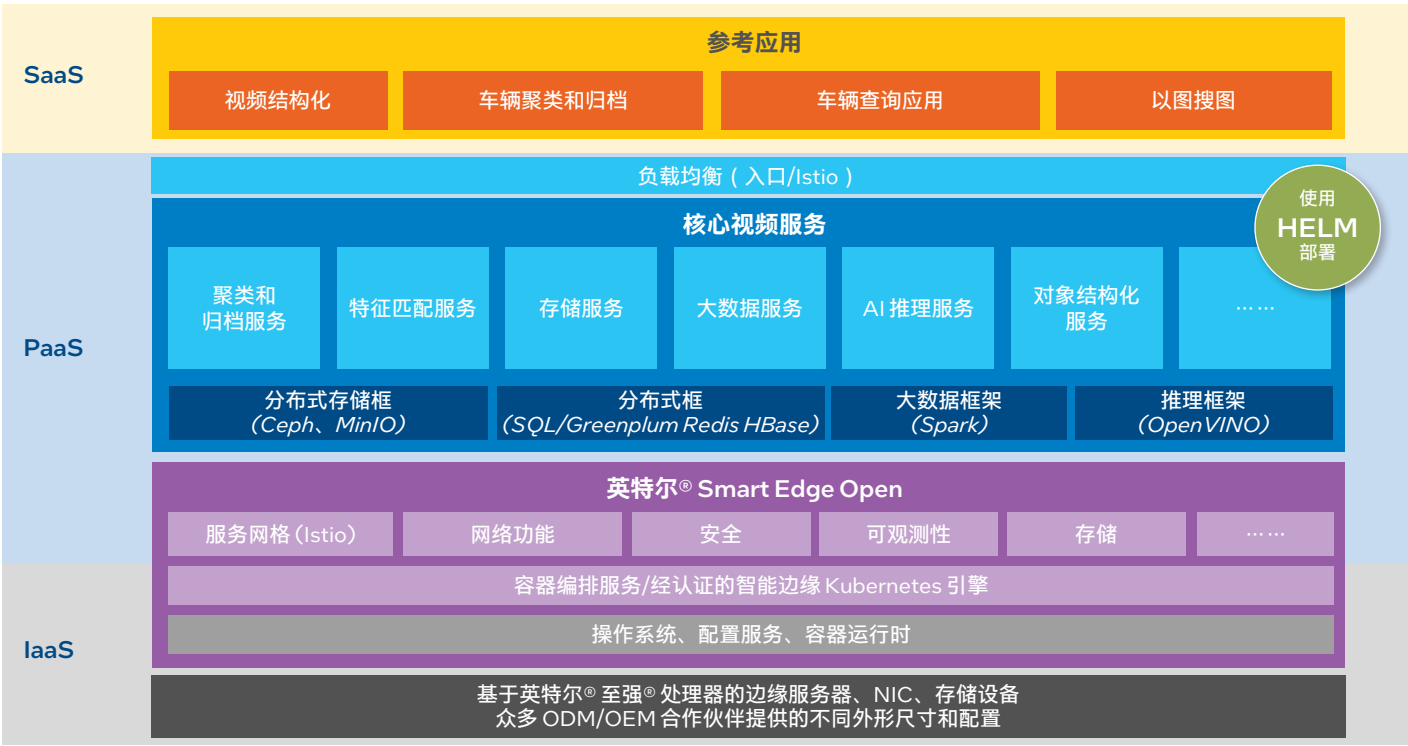
随着企业和机构纷纷着手构建面向视频分析的软件定义边缘原生实现方案，他们也面临着一系列挑战。许多企业和机构在优化性能和资源编排的硬件功能方面欠缺深厚专业知识，这进一步加重了开发和验证负担，使其难以构建高质量的解决方案。供应商中立的开源工具和组件支持有限，可能会增加解决方案的开发难度，甚至可能落后于其他行动更快、上市时间更短的竞争对手，而失去竞争优势。

英特尔® Edge AI Server 参考架构便是为应对这些挑战而设计，其基础为经验证的现成硬件配置，可根据各种边缘视频分析工作负载灵活定制。开源代码能够保证软件灵活性，可以支持基于英特尔® 至强® 处理器的服务器之间的兼容性和可移植性，以及英特尔® 以太网 800 系列网络适配器提供的高性能网络连接。这一硬件可在受限的边缘环境中支持各种外形尺寸和多节点拓扑。边缘原生架构基于 Kubernetes 和微服务，可从边缘到集群再到云端无缝部署可扩展、灵活和统一的服务。

AI 必须在 GPU 上运行这一传统观念不再是绝对。尽管 GPU 及其他专用加速器仍是许多高密度、计算密集型用例的理想硬件，但 CPU 应对中低复杂度 AI 工作负载的能力也日益提升。事实上，性能均衡的英特尔® CPU 平台在采用多个内置硬件加速器后，性能突飞猛进。如今，英特尔® CPU 已成为比 GPU 更灵活、更具成本效益的基础设施平台，可以更好地满足多种 AI 工作负载的需求。第五代英特尔® 至强® 可扩展处理器专为 AI 加速而生，可提供更出色的 AI 性能。

第五代英特尔® 至强® 可扩展处理器是从边缘扩展到数据中心再到云端的综合 AI 就绪计算环境中的关键硬件组件。英特尔® Edge AI Server 参考架构可帮助解决方案开发团队在边缘快速部署针对平台高度优化的视频分析。

开放编程模型具有硬件可移植性，且避免了供应商绑定风险，相比 CUDA 等专有模型，可使解决方案更好地适应未来需求。这一参考架构提供了工作负载导向型可定制的视频服务，已针对英特尔® 至强® 处理器进行了性能和能效预先优化，包括利用了英特尔® AMX 来加速 CPU 的 AI 推理。



英特尔® Edge AI Server 参考架构

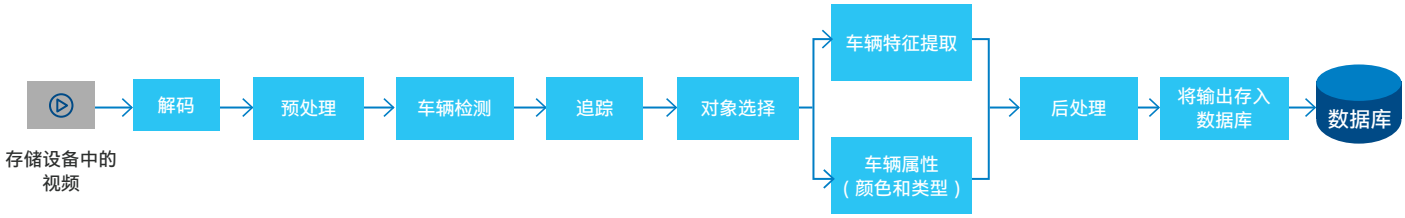
视频分析工作负载

高达 **2.1 倍**

视频分析流提升
与英特尔® 至强® 金牌 6348 处理器对比[1]

由参考架构实现的管线和服务

英特尔® Edge AI Server 参考架构基于容器化视频服务，提供模块化、可移植、可定制的拓扑。它让解决方案开发人员能够从一系列已针对英特尔® 至强® 处理器的硬件级加速进行了全面优化的软件中挑选和配置他们所需的组件。这一参考架构最突出的服务是 AI 推理、特征匹配和聚类。下文将逐一阐述这些服务。



AI 推理服务构建的运行时管线示例

AI 推理服务

AI 推理服务使用管线配置文件中的信息，构建从视频输入和解码到推理和输出的运行时管线，然后处理对管线的请求以进行媒体处理和 AI 推理作业。管线本身使用名为异构视频分析（HVA）的序列化拓扑，其中每个节点执行一个特定任务，并针对英特尔® 架构进行了高度优化。AI 推理服务集成了车辆检测、属性识别、对象跟踪、对象质量选择、特征提取等典型 AI 算法。

在本示例中，AI 推理服务检测输入流中的车辆，跟踪并选择单个车辆作为帧中的移动对象，然后生成并存储特征矢量。这一服务可以使用容器化微服务在各个边缘平台上进行打包和部署。

特征匹配服务

特征矢量是视觉对象的数学表征。英特尔® Edge AI Server 参考架构的特征匹配服务旨在查询巨大的数据集，并在数毫秒内返回最相似的矢量。存储特征矢量的数据库往往非常庞大，并且流、对象和特征的数量成倍增加，使得每个视频通道的特征匹配工作负载远大于 AI 推理服务的工作负载。因此，特征匹配服务能够通过将特征数据集分割到多个服务器或 Worker 实例上来分配任务，为应用提供 RESTful 服务 API，以便通过编程来访问该功能。

这项功能使服务能够部署更多 Worker 实例以支持更大的特征数据库或更高的吞吐量要求。开发人员无需重写服务代码，只需通过为不同用例创建特征存储客户端便可扩展特征匹配服务，从而为更广泛的解决方案生命周期提供了可扩展性。特征匹配服务的关键价值在于使解决方案能够以高效、可扩展、可重复的方式轻松比较多组视觉信息的内容。

聚类服务

边缘视频分析管线产生了海量非结构化数据，而聚类正是数据分析的重要一环。在安全、安保和设施管理等边缘视频用例中，IP 摄像头通常会传播大量数据，并将这些数据传输到边缘服务器，而几乎或完全不考虑结构或情境。聚类服务对特征矢量集进行分析，并利用这些计算结果将对象进行分组。

这些分组依据特征匹配、车辆跟踪或克隆车牌检测等特定用途的特征而分门别类。通过对数据进行聚类，可以大大缩小基于数据的分析范围，从而降低计算要求，节省计算时间。

预先验证的解决方案构建模块

英特尔® Edge AI Server 参考架构是一个完整的组件堆栈，为解决方案开发人员提供了特定领域边缘视频分析解决方案的起点。该堆栈的基础是第五代英特尔® 至强® 可扩展处理器，可支持负责提供上述服务的经过优化的软件工具和组件。这款处理器是一个性能均衡的平台，提供了全面技术升级，包括以下方面：

- **高吞吐量、高效的执行资源**：与上一代产品相比每核性能更高，配备众多内置加速器，并利用经优化的电源模式降低能耗。
- **更强大的内存子系统**：与上一代产品相比，更快的 DDR5 内存将内存速度提升高达 16%，共享缓存容量提升高达 3 倍（特定型号 SKU 上支持）。
- **速度更快、处理能力更强的 I/O**：每路多达 80 条 PCIe 通道，英特尔® 超级通道互联（Intel® Ultra Path Interconnect，英特尔® UPI）2.0 速度高达 20 GT/s，并支持 Compute Express Link（CXL）Type 1、Type 2 和 Type 3。

英特尔® AMX 是内置于第五代英特尔® 至强® 可扩展处理器的硬件加速器，用以执行深度学习训练和推理。它能够加速对于 AI 计算至关重要的矢量运算，无需借助其他独立硬件即可提高吞吐量。它能从处理器内核卸载这些运算，以更高的能效完成这些任务，从而释放处理器内核以处理其他工作。

开发人员可以优化代码，利用英特尔® AMX 为 AI 工作负载实现峰值性能，同时支持共享硬件上的通用工作负载。英特尔® Edge AI Server 参考架构利用英特尔的 AI 生态系统，通过跨硬件平台的免费开源编程软件组件，提供更大的灵活性，可更好地应对未来需求，且避免了供应商绑定风险。下文将介绍英特尔® Edge AI Server 生态系统的关键组成。

OpenVINO™ 工具套件

OpenVINO™ 工具套件支持一次编写，随处运行，可以简化开发工作。开发人员可以轻松转换和优化经 TensorFlow、PyTorch 和 Caffe 等主流框架训练的深度学习模型，以在本地、边缘或云端各种英特尔的硬件和环境中进行部署。

EPIC iO 采用了 OpenVINO™ 工具套件，这为其 AI 管线开发流程提供了关键差异化优势。随着该公司 AI 开发团队规模不断扩大，经验证的标准可帮助公司更高效地培训开发人员，使其更快达到所需的熟练度。

英特尔® Feature Matching Acceleration Library

英特尔® Feature Matching Acceleration Library 提供了可量产的分布式特征匹配解决方案，能够跨多个服务器处理非常庞大的特征集，为基于英特尔® 架构的边缘服务器提供了高度性能优化。

EPIC iO 在其新一代模型开发和“属性”机制中采用了英特尔® Feature Matching Acceleration Library 的功能。跨摄像头并处理节点聚类这一重要功能可为各行各业提供更好的视角和关联性。

结论

解决方案开发人员可以利用 AI 推理、特征匹配和聚类服务等功能，来实施英特尔® Edge AI Server 参考架构，从而加速边缘高性能 AI 增强型视频分析的投产。由于第五代英特尔® 至强® 可扩展处理器为边缘视频分析提供了多项突破性功能，使得该解决方案架构体现出在 CPU 上运行 AI 推理的新标准。这一转变有望帮助解决方案提供商及其客户普及 AI，利用数据来表现视觉世界，并从所见中发掘洞察、创造价值。

面向第四代英特尔® 至强® 可扩展处理器的英特尔® Edge AI Server 参考架构可在保密协议（NDA）规定下下载。面向第五代英特尔® 至强® 可扩展处理器的修订版预计将于 2024 年推出。更多详情，请见“Intel Edge Video Infrastructure Reference Architecture Get Started Guide（英特尔® 边缘视频基础设施参考架构入门指南）”（文件编号：767995）。

了解更多信息：

英特尔® 边缘计算解决方案和技术

关于 EPIC iO Technologies

EPIC iO Technologies 是一家以软件为核心的科技公司，致力于将 5G 就绪连接技术与人工智能物联网（AIoT）解决方案相结合。该公司的 DeepInsights 平台提供了一个云原生的联合软件平台，集成了 AI、计算机视觉、物联网传感器和遥测聚合、视频管理和可视化功能。

EPIC iO 为智慧城市、交通、零售、医疗、现场安全、企业智能空间、金融和酒店等用例提供模板化解决方案。EPIC iO 与英特尔建立了战略合作关系，以确保能在计算机视觉、边缘解决方案、数据处理安全性以及云边协同扩展方面立足前沿，面向未来。

intel.

[1] 详情请见以下网址的 [E6]：intel.com/processorclaims（第五代英特尔® 至强® 可扩展处理器）。结果可能不同。

实际性能受使用情况、配置和其他因素的差异影响。更多信息请见英特尔的性能指标网页。

性能测试结果基于配置信息中显示的日期进行的测试，且可能并未反映所有公开可用的安全更新。详情请参阅配置信息披露。没有任何产品或组件是绝对安全的。

具体成本和结果可能不同。

英特尔并不控制或审计第三方数据。请您审查该内容，咨询其他来源，并确认提及数据是否准确。

英特尔技术可能需要启用硬件、软件或激活服务。