Solution Brief

Intel® Edge AI Server Reference Architecture

Streamlining Development of High-Performance Edge-Native Video Analytics

Intel® Edge AI Server reference architecture provides validated hardware and software design patterns for building edge-to-cloud video analytics services including AI inference, structuring, clustering and feature matching. The solution accelerates AI inference feature matching and clustering using Intel Advanced Matrix Extensions (Intel AMX) to optimize performance on Intel® Xeon® processors.

Extending cloud-native architecture into a cloud-to-edge model builds on the cloud's agility and scalability, with edge advantages such as low latency, high reliability, enhanced privacy and reduced transmission costs. Massive amounts of video data being generated at the edge are driving the need for AI-enhanced edge analytics, to turn that data into usable information and insights.

An edge-native development approach draws on the idea of the edge as a natural extension of the cloud, to expand infrastructure into a cloud-to-edge continuum. Edge video servers based on this type of architecture therefore take advantage of benefits from both cloud-like infrastructure and edge computing:

- **High density computing** with cloud-native flexibility and scalability.
- **Designed and optimized for video workloads**, bringing better power efficiency and performance per dollar.
- **Process massive visual data locally** to reduce latency and transmission cost, with improved data security.

The Intel® Edge AI Server reference architecture helps enable and accelerate this transition with pre-validated hardware and software building blocks. Solution providers can streamline their adoption of edge-native architectures for video analytics usages such as traffic flow management, defect detection in manufacturing and retail store traffic pattern analysis, with lower development cost and faster time to market. Those solutions automatically inherit optimizations for 5th Gen Intel® Xeon® Scalable processors, which unlock new edge opportunities with built-in accelerators, better efficiency and lower TCO.

> *Built-in acceleration based on Intel Advanced Matrix Extensions (Intel AMX) is extending the capabilities for AI inference on the CPU.*

Accelerating video innovation at the edge

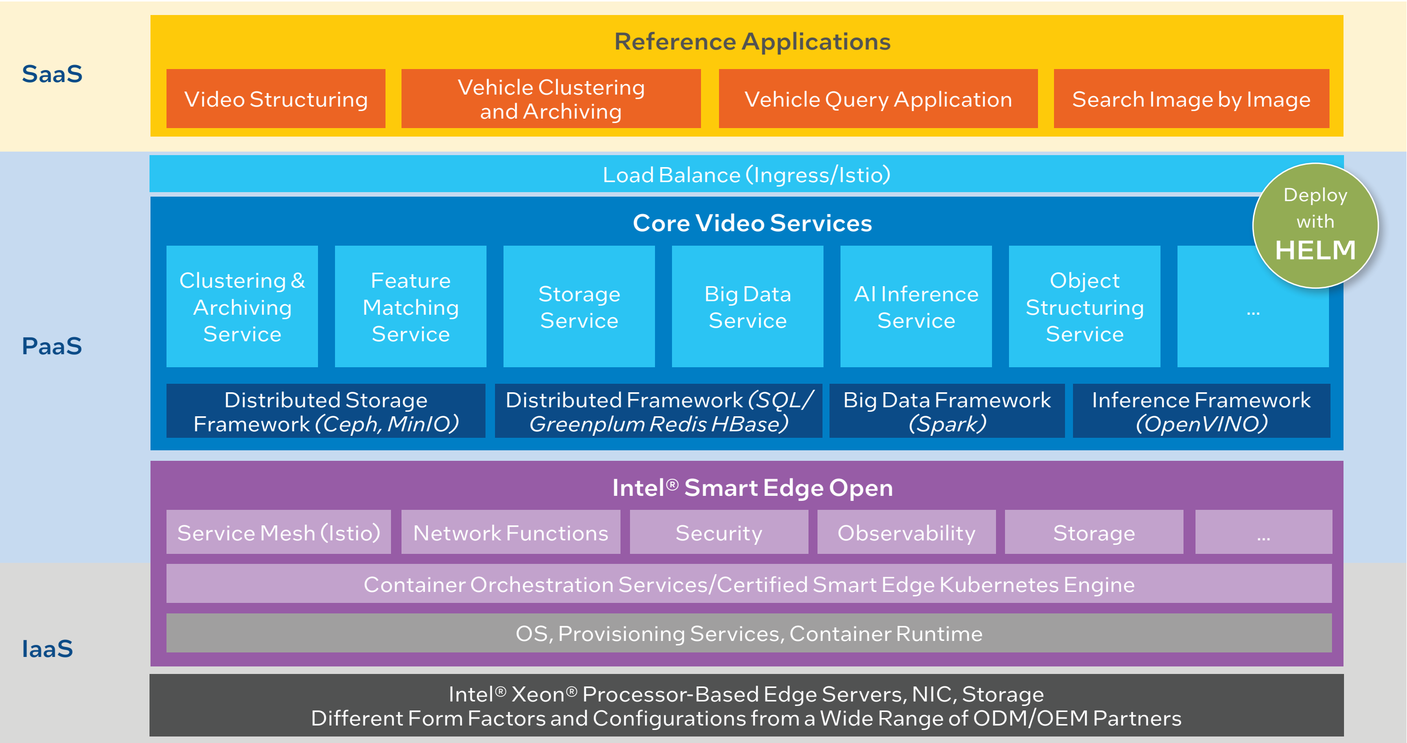
As organizations set out to build software-defined, edge-native implementations for video analytics, they face an array of challenges. Many lack deep expertise with hardware features for optimizing performance and resource orchestration, adding to the development and validation burdens to create high-quality solutions. Limited support for vendor-neutral and open source tools and components may drive up solution-development requirements, possibly even yielding competitive advantage to others who can move more quickly and get to market sooner.

The Intel Edge AI Server reference architecture is designed to neutralize those challenges, with a flexible foundation of validated, off-the-shelf hardware configurations that are tailored for a range of video analytics workloads at the edge. Software flexibility is guaranteed by the open source nature of the code, with compatibility and portability across Intel Xeon processor-based servers with high-performance network connectivity provided by Intel Ethernet 800 Series Network Adapters. This hardware supports various form factors and multi-node topologies for the constrained edge environment. The edge-native architecture is based on Kubernetes and microservices, to seamlessly deploy scalable, flexible and unified services from edge to cluster to cloud.

The historic viewpoint that AI has to run on GPUs is no longer absolute. While GPUs and other dedicated accelerators remain the preferred hardware for many high-density, compute-intensive usages, CPUs are increasingly capable for AI workloads of low to medium complexity. In fact, advances across balanced Intel CPU platforms, with built-in hardware accelerators, now make Intel CPUs a more flexible, cost-effective infrastructure platform for many AI workloads than GPUs. 5th Gen Intel Xeon Scalable processors are designed for AI, delivering AI performance that's unmatched by any other CPU.

5th Gen Intel Xeon Scalable processors are the key hardware ingredient for a comprehensive AI-ready compute environment that stretches from edge to data center to cloud. Intel Edge AI Server reference architecture gives solution development teams a fast track to deploying video analytics at the edge that are highly optimized for the platform.

An open programming model helps make solutions more future-ready than proprietary models such as CUDA, with portability across hardware and no vendor lock-in. The reference architecture provides workload-oriented, customizable video services that are pre-optimized for performance and power efficiency on Intel Xeon processors, including taking advantage of Intel AMX, which accelerates AI inference on the CPU.


SaaS
Reference Applications
Video Structuring
Vehicle Clustering and Archiving
Vehicle Query Application
Search Image by Image
PaaS
Load Balance (Ingress/Istio)
Deploy with HELM
Core Video Services
Clustering & Archiving Service
Feature Matching Service
Storage Service
Big Data Service
AI Inference Service
Object Structuring Service
...
Distributed Storage Framework (Ceph, MinIO)
Distributed Framework (SQL/ Greenplum Redis HBase)
Big Data Framework (Spark)
Inference Framework (OpenVINO)
Intel® Smart Edge Open
Service Mesh (Istio)
Network Functions
Security
Observability
Storage
...
Container Orchestration Services/Certified Smart Edge Kubernetes Engine
IaaS
OS, Provisioning Services, Container Runtime
Intel® Xeon® Processor-Based Edge Servers, NIC, Storage
Different Form Factors and Configurations from a Wide Range of ODM/OEM Partners

Intel® Edge AI Server reference architecture.

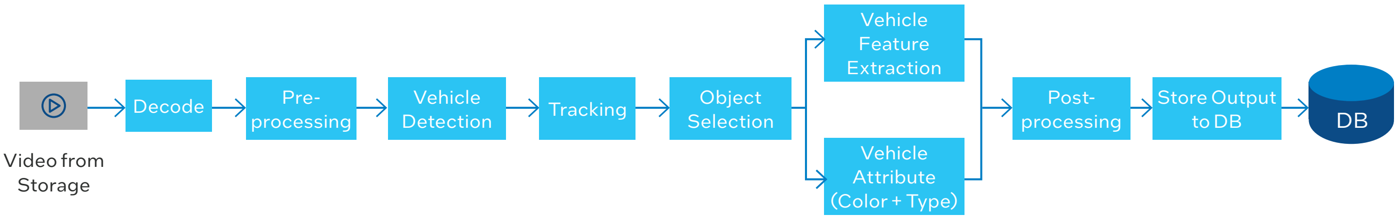
Video Analytics Workload

UP TO **2.1x**

More video analytics streams vs Intel Xeon Gold 6348 processor[1]

Pipelines and services implemented by the reference architecture

The Intel Edge AI Server reference architecture is based on containerized video services that offer a modular, portable, customizable topology. It gives solution developers the ability to pick and configure the components they need, from a collection of software elements that are fully optimized for hardware-level acceleration on Intel Xeon processors. The architecture's most prominent services are AI Inference, Feature Matching and Clustering, each of which is discussed below.



Sample runtime pipeline built by AI Inference service.

AI Inference service

Using information in the pipeline configuration file, the AI Inference service builds the runtime pipeline, from video input and decoding to inference and output, then processes requests to the pipeline for media processing and AI inference jobs. The pipeline itself uses a serialized topology called Heterogeneous Video Analytics (HVA), where each node performs a certain task, with high optimization for Intel architecture. The AI Inference service integrates typical AI algorithms such as vehicle detection, attribute recognition, object tracking, object quality selection and feature extraction.

In the example shown, the AI Inference service detects vehicles within the input streams, tracks and selects individual vehicles as moving objects with in the frame, then generates and stores feature vectors. The service can be packaged and deployed using containerized microservices on individual edge platforms.

Feature Matching service

Feature vectors are mathematical representations of visual objects. Intel Edge AI Server's Feature Matching Service is designed to query against huge datasets and return the most similar vectors within milliseconds. The databases that store feature vectors tend to be very large, and the numbers of streams, objects and features multiply together to make feature matching workloads substantially larger on a per-video-channel basis than the workloads for the AI inference service. The Feature Matching service therefore has the ability to distribute the work by splitting the feature dataset onto multiple servers or worker instances, providing a RESTful service API for applications to programmatically access that ability.

This functionality enables the service to deploy more worker instances to support larger feature databases or higher throughput requirements. Developers can extend the Feature Matching service by creating feature storage clients for different use cases without having to rewrite the service, providing extensibility for the broader solution lifecycle. The key value of the Feature Matching service is to enable the solution to readily compare the contents of multiple sets of visual information, in an efficient, scalable, repeatable manner.

Clustering service

Clustering is a critical aspect of giving meaning to the massive amount of unstructured data that is generated by edge video analytics pipelines. It is common in edge video use cases such as safety, security and facilities management for IP cameras to propagate massive amounts of data and deliver it to edge servers with little or no structure or context. The Clustering service conducts analytics on sets of feature vectors and uses the outcomes of those calculations to place objects into groups.

The categories that underlie those groups are based on characteristics that are tailored to specific purposes, such as feature matching, vehicle tracking or cloned license plate detection. By clustering the data, the scope of analysis based on that data can be dramatically reduced, saving on compute requirements and time.

Pre-validated solution building blocks

The Intel Edge AI Server reference architecture is a full stack of components that give solution developers a starting point for domain-specific edge video analytics solutions. The foundation of this stack is 5th Gen Intel Xeon Scalable processors, supporting optimized software tools and components responsible for delivering the services described above. The processor delivers enhancements across the balanced platform, including the following:

- **High-throughput, high-efficiency execution resources**. Improved per-core performance over its predecessor, with the industry's most built-in accelerators and reduced energy usage with optimized power mode.
- **Enhanced memory subsystem**. Up to 16% increased memory speeds using faster DDR5 memory than the previous generation and up to 3x larger shared cache (available on select SKUs).
- **Fast, high-capacity I/O**. Up to 80 lanes of PCIe per socket, with Intel UltraPath Interconnect (Intel UPI) 2.0 speeds up to 20 GT/s and support for Compute Express Link (CXL) Types 1, 2 and 3.

Intel AMX is a hardware accelerator for deep learning training and inference that is built into 5th Gen Intel Xeon Scalable processors. It accelerates the vector operations that are at the heart of AI computations to improve throughput, without additional discrete hardware. It offloads those operations from the processor cores, completing them with better power efficiency and freeing up cores for other work.

Developers can optimize code to take advantage of Intel AMX for peak performance on AI workloads while also supporting the general-purpose workloads that accompany them, on the same shared hardware. The reference architecture draws on the Intel AI ecosystem of free, open software components for programming across hardware platforms with maximum flexibility and future-readiness without vendor lock-in. Key aspects of that ecosystem for Intel Edge AI Server are described in the remainder of this section.

OpenVINO™ toolkit

The OpenVINO toolkit makes it easier to write once and run anywhere. Developers can readily convert and optimize deep learning models trained using popular frameworks including TensorFlow, PyTorch, and Caffe, to deploy across a mix of Intel hardware and environments, on-premises, at the edge or in the cloud.

EPIC iO has adopted the OpenVINO toolkit as a critical differentiator in its AI pipeline development process. As the company continues to scale its AI development team, validated standards enable it to onboard developers more efficiently and bring them up to the needed level of proficiency at a much faster pace.

Intel Feature Matching Acceleration Library

The Intel Feature Matching Acceleration Library provides a production-ready, distributed feature-matching solution with the capability to handle very large feature sets across multiple servers, with a high degree of performance optimization for edge servers based on Intel architecture.

EPIC iO is leveraging the Intel Feature Matching Library capability in some of its next-generation model development and "attribution" mechanisms. Cross-camera and processing node clustering is a required feature that provides all industries with enhanced perspectives and correlation.

Conclusion

Using capabilities that include the AI Inference, Feature Matching and Clustering services, solution developers can implement the Intel Edge AI Server reference architecture to accelerate their time to production for high-performing AI-enhanced video analytics at the edge. This solution architecture manifests the emerging norm of running AI inference on CPUs, with 5th Gen Intel Xeon Scalable processors offering breakthrough capabilities for edge-based video analytics. This shift promises to aid solution providers and their customers as they implement pervasive AI to represent the visual world using data and then create insight and intelligence from what they see.

Intel Edge AI Server reference architecture for 4th Gen Intel Xeon Scalable processors can be downloaded under NDA. A revision for 5th Gen Intel Xeon Scalable processors is expected in 2024. More details are available in the Intel Edge Video Infrastructure Reference Architecture Get Started Guide (Doc # 767995).

Learn more:

Intel® Edge Computing Solutions & Technologies

About EPIC iO Technologies

EPIC iO Technologies is a software-focused technology company that combines 5G-ready connectivity with AIoT solutions. The company's DeepInsights platform delivers a cloud-native, federated software platform with integrated AI, computer vision, IoT sensor and telemetry aggregation, video management and visualization.

EPIC iO offers templated solutions for use cases including smart cities, transportation, retail, health care, site security, enterprise smart spaces, finance and hospitality. The company has aligned itself strategically with Intel to ensure that its computer vision, edge solutions, processing security and edge-to-cloud scale are leading-edge today and ready for tomorrow.

[1] See [E6] at intel.com/processorclaims: 5th Gen Intel® Xeon® Scalable processors. Results may vary.

Performance varies by use, configuration and other factors. Learn more on the Performance Index site.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Intel technologies may require enabled hardware, software or service activation.

1223/LH/MESH/353960-001US