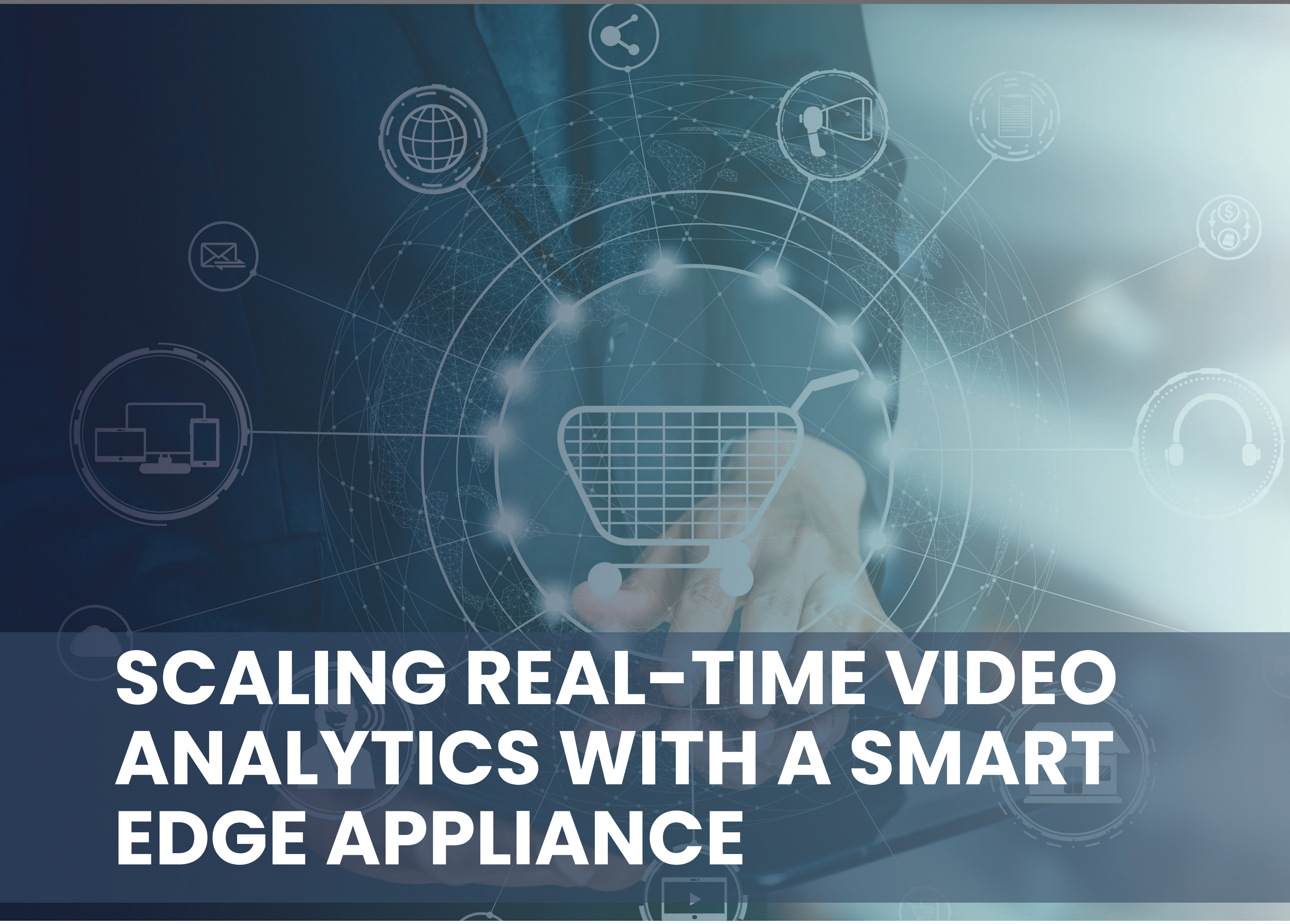
kinara



SCALING REAL-TIME VIDEO ANALYTICS WITH A SMART EDGE APPLIANCE

WHITE PAPER

Introduction

If data is the new oil, then video is its shipping container. According to IDC, more than 40% of all enterprise data produced worldwide originates from video cameras (IDC, 2022). By 2025, 1 million minutes of video will be streamed across the internet every second from various sources, including security systems, traffic cameras, drones, social media, and many others. The abundance of video data presents opportunities to extract meaningful, actionable insights in real-time using video analytics. Figure 1 demonstrates some such use cases.

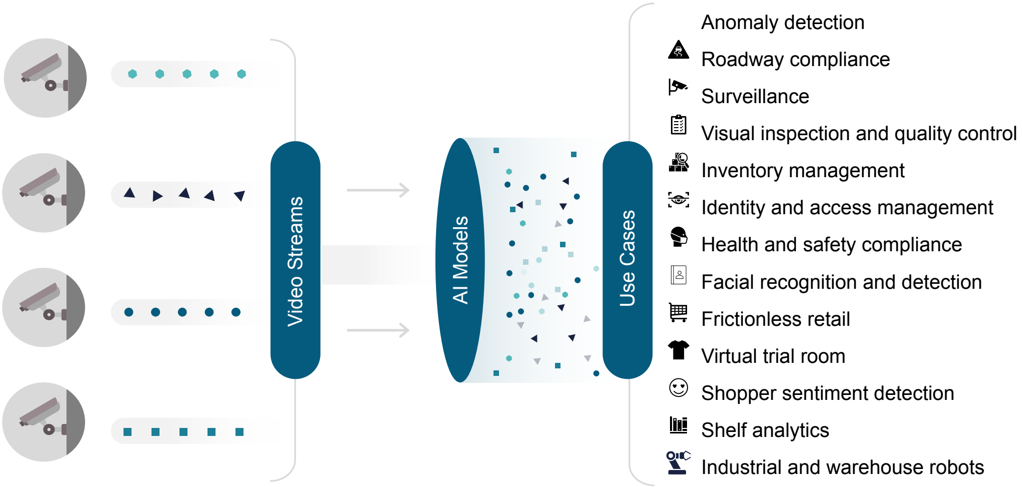

Video Streams
AI Models
Use Cases
Anomaly detection
Roadway compliance
Surveillance
Visual inspection and quality control
Inventory management
Identity and access management
Health and safety compliance
Facial recognition and detection
Frictionless retail
Virtual trial room
Shopper sentiment detection
Shelf analytics
Industrial and warehouse robots

Figure 1.
Some example applications of real-time video analytics

Video analytics can be performed in the camera, the cloud, or the edge server and its lower-cost cousin, the edge appliance. The edge appliance, located on-premises, is typically used for smaller multi-camera installations. This whitepaper will review an edge compute appliance – the EAIA - powered by the AMD Kria™ System on Module (SOM) combined with Kinara Ara-1 AI processors. The appliance provides a cost-effective, high-performance video analytics solution for multi-camera installations.

Some Applications of Video Analytics

We will first describe the applications of video analytics across various sectors.

Security & Surveillance: AI-driven video analytics enable real-time face recognition, intrusion detection, loitering detection, and crowd monitoring. These insights aid in threat prevention and timely response, enhancing public safety. For example, intelligent algorithms can recognize patterns and anomalous behavior in crowded areas, helping to prevent potential threats.

Retail: Stores can use video analytics for customer behavior analysis, aiding in inventory management, store layout optimization, and enhancing customer experiences. Cashier-less stores are an emerging trend where surveillance cameras track customers' interactions with items, powering an automated payment process and creating a seamless shopping experience.

Transportation: In intelligent traffic systems, AI-based video analytics provide real-time license plate recognition, vehicle classification, traffic flow optimization. These insights contribute to efficient traffic management, predicting congestion points, and promoting safer road conditions.

Manufacturing: Video analytics can significantly improve quality control on assembly lines. Real-time defect detection and anomaly recognition ensure efficient operation and minimize production downtime. Similarly, video analytics can guide robotic systems, facilitating automation and increasing manufacturing precision.

Intelligent Buildings: Surveillance cameras and AI-based video analytics improve building security with facial recognition-enabled access control. Occupancy detection helps optimize energy usage, contributing to efficient, sustainable building management.

Healthcare: Remote patient monitoring and diagnostics benefit significantly from video analytics. Real-time vital sign monitoring, anomaly detection, and medical imaging analysis can improve patient outcomes and reduce healthcare costs.

The Case for an Edge Appliance for Video Analytics

Modern video analytics applications rely heavily on artificial intelligence algorithms to enable advanced features. Most applications also need to run multiple AI models per stream. Processing multiple simultaneous video streams further exacerbates the computational demands, as each additional stream introduces more data to analyze and insights to derive using AI models.

Processing the video data in the cloud is often impractical. This is due to the real-time nature of many of these applications, where the latency of the cloud is too high. Furthermore, a high-bandwidth connection is required to transfer entire video frames to the cloud, which can not only be expensive but also raises concerns about security risks and privacy.

Performing video analytics at the edge resolves these issues. Analyzing the video data locally reduces both cost and latency, enabling real-time video analytics for various applications. It also enhances privacy, reducing the amount of sensitive data transmitted and stored remotely.

Traditionally, video analytics at the edge has been implemented in two ways.

1. **Smart Cameras**: These cameras have built-in AI capabilities to perform analysis directly on the camera. They can run AI models to detect objects, movements, or suspicious activities.

2. **Edge Servers with GPUs**: Video feeds from multiple cameras are sent to an on-premises edge server equipped with a GPU. The edge server performs the video analysis, using more complex models than AI-enabled cameras due to significantly higher computational power available with the GPUs.

Both approaches have significant limitations (Figure 2).

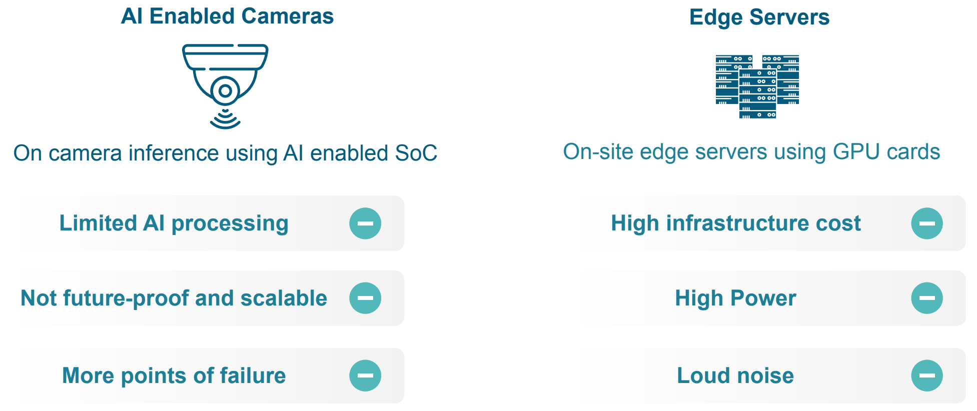


Figure 2
Challenges of real-time video analytics with AI cameras and edge servers

Smart cameras have limited computational power, especially for AI. Embedded system-on-chips (SoC) that power smart cameras today sometimes include a dedicated AI engine to handle AI inference. However, these processing engines share the resources of the SoC. This limits their ability to run multiple AI models per stream. The AI models are also constantly evolving in their complexity and capabilities. Frequently, there is a need to enhance the system's AI processing capabilities. To attain this functionality with smart cameras, the existing cameras will have to be replaced with more powerful ones, which can be cost-prohibitive.

Edge Servers with GPUs are expensive. They cost thousands of dollars and consume a lot of power leading to high energy and cooling costs, especially in large-scale deployments. It makes them unsuitable for deployments where cost and power budgets are major constraints.

This leads us to a third option: the edge appliance. An edge appliance is a video analytics system that is not only optimized for cost and power but also supports high-performance AI inference on multiple video streams. An appliance utilizing an embedded processor with power- efficient AI accelerators provide the same performance as a GPU at an order of magnitude lower cost and lower power. The edge appliance is highly scalable. It can be easily configured to support additional video streams and increased model complexity by adding more AI accelerators.

Example: Edge AI Appliance with AMD Kria™ SOM and Kinara Ara-1

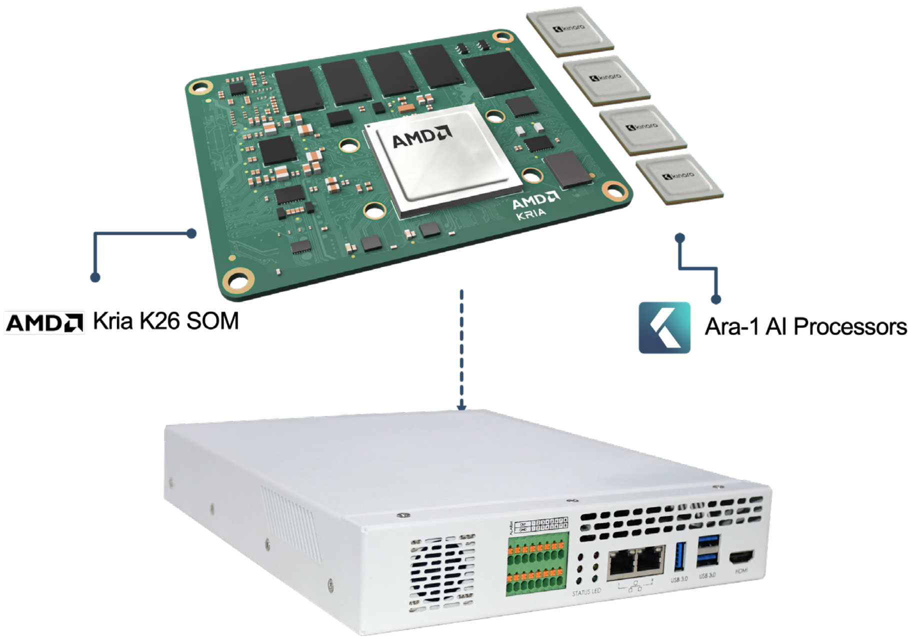


Figure 3
Edge compute appliance with Kria™ SOM and Kinara Ara-1

The edge compute appliance (or Edge AI Appliance or EAIA) is one example of a cost-effective and power-efficient video analytics system. At the heart of it is the AMD Kria™ SOM and four Kinara Ara-1 AI processors, (Figure 3).

The AMD Kria SOM is built around the Zynq™ UltraScale+™ MPSoC architecture, which combines 4 Arm® Cortex®-A53 cores™ with programmable FPGA logic. This enables developers to create custom hardware accelerators for specific tasks, significantly improving the efficiency and performance of edge applications. The Ara-1 processors provide high-performance AI processing capabilities with low latency and low power consumption. The combined capabilities of the Kria processor and Ara-1 accelerators result in an appliance that can perform complex AI workloads on eight 1080p video streams.

To understand how the EAIA performs video analytics on the incoming video streams, we need to describe an AI inference pipeline. An AI inference pipeline typically consists of three steps: pre-processing, inference, and post-processing, as shown in Figure 4.

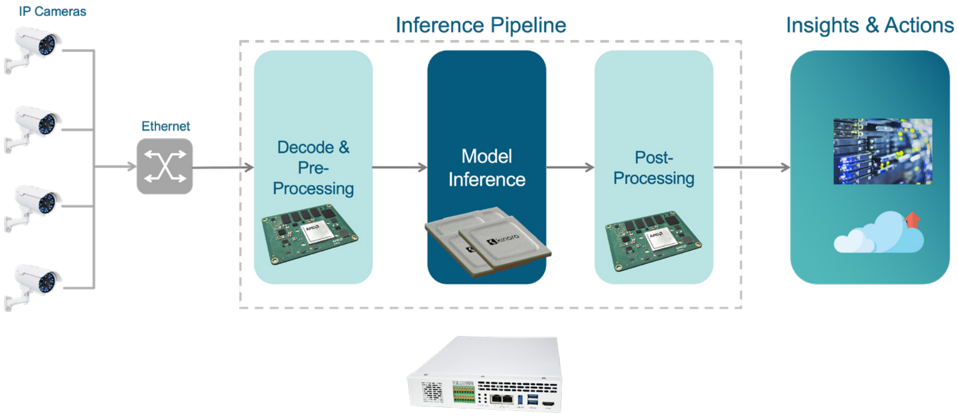


Figure 4
Edge AI Inference pipeline

Pre-processing: The pre-processing step prepares the input data for the specific AI model being executed by formatting it to match the model's input requirements. These include image resizing, normalization, and augmentation.

Model Inference: The inference step involves running the pre-processed video data through one or more trained AI models to make predictions or classifications. Typically models include object detection, activity recognition, facial recognition, image segmentation or any other task that the AI model has been trained to perform.

Post-processing: The post-processing step filters the inference results for use by the analytics application. This may involve adjusting the predictions, filtering out outliers, or providing additional information about the predictions. Some of the common post-processing steps include bounding box adjustments, confidence score thresholding, and so on.

The edge compute appliance uses the Kria host processor to perform pre-processing and post-processing tasks while the Ara-1 processors are used for inference as illustrated in Figure 5. The video streams from cameras are decoded and processed by the Kria processor, using the Vitis Vision libraries. After the pre-processing step, data is transferred to the Kinara Ara-1 processors to run inference workload in real time. Post-processing steps, such as annotating video feeds with bounding boxes or labels, along with the business logic, can be executed either by the Kria or an external management server.

The management of the video analytics pipeline is done via the widely used GStreamer media streaming framework. Both AMD and Kinara provide hardware accelerated operations and model inference through open source GStreamer plugins. With these plugins, users can quickly create deployment-ready AI solutions and deploy them as microservices.

Depending on the model, the edge AI appliance can be up to 7X more cost effective compared to a GPU-based edge appliance like NVIDIA Jetson series. In other words, it can provide the same performance per stream at a 7X lower cost (Figure 5).

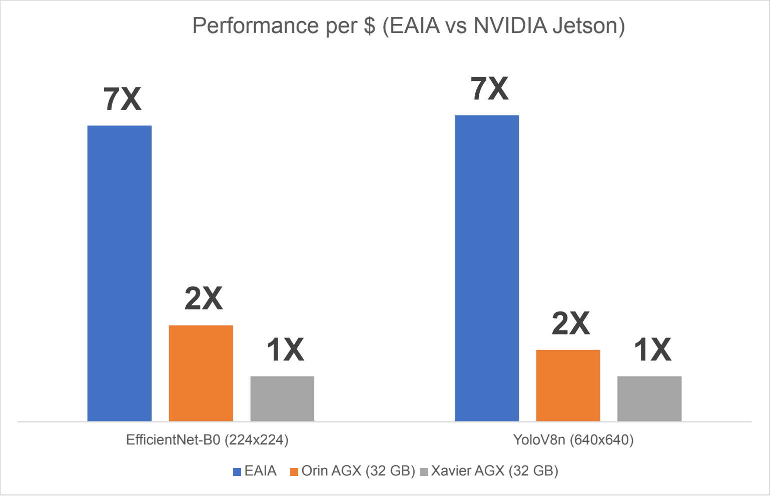


Figure 5
Performance per $ comparison between edge AI appliance and NVIDIA Jetson devices

Note: Jetson performance for the benchmarks are for batch size of 1. The pries are for the full box listed online (through AAEON)

Remote Management using Kubernetes and GRPC

In many deployment scenarios, the edge appliances need to be managed and monitored remotely. AMD and Kinara have employed a robust, scalable, and resilient framework utilizing Kubernetes and gRPC (Google Remote Procedure Call) to meet this requirement, as shown in Figure 6. This crafts a complete edge-to-cloud solution for real-time inference.

Kubernetes is an open-source platform that automates the deployment, scaling, and management of containerized applications across a cluster of hosts. An edge deployment with many cameras will require multiple edge compute appliances. Each of these appliances can be considered a 'node' in Kubernetes – a worker machine, that is the smallest unit for deployment.

The edge compute appliance comes with the Kubernetes infrastructure. This enables seamless scaling of these appliances by dynamically adding or removing nodes based on workload. If traffic to a particular node is high, Kubernetes is also able to load balance and distribute the video streams across other nodes to keep the deployment stable. This flexibility is particularly important as video analytics demand fluctuates during the day. It ensures that the system adapts to maintain optimal performance.

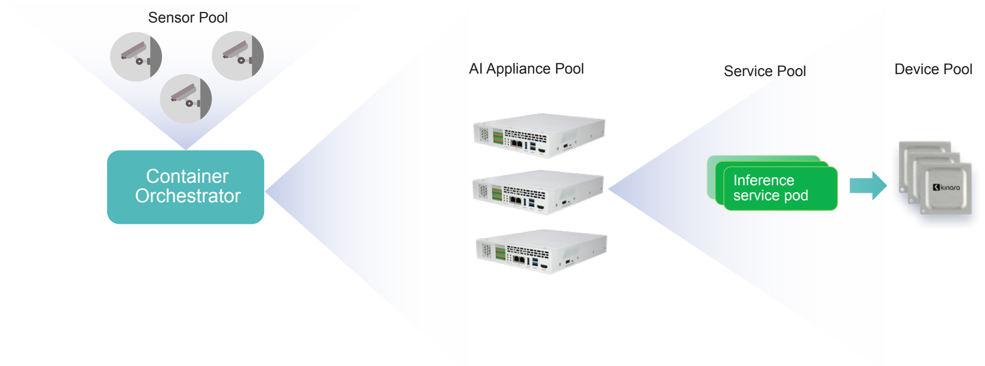


Figure 6
Orchestration and management for EAIA deployment

The gRPC protocol further enhances this framework – enabling users to create a microservice-oriented architecture for real-time inference with useful features like load balancing, health checking, and failure recovery. It facilitates smooth, efficient communication among all the EAIA devices and a central server, ensuring reliable control and status updates.

The ability to configure the number of cameras and edge AI appliances and to dynamically add inference pipelines for the video streams helps in scalability and resilience of the infrastructure. If any device goes offline or has an error, the Kubernetes infrastructure can dynamically track and ensure a failsafe recovery.

Conclusion

AI powered analytics solutions that extract meaningful information from the vast amount of video data has the potential to transform many industries – from security to retail, healthcare, transportation, and manufacturing. However, performance, cost, and privacy considerations necessitate processing video streams at the edge rather than in the cloud.

The Edge AI Appliance (EAIA) discussed in this paper offers an optimal edge solution for multi-stream video analytics. By combining the AMD Kria system-on-module with Kinara Ara-1 AI processors, the EAIA achieves high throughput video analytics with low latency, low power, and optimal cost effectiveness.

The EAIA's seamless software integration using GStreamer and toolchains like Vitis Vision and Kinara SDK simplifies the deployment of multiple concurrent AI inference pipelines on incoming video streams. The Kubernetes and gRPC based management framework provide flexibility to scale the solution by adding appliances and facilitates resilient operations.