



Intel® AI Edge Systems Reference Architecture – AI Accelerated Real-time Video Insights on RHEL 10.2

Efficiency Optimized Edge AI Systems with Intel® Core™ Ultra Series 3 Processors

Reference Architecture

August 2026

Intel Confidential

Authors:
Abhijit Sinha
Yuan Kuok Nee
Edel Curley



You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or visiting the [Intel Resource and Documentation Center](#).

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No product or component can be absolutely secure. Check with your system manufacturer or retailer or learn more at [intel.com](#).

No product or component can be absolutely secure.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

Contents

1.0	Introduction	5
1.1	Executive Summary	5
1.2	Intel's Edge AI Portfolio	5
1.3	Red Hat's Software Portfolio	6
1.4	Terminology	7
2.0	Reference Architecture	8
2.1	Reference Software	8
2.2	Use Case	8
2.3	Solution Architecture	9
2.3.1	Red Hat AI	10
3.0	Installation Instructions	11
3.1	Scalable Solution Across Xeon and Core	11
3.2	Installation Instructions	11
4.0	Hardware and Software Configuration	13
4.1	Hardware BOM	13
4.2	Software BOM	13
4.3	Platform Security	15
4.4	Side Channel Mitigation.....	15
5.0	Summary.....	16

Figures

Figure 1.	Video Search and Summarization Use Case.....	9
Figure 2.	Solution Architecture.....	9
Figure 3.	System Architecture	11

Tables

Table 1.	Terminology	7
Table 2.	Platform Configuration Intel® Core™ Ultra Series 3 Processor	13
Table 3.	Software Configuration	13



Revision History

Date	Revision	Description
August 2026	1.1	Added Red Hat Quickstart link.
June 2026	1.0	Initial release.

1.0 Introduction

1.1 Executive Summary

Red Hat's* leadership in providing consistent application platform delivery from cloud to far-edge, together with the breadth and performance of Intel's processor portfolio—spanning Xeon, Core, and Atom—offers a powerful and compelling collaboration to accelerate AI adoption and unlock new opportunities. This reference architecture, jointly developed by Intel and Red Hat*, strives to reduce integration complexity by providing design guidance for a real-time captioning use case.

1.2 Intel's Edge AI Portfolio

Intel's Edge Silicon Portfolio is designed to accelerate AI adoption at the edge using not only our CPUs but also our integrated GPUs and our integrated NPUs and the discrete GPUs. It is structured along a number of swim lanes.

Entry and Mainstream Edge: Suitable for low-end workloads and small AI workloads, Addresses the Low-end space with limited integrated GPU capability.

Efficiency Optimized Edge AI - Core Ultra is our flagship product with significantly integrated GPU space, ideal for customers starting their AI journey. In many Edge use cases, integrated AI acceleration may be all you need.

Scalable performance Edge AI - with desktop parts, discrete GPUs, as well as our Xeon parts, offering a path for customers to grow with Intel's AI technologies

Intel has been a key technological partner at the edge and has built a broad, open [ecosystem](#).

Intel is accelerating Edge AI readiness through programs like **Intel® AI Edge Systems and Intel Edge AI Suites**.

Intel® AI Edge Systems are a range of optimized commercial AI systems delivered and sold through OEM/ODM in the Intel® ecosystem. They are commercial platforms, verified, configured, tuned, and benchmarked using Intel's reference AI software application on Intel hardware to deliver optimal performance for Edge workloads.

Intel AI Edge Systems provide a solid foundation for rapid development and deployment, so you can streamline deployments with pre-optimized, right-sized systems.

Intel Edge AI suites offer reference pipelines and libraries to accelerate our partners in optimized Edge AI application development with the aim of bringing industry-ready, AI-driven capabilities to market fast. OEMs can use the software available in these vertical suites to evaluate and demonstrate performance. The following verticals are represented in the latest release: Metro AI suite, Manufacturing AI suite, Retail AI suite, Education AI suite and Health and Life Sciences, and Robotics.

For more information, refer to the [GitHub Open Edge Platform AI Suites](#) collections.

Intel's OpenVINO™ is an open-source toolkit designed for optimizing and deploying deep learning models, accelerating AI inference across Intel hardware, including CPUs, GPUs, and NPUs.

Intel AI Edge Systems and Intel Edge AI Suites simplify onboarding for Edge AI ISVs and fast-track performance optimization for specific vertical use cases.

This reference architecture features reference software from Intel's Edge AI Suites for an AI-accelerated video captioning use case running on the latest **Intel® Core™ Ultra Series 3 processor** – Intel® Core™ Ultra Series 3 X7 358H, with:

- P-Cores, E-Cores, LP E-Cores
- Integrated Graphics: Intel® Arc™ B390 Graphics
- NPU (V5): For Power-Efficient AI Inferencing

1.3 Red Hat's Software Portfolio

Red Hat, the leading provider of enterprise open-source solutions, offers a comprehensive and integrated portfolio of technologies designed to modernize enterprise IT operations, accelerate innovation, and reduce complexity across hybrid cloud, data centers, and edge environments.

Edge devices such as Self-checkout terminals and IoT gateways need to operate with ruthless efficiency, with compute resources constrained, alongside power and connectivity. Red Hat Device Edge offers operational consistency across far-edge workloads and devices. It features:

- Consistence from Edge to Cloud

Platforms designed to allow for ultimate flexibility from the far edge to cloud environments.

- Edge Management Platform

Easy-to-use fleet management with Red Hat Edge Manager enables extending declarative device orchestrations to thousands of edge devices.

- Edge Optimized OS Capabilities



Automated device onboarding, customizable image generation, over-the-air updates, and intelligent rollbacks provided by Red Hat Enterprise Linux minimize the need for on-site expertise.

- Red Hat Extensions

The RHEL extensions repository is a collection of software, curated by Red Hat, that includes developer tools, open-source libraries, and niche utilities. Its primary goal is to deliver high-value, fast-moving software through Red Hat's secure supply chain. This means you get access to software that's up to date and built and signed by Red Hat (including OpenVINO™ when available).

- Together, Red Hat and Intel empower the Edge AI ecosystem to build resilient, future-ready platforms that support digital transformation, edge computing, and AI innovation.

1.4 Terminology

Table 1. Terminology

Term	Description
FFMPEG	Fast Forward Moving Picture Experts Group
GPU	Graphics Processing Unit
iGPU	Intel integrated Graphics (iGPU)
Intel® TXT	Intel® Trusted Execution Technology
LLM	Large Language Model
MQTT	Message Queuing Telemetry Transport
NPU	Neural Processing Unit
RA	Reference Architecture
RAG	Retrieval-Augmented Generation
RHEL	Red Hat* Enterprise Linux*
RTSP	Real-Time Streaming Protocol
TPM	Trusted Platform Module
VLM	Vision Language Model

2.0 Reference Architecture

Red Hat* and Intel have joined forces to create this Reference Architecture for our partners, aiming to minimize risk by validating an integrated architecture that ensures compatibility between various components. The solution also addresses pain points by providing documented design guidance, deployment guidance, and support that can be used at various stages (planning, designing, and implementation) of a business project targeting Edge deployments. The software stack, hardware BOM, and guidance installation instructions are documented.

2.1 Reference Software

To create a validated system recipe and demonstrate the compatibility of system components, live video search and live video alert agent in the latest release of the *Smart NVR sample application* from Intel's Edge AI suites are leveraged.

- [Metro AI Suite](#)
- [Video Alert Agent](#)
- [Live Video Captioning](#)

This **Reference Architecture (RA)** defines the “what” and “how” for a class of deployments—compute, acceleration, software stack, operational model.

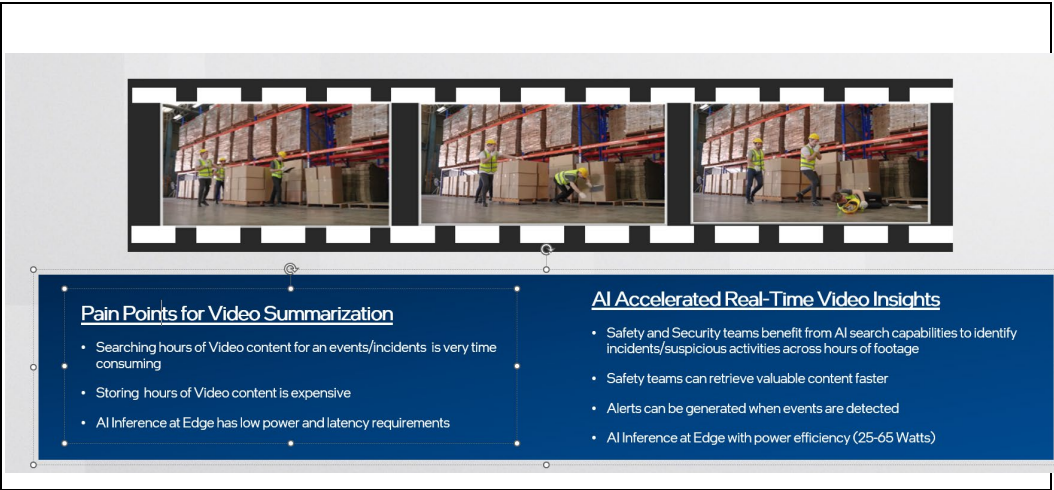
Intel also develops **Verified Reference Blueprints**, where a reference architecture transitions to a recipe with verified performance, benchmarked and validated to deliver optimized, scalable, repeatable Edge AI performance.

2.2 Use Case

Video Search and summarization are useful for rapid access to relevant video content. However, they can be time-consuming. Current solutions can perform object detection on hours of stored footage to retrieve a useful frame. The challenge increases when the analysis must be done across multiple camera feeds. The AI Accelerated Real Time Video insights solution performs captioning in almost real time - captioning the video frame using a Vision Language Model (VLM) and an embedding model. Storing this library of ‘shortlisted’ video clips at Edge rather than the entire video stream allows for streamlined, reactive local/Edge solutions. The database can be queried, and because the clips are captioned, they can be more efficiently utilized for text search cues by the central control center. Alerts can also be generated when certain events are detected for immediate incident response.

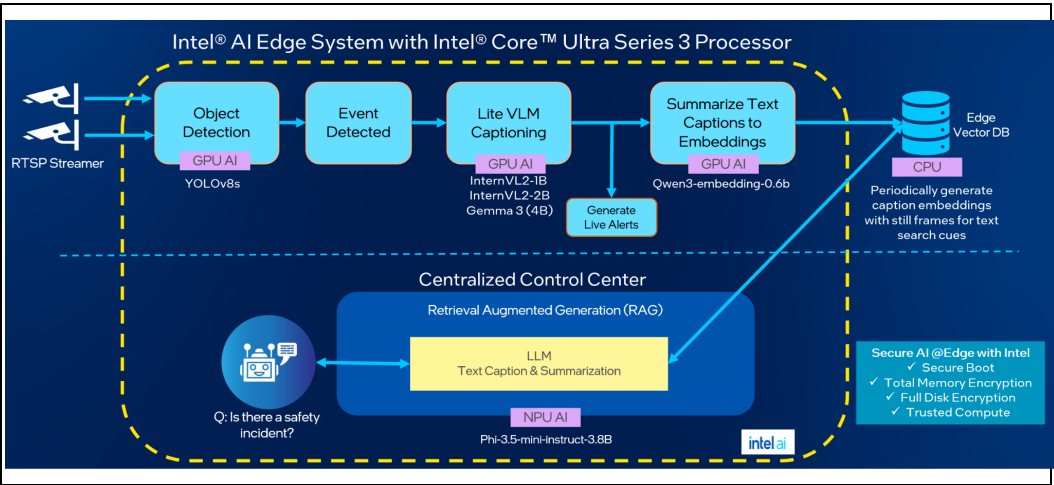
The AI compute power offered by Intel® Core™ Ultra Series 3 processor integrated graphics enables a VLM to run with lower latency than previous generations, thus captioning the camera stream at a higher latency.

Figure 1. Video Search and Summarization Use Case



2.3 Solution Architecture

Figure 2. Solution Architecture



Combining YOLO-based detection with VLM-generated captions delivers richer context and fewer false positives, enabling faster alerting and operational decisions in environments such as industrial safety. This architecture implements a lite VLM approach where only frames of interest are passed to the VLM to be captioned

The NPU-powered chat UI retrieves relevant keyframes and timestamps using vector similarity, enabling rapid “ask the video” queries for search, summarization, and review across one or more camera streams.

Intel Core Ultra Series 3 (Panther Lake) enables real-time video captioning and multi-modal inferencing by distributing work across CPU, Xe3 GPU, and NPU—reducing the need for server/workstation deployments

2.3.1 Red Hat AI

For more information, refer to the [Red Hat AI developer](#) page.

Red Hat AI Inference Server

Deploy your preferred models faster and more cost-effectively across the hybrid cloud with Red Hat AI Inference Server.

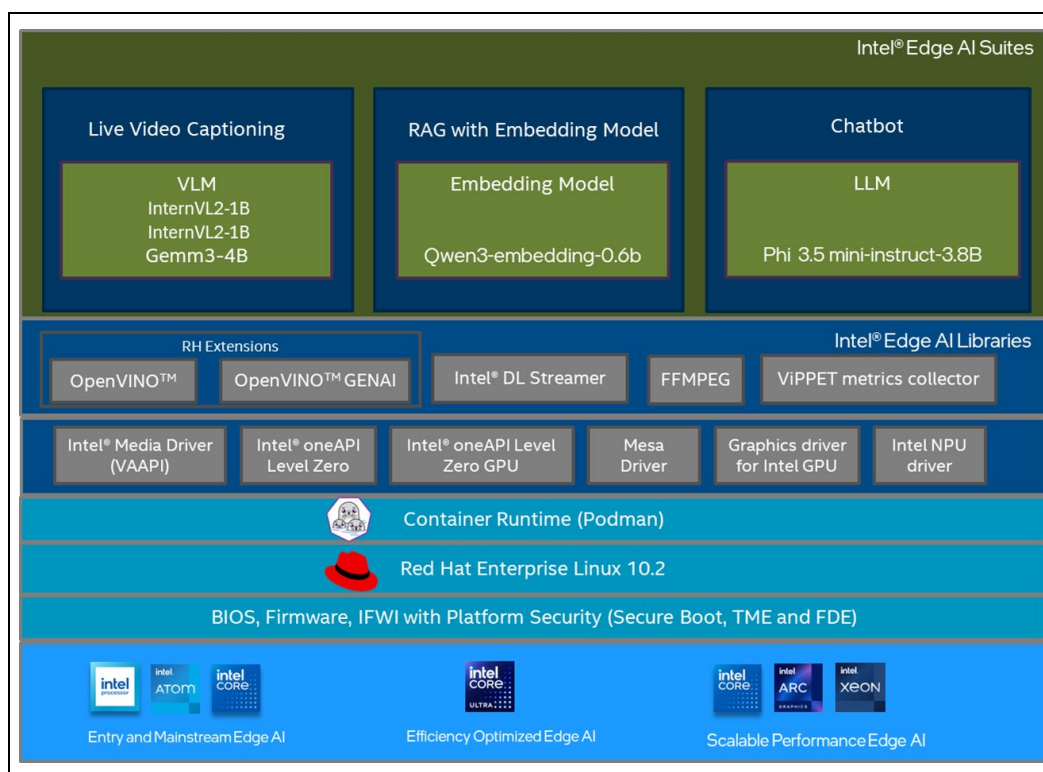
Features and benefits include:

- vLLM runtime maximizes inference throughput and minimizes latency.
- Pre-optimized model repository for rapid model serving.
- LLM compressor reduces compute costs without sacrificing accuracy.

3.0 Installation Instructions

3.1 Scalable Solution Across Xeon and Core

Figure 3. System Architecture



3.2 Installation Instructions

1. Use an Intel® Core™ Ultra Series 3 processor-based platform.
2. Install RHEL 10.2 GA from Red Hat*.
3. Intel® Core™ Ultra Series 3 processor (Code named Panther Lake) kernel with GPU and NPU drivers:
 - [Kernel](#)
 - [NPU driver packages](#)
 - Install OpenVINO™ by following the installation steps on the [OpenVINO™ page](#).

4. Application Installation:

- a. AI Accelerated Real-Time Video insights: The sample application processes Real-Time Streaming Protocol (RTSP) streams and generates live captions using a Vision-Language Model. The overall infrastructure involves ingesting an RTSP stream, processing it through a DLStreamer pipeline that samples frames for VLM inference, and delivering the resulting insights to the dashboard. The software is open-source and can be installed using the [Intel Open Edge Platform Edge AI Suites](#).
- b. Chat-question-and-answer-core: Chat Question-and-Answer sample application is a foundational Retrieval-Augmented Generation (RAG) pipeline that allows users to ask questions and receive answers, including those based on the captioned video data corpus. The software is open-source and can be installed using the [Intel Open Edge Platform Edge AI Libraries](#).

All applications are containerized using Podman. [Figure 1](#) shows the architecture diagram of the blueprint.

For the demo, refer to the [Red Hat Quickstart](#).



4.0 Hardware and Software Configuration

The hardware BOM used to develop the Verified Reference Blueprint is listed below.

4.1 Hardware BOM

Table 2. Platform Configuration Intel® Core™ Ultra Series 3 Processor

Ingredient	Requirement	Required/ Recommended	Quantity
Processor	Intel® Core™ Ultra Series 3 Processor*	Required	1
Memory	32GB (2x16GB) DDR5 5600MT/s SO-DIMMs or higher	Required	2
	OR 32GB LPDDR5x, 8533 MT/s or higher	Recommended	
Network	Intel® Ethernet Network Adapter i226-V/LM/IT (2.5 Gbps)	Required	1
Storage (Boot/Capacity Drive)	1 TB M.2 NVME or equivalent	Required	1
iGPU	Intel® Arc™ graphics (integrated in SoC)	Required	1
NPU	Intel® AI Boost (integrated in SoC)	Required	1

Note: * Recommended CPU SKU is Intel Core Ultra 7 366H (4 Xe cores) for two camera streams or Intel Core Ultra X7 358H (12 Xe cores) for up to four concurrent camera streams.

4.2 Software BOM

Table 3. Software Configuration

SW Components (Intel Open Edge Platform)		Live Video Captioning with RAG	Chat QnA
AI Models	Vision Language Model (VLM)	OpenGVLab/InternVL2-1B	N/A
		OpenGVLab/InternVL2-2B	

		Google/Gemma-3-4b-it	
	Large Language Model (LLM)	N/A	microsoft/Phi-3.5-mini-instruct
	Embedding Model	QwenText/qwen3-embedding-0.6b	N/A
	Object Detection	YOLO-v8s	N/A
Frameworks and Runtimes	OpenVINO™ Toolkit	OpenVINO™ GenAI 2025.4.1	OpenVINO™ GenAI 2025.4.1
	DL Streamer Pipeline Server	2026.0.0-ubuntu24	N/A
	VDMS Vector DB	intellabs/vdms:v2.11.0	N/A
	Intel multimodal-embedding-serving	1.3.2	N/A
	FFMPEG	6.1.1	N/A
	MQTT Broker	2.0	N/A
	Intel ViPPET Collector	2025.2.0	N/A
	MediaMTX (WebRTC signaling)	1.16.2	N/A
	NGINX (reverse proxy)	N/A	1.29.2
Low-Level Libraries & Drivers	Intel Media Driver (mesa-va-driver)	25.4.6-1	
	Intel NPU Driver	1.28.0	
	Intel Graphics and OpenCL Driver	26.01.36711.4-0	
Development Tools	Python	3.12.3	
Containerization Tools	Podman Engine	5.6.0	
Operating System and Kernel		RHEL 10.2 GA, 6.17	

4.3 Platform Security

It is recommended that Intel® Boot Guard Technology be enabled on Edge AI hardware to verify the platform firmware is suitable during the boot phase.

In addition to protecting against known attacks, all Intel® Accelerated Solutions recommend installing the Trusted Platform Module (TPM). The TPM module enables administrators to secure platforms for a trusted (measured) boot with known trustworthy (measured) firmware and OS. This allows local and remote verification by third parties to advertise known safe conditions for these platforms through the implementation of Intel® Trusted Execution Technology (Intel® TXT).

4.4 Side Channel Mitigation

Intel recommends checking your system's exposure to the "Spectre" and "Meltdown" exploits. This reference implementation has been verified with Spectre and Meltdown exposure using the latest Spectre and Meltdown Mitigation Detection Tool, which confirms the effectiveness of firmware and operating system updates against known attacks.

The [spectre-meltdown-checker tool](#) is available for download.

5.0 Summary

This reference architecture outlines a practical approach to deploying Red Hat* OpenShift AI on Intel® Core™ Ultra Series 3 processor, ensuring that the integrated GPU component (iGPU) is exposed to the software.

As workloads evolve toward higher throughput and agentic AI patterns, the architecture scales without application refactoring or platform redesign. Intel® Core Ultra 3 integrated GPU capabilities, combined with OpenShift AI, enable efficient, optimized inference performance through validated tuning and runtime practices. Support for Secure boot, Total memory encryption, Full disk encryption, and Trusted Compute further extends the platform's capabilities for security-sensitive use cases.

Collectively, these capabilities provide a balanced, future-ready AI platform that prioritizes flexibility, performance, and investment protection, while enabling support for the full AI lifecycle.

For more information, refer to the *AI on Intel® Xeon® Processors with Red Hat OpenShift AI* ([Document number: 919242](#)).