Intel® AI Edge Systems Verified Reference Blueprint with Retail Use Case and GEN AI

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

Reference Architecture

April 2026

Intel Confidential

Authors:
Abhijit Sinha
Yuan Kuok Nee
Edel Curley

Document Number: 916799-1.0

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.

No license (express or implied, by estoppel or otherwise) to any Intel® lectual 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.

 Intel® warrants performance of its FPGA and semiconductor products to current specifications in accordance with Intel® 's standard warranty but reserves the right to make changes to any products and services at any time without notice.

Intel® assumes no responsibility or liability arising out of the application or use of any information, product, or service described herein except as expressly agreed to in writing by Intel® . Intel® customers are advised to obtain the latest version of device specifications before relying on any published information and before placing orders for products or services.

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

Contents

1 Introduction 5
1.1 Terminology 5
1.2 Verified Reference Blueprint Architecture 6
1.3 GEN AI 7
2 Hardware and Software Configurations 9
2.1 Hardware BOM 9
2.2 Software BOM 9
2.3 Platform Security 10
2.4 Side Channel Mitigation 11
3 Verified Performance Data 12
3.1 Retail Use Case: Verified Performance 12
3.2 Gen AI Verified Performance 13
4 Summary 15
Appendix A Appendix 16
A.1 Automated Self-Checkout Test Methodology 16
A.2 Gen AI Test Methodology Using OpenVINO™ with Gen AI 18

Tables

Table 1. Terminology 5
Table 2. Gen AI Models 8
Table 3. Platform Configuration Intel® Core™ Ultra Processor Series 3 9
Table 4. Software Configuration 9
Table 5. Gen AI Models 13

Revision History

Revision Date	Revision	Description
April 2026	1.0	Initial Release.

§

1 Introduction

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 enable our partners to jumpstart development through a hardened system foundation verified by Intel® and to increase the trust in their system performance.

Verified Reference Blueprints (VRB) include Hardware BOM, Foundational Software configuration (OS, Firmware, Drivers), and are tested and verified with a supported Software stack (software framework, libraries, orchestration management).

Intel® AI Edge Systems Verified Reference Blueprint with Retail Use Case and GEN AI – Efficiency Optimized Edge AI on Intel® Core™ Ultra processors Series 3 features reference software from Intel's Open Edge Platforms Edge AI Suites - Intel® Automated Self-Checkout software. GEN AI models have been profiled with a recent release of Intel's OpenVINO™ (LLM test bench). All data has been verified with established performance benchmarks.

This blueprint features the latest **Intel® Core™ Ultra (Series 3)** – Intel Core Ultra Series 3 X7 358H with the following:

- 4 P-Cores, 8 E-Cores, 4 LP E-Cores
- Graphics: 12Xe @ 2.50GHz for AI intensive workloads
- NPU (V5): for Power-efficient AI inferencing

The Bill of Materials (BOM) details for the configurations and software used are provided in Chapter 2 of this document.

1.1 Terminology

Table 1. Terminology

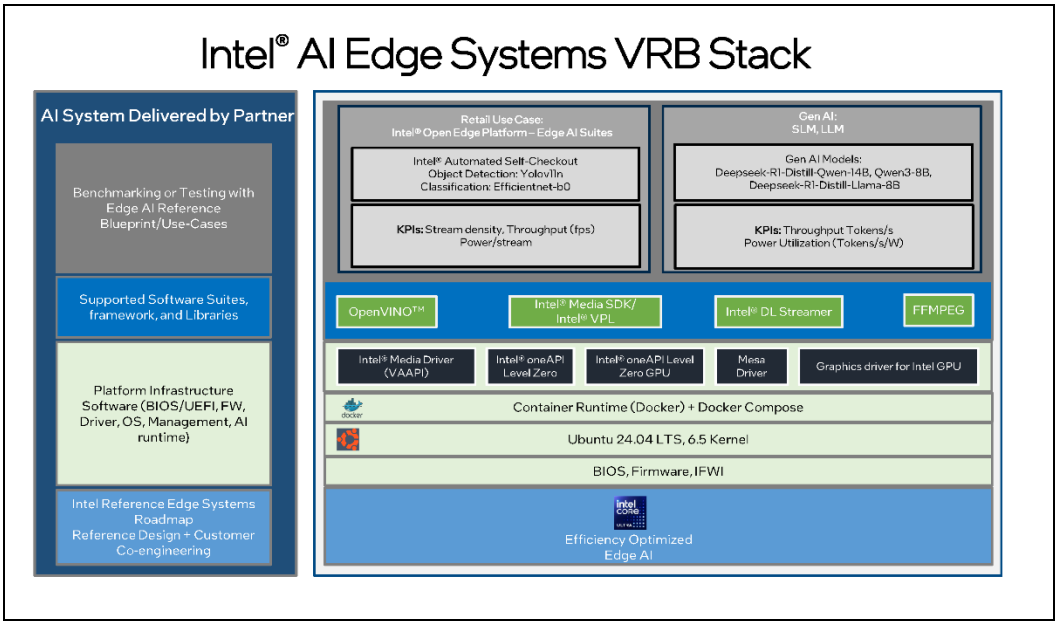
Term	Description
BOM	Bill of Materials
Intel® TPM	Intel® Trusted Platform Module
Intel® TXT	Intel® Trusted Execution Technology
ODM	Original Design Manufacturer
OEM	Original Equipment Manufacturer
VRB	Verified Reference Blueprint

1.2 Verified Reference Blueprint Architecture

The software stack used in this blueprint profiles Verified Performance Data for Vision AI using a Retail use case and for Gen AI using OpenVINO™. The Retail use case leverages software from Intel's Edge AI Suites available on the Intel Open Edge Platform.

All applications are containerized using Docker. Figure 1 shows the architecture diagram of the blueprint.

Figure 1. Architecture of the Intel® AI Edge Systems Verified Reference Blueprint



Self-Checkout in grocery retail is moving from manual Barcode scanning only to vision-enabled or vision-only checkout. Modern AI-powered self-checkout systems use cameras to recognize products visually, eliminating the need to locate and scan a barcode for each item. For non-packaged items, like fruits and vegetables, self-checkout can be a painstaking process. You must search for the item name on the display or find the sticker on a piece of produce, and then manually enter it into the system. It can be difficult to identify the correct option among many, and what happens if the system doesn't recognize the item, or you make a mistake in entering the code? You likely will have to wait for assistance. Given the variety of non-packaged items, staff need to be trained on the available options. When image recognition capabilities are incorporated at the register, it offers better customer experience and lower costs for retailers

The Intel® Automated Self-Checkout reference software from Intel's Open Edge Platform - Edge AI Applications, implements detection and classification features to identify a product during checkout. The software KPI is stream density, i.e the number of camera streams supported on the platform at the target 14.95FPS. The video data is ingested and pre-processed before each inferencing step. The AI inference is performed using two models: YOLOv11 and EfficientNet. The YOLOv11 model does product detection, and the EfficientNet model does product classification.

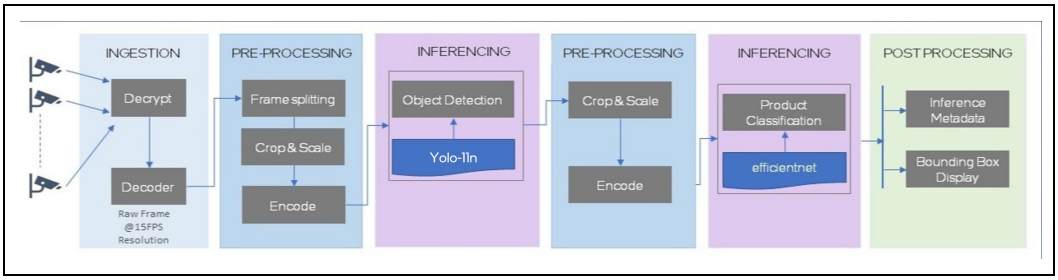
ISVs have previously implemented AI workloads on larger GPUs, but Intel's Core Ultra Series 3 iGPU accelerates AI inferencing and provides a power-efficient alternative to discrete GPUs.

The Automated Self-Checkout Reference Implementation provides critical components to build and deploy a self-checkout use case using Intel® hardware, software, and other open-source software such as OpenVINO™.

For instance, in this case, all the models in the pipeline are converted into OpenVINO™ format. In addition, this proxy Edge AI application makes use of both GStreamer for media processing and DLStreamer for inferencing, which includes detection and classification, accounting for all stages within the processing pipeline. For more details, see Appendix.

The video stream is cropped and resized to enable the inference engine to run the associated models. This reference software supports either running directly on the CPU or fully offloading to the iGPU or NPU, including encoding/decoding and inferencing. (For the first rev. of this document, the software runs on CPU or iGPU only. The next rev. supports offload to NPU).

Figure 2. Vision AI Video Analytics Pipeline



The Yolo11n model is used for Object detection, and the Efficient-b0 model for Object classification, which are optimized using OpenVINO™ at INT8 precision. Individual system results may vary as power and performance are affected by use, configuration, and other factors. Details are at Intel Product Performance Index.

For this Verified Reference Blueprint, we have also measured the average package Power dissipated in Watts. The stream density measurement is recorded as the number of camera streams.

1.3 GEN AI

The Gen AI **Verified Reference data** leverages the OpenVINO™ Gen AI LLM Benchmarking

The results shown include performance values for throughput (tokens/s) and the time to generate the first token. Also, the average power measured during AI inferencing is recorded.

Table 4 shows the software configuration for this Blueprint.

The large language model (LLM) proxy workload highlights the Gen AI processing capabilities on iGPU and NPU.

The GEN AI testing leveraged the models in the table below.

Table 2. Gen AI Models

Model Name	Parameters
DeepSeek R1-Distill-Qwen-14B	14B
Qwen3-8B	8B
DeepSeek R1-Distill-Llama-8 B	8B

§

2 Hardware and Software Configurations

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

2.1 Hardware BOM

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

Ingredient	Requirement	Required/ Recommended	Quantity
Processor	Intel® Core™ Ultra Series 3 X7 358H with 4 P-Cores, 8 E-Cores, 4 LP E-Cores	Required	1
Memory	LPDDR5, 8533 MT/s	Required	?
Network	Intel® Ethernet Network Adapter i226-V/LM/IT (2.5 Gbps)	Required	1
Storage (Boot/Capacity Drive)	1 TB or equivalent	Required	1
iGPU	Intel® Arc™ graphics (part of SoC)	Required	1
NPU	Intel® AI Boost (part of SoC)	Required	1

2.2 Software BOM

Table 4. Software Configuration

SW Components		Vision AI	Gen AI
Reference Workloads		Intel Automated self-checkout v3.6.4 (Intel Open Edge Platform - Retail AI Suite)	OpenVINO™ Gen AI Benchmarking Suite (LLM Bench)
Frameworks and Runtimes	OpenVINO™ Toolkit	OpenVINO™ 2024.6.0	OpenVINO™ GenAI 2025.4.1
	Optimum-intel	N/A	1.27.0
	DLStreamer	dlstreamer:2025.2.0-ubuntu24	N/A

SW Components		Vision AI	Gen AI
	FFMPEG	6.1.1	N/A
Low-Level Libraries & Drivers	Intel Video Processing Library (VPL)	2.16.0	
	Intel LibVA library (Video Acceleration API)	2.22.0	
	Intel Media Driver (mesa-va-driver)	25.4.6-1	
	Intel NPU Driver (only needed on SE 60n)	1.28.0	
	Intel Graphics and Open CL Driver	26.01.36711.4-0	
Development Tools	Python	3.12.3	
Containerization Tools	Docker Engine	29.2.1	
	Docker Compose	v5.0.2	
Operating System and Kernel		Ubuntu 24.04.4 LTS, 6.17.0-14-generic	

2.3 Platform Security

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

In addition to protecting against known attacks, all Intel® Accelerated Solutions recommend installing the Intel® Trusted Platform Module (Intel® 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).

2.4 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 at GitHub Spectre-Meltdown Checker.

§

3 Verified Performance Data

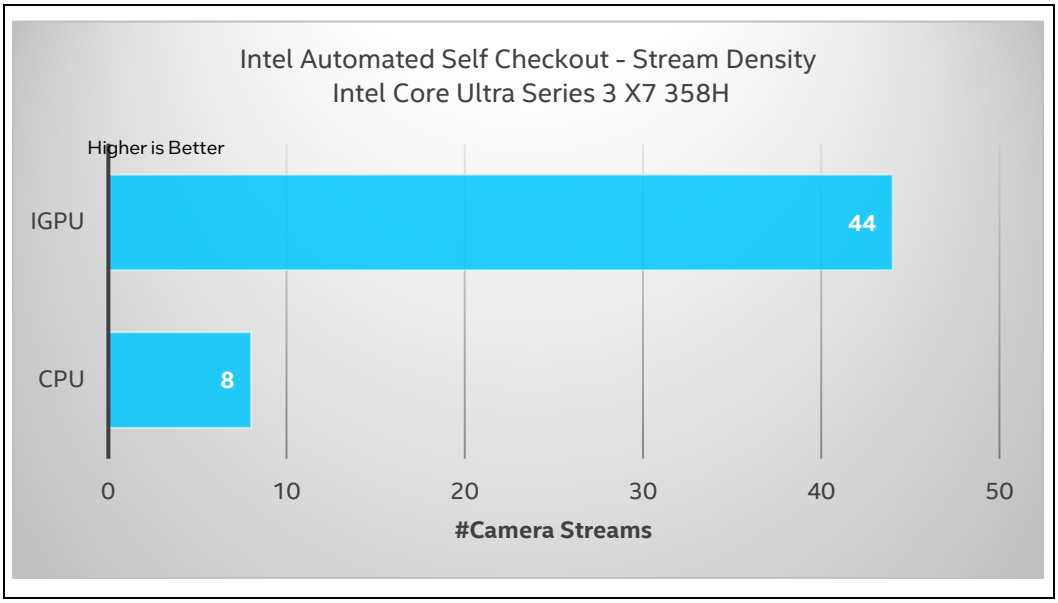
The Verified Performance Data is shown below.

The solutions were tested in January 2026, with the hardware and software configurations listed in Chapter 2.

3.1 Retail Use Case: Verified Performance

The number of camera streams (stream density) for Intel's Automated Self-Checkout reference software is shown below.

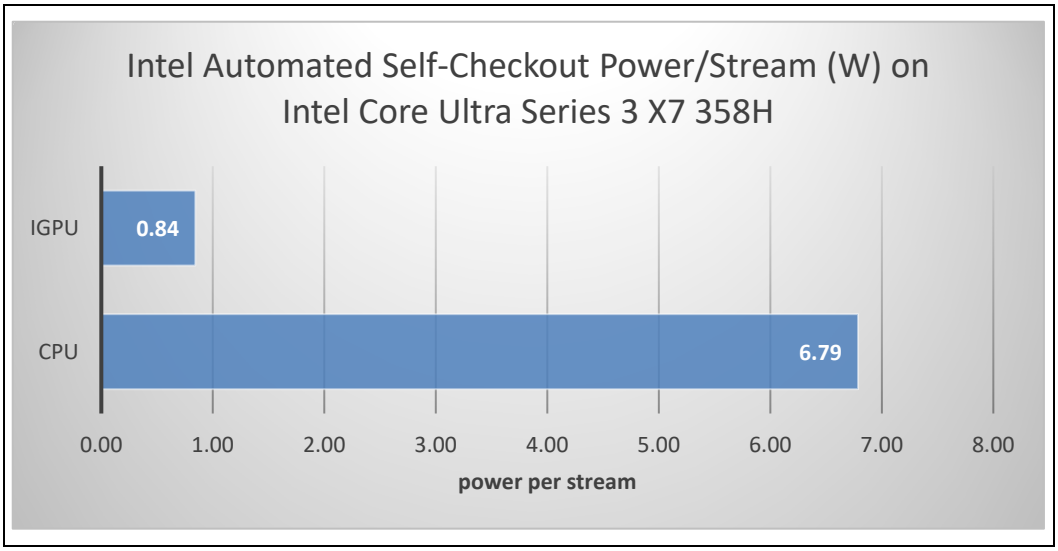
Figure 3. Retail Use Case - Intel® Core™ Ultra (Series 3) – Core Ultra Series 3 X7 358H on CPU and iGPU



The graph above shows that the Intel® Core™ Ultra Series 3 X7 358H can support up to 8 streams on CPU, 44 streams on iGPU @14.95 fps target FPS with **batch size 8.**

The power dissipated during the use case, per camera stream, is shown in the figure below.

Figure 4. Retail Use Case – Power per stream



The Power per stream was 0.84 watts of power per stream on Intel® Core™ Ultra Series 3 X7 358H iGPU with batch size 8 with the Vision AI (Retail Use Case) pipeline.

The Power per stream was 6.79 watts of power per stream on Intel® Core™ Ultra Series 3 X7 358H CPU with batch size 8 with the Vision AI (Retail Use Case) pipeline.

3.2 Gen AI Verified Performance

Using Intel's OpenVINO™ benchmarking tool, verified performance values for throughput (tokens/s) and the time to generate the first token were measured. Also, the average power measured during AI inferencing is recorded.

The GEN AI testing leveraged the models in the table below.

Table 5. Gen AI Models

Model Name	Parameters
DeepSeek R1-Distill-Qwen-14B	14B
Qwen3-8B	8B
DeepSeek R1-Distill-Llama-8 B	8B

The results for the GEN AI Workload (Throughput) and Power Efficiency on both iGPU and NPU are shown in Figure 5.

Figure 5. GEN AI Throughput (Tokens/s)

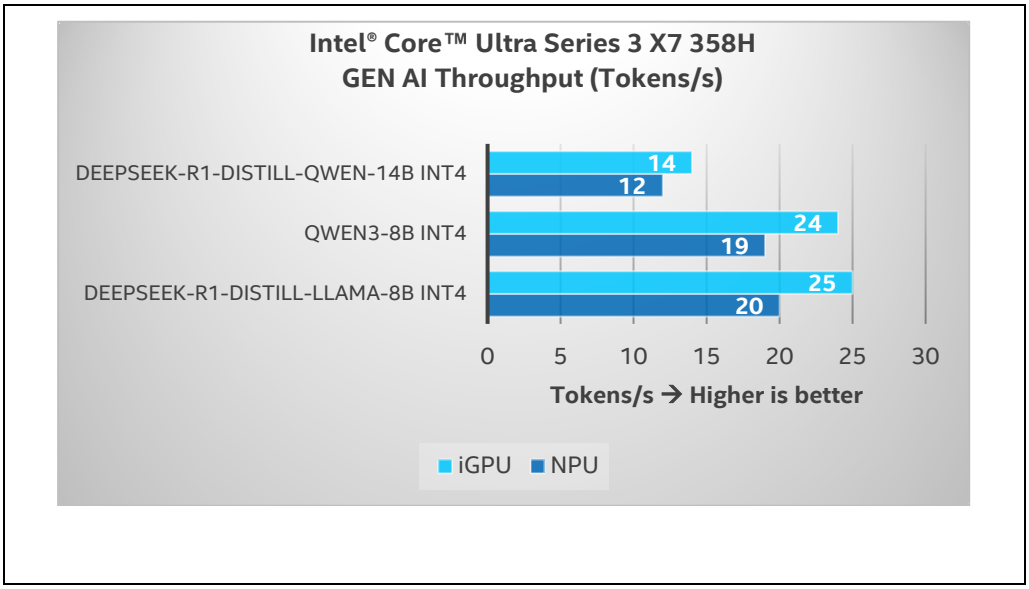
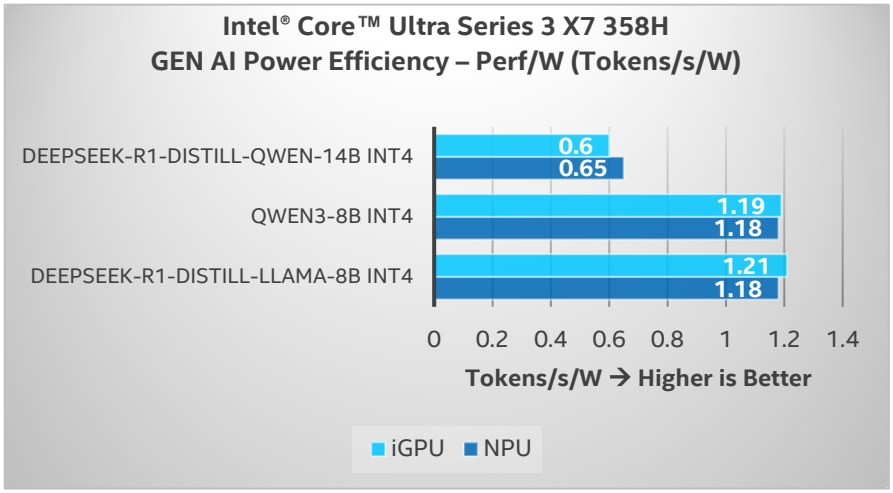


Figure 6. Power Efficiency



The Intel® Core™ Ultra Series 3 X7 358H can run AI Inferencing for **DeepSeek R1-Distill-Qwen-14 B** up to 12 tokens/s on iGPU with Batch size 1 at INT4 precision with a Power Efficiency of 0.65 Tokens/s/W.

The Intel® Core™ Ultra Series 3 X7 358H can run AI Inference for **Qwen3-8B** at up to 19 tokens/s on the iGPU with a batch size of 1 at INT4 precision with a Power Efficiency of 1.18 Tokens/s/W.

The Intel® Core™ Ultra Series 3 X7 358H can run AI Inference for **DeepSeek R1-Distill-Llama-8B** up to 20 tokens/s on the iGPU with a batch size of 1 at INT4 precision with a Power Efficiency of 1.18 Tokens/s/W.

4 Summary

Intel's drive to accelerate Edge AI system readiness with OEMs by working with the ecosystem to deliver Sized, verified, benchmarked, and scalable Intel® AI Edge Systems from partners.

The Intel® AI Edge Systems Verified Reference Blueprint – with Retail Use Case and GEN AI on Intel Core Ultra Series 3 shows verified performance data on commercial hardware that partners can demonstrate to their downstream customers. It uses Edge AI Suites – Intel Automated Self-Checkout, which provides an optimized reference pipeline, to produce Verified Performance data – number of camera streams at 14.95 fps and Power/stream. The OpenVINO™ testbench produces Verified performance for GEN AI models.

The system was tested in January 2026.

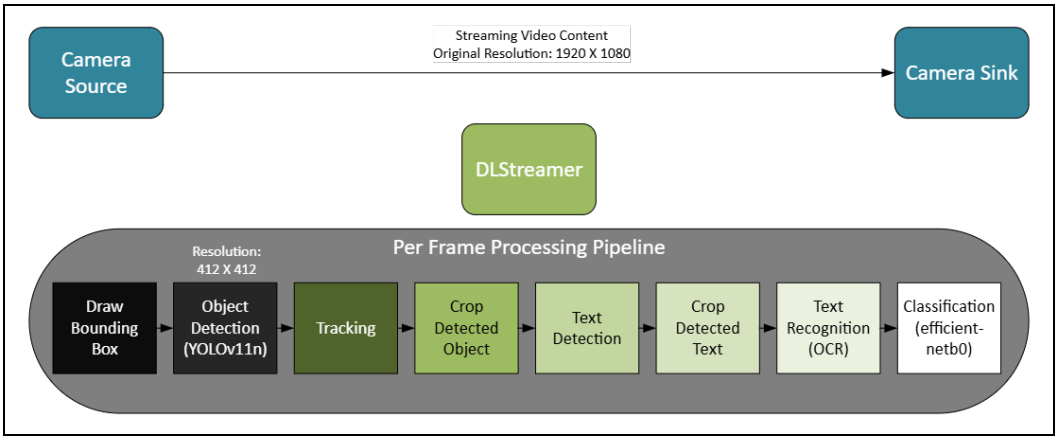
§

Appendix A Appendix

The following section provides detailed instructions for benchmarking a platform with each of the proxy workloads for Vision AI, Gen AI, and Network Security AI. The benchmarking process leverages the tools and scripts provided as part of the Intel® AI Edge Systems Verified Reference Blueprint will be available later. Reach out to your Intel Field Representative for access.

A.1 Automated Self-Checkout Test Methodology

Figure 7. Test Methodology for the Automated Self-Checkout Proxy Workload



The Intel® Automated Self-Checkout Reference Package provides critical components required to build and deploy a self-checkout use case using Intel® hardware, software, and other open-source software. It is a part of Intel® 's Edge AI Suites – a collection of building blocks, industry-specific libraries, and sample applications designed to help develop optimized AI solutions - GitHub - Open Edge Platform Edge Ai Suites.

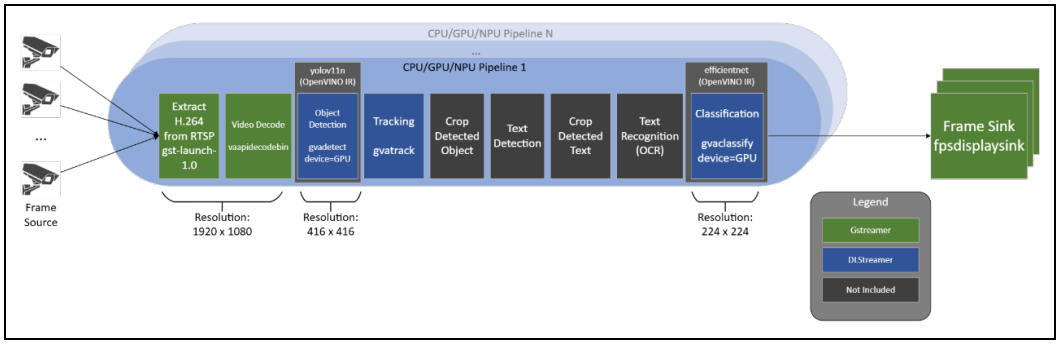
Vision workloads are large and complex, and need to go through many stages. For instance, in the pipeline shown in the figure below, the video data is ingested, pre-processed before each inferencing stage, inferenced using two models – YOLOv11 and EfficientNet, and post-processed to generate metadata along with drawing the bounding boxes for each frame. The camera source plays back pre-recorded video content, which is then processed by the media analytics pipeline. The video stream is decoded within the CPU or iGPU pipeline using software-based decodebin API calls, while for the GPU pipeline the decoding is offloaded using vaapidecodebin API calls. The video content is freely available at https://www.pexels.com .

The Intel® Automated Self-Checkout Reference makes use of Intel® Deep Learning Streamer (Intel® DL Streamer), which leverages the open-source media framework GStreamer to provide optimized media operations along with the Deep Learning Inference Engine from the OpenVINO™ Toolkit to provide optimized inference. DLStreamer accelerates the media analytics pipeline for the Vision AI use case and allows for offloading to the underlying Intel® ARC™.

The media analytics pipeline for Vision AI utilizes DLStreamer to performs object classification on the Region(s) of Interest (ROI) detected by gvadetect using the gvaclassify element and Intermediate Representation (IR) formatted object classification model. The models used for detection are in OpenVINO™ Intermediate Representation format, which is optimized for Intel® CPUs and GPUs. One advantage of the OpenVINO™ IR format is that the models can be used as-is without the need for retraining to leverage Intel® CPUs and GPUs. The Vision AI pipeline also uses object tracking for reducing the frequency of object detection and classification, thereby increasing the throughput, using gvatrack. The pipeline publishes the detection and classification results within a JSON file, which is then parsed, and the final results are reported in a log file.

Note: The GStreamer multi-media framework is used to stream video content by the frame source and the frame sink endpoints. The current release does not make use of the underlying media engines; offloading to the media engines is planned for future releases of the Intel® Automated Self-Checkout Reference.

Figure 8. Detailed Test Methodology for Retail Self-Checkout Pipeline



The test measures the number of streams that the system can sustain at the target FPS. For each test iteration, the number of camera streams is increased monotonically until the currently measured FPS value falls below the target FPS value. The number of streams is then monotonically decremented until the target FPS is met.

- Upon test completion, the results are captured for the average FPS, the cumulative FPS, and the peak number of streams achieved at the target FPS.

To run the automated self-checkout test, go to [GitHub - Intel Retail Automated Self-Checkout](#).

A.2 Gen AI Test Methodology Using OpenVINO™ with Gen AI

The Gen AI benchmark leverages the OpenVINO™ Gen AI LLM Benchmarking framework and is deployed in a containerized manner.

Note: Version OpenVINO™ 2025.4 was used for data collection in this document.

The LLM models were quantified for NPU. For more information on using LLMs on NPU, go to OpenVINO™ GenAI on NPU.

The list of models optimized for NPU is shown on the LLMs Optimized for NPU site.

§