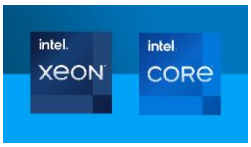
White Paper

Intel® AI Edge Systems Verified Reference Blueprint with Supermicro –Vision AI



Authors

Ecosystem Edge AI System Architect:

Abhijit Sinha

Timothy Miskell

Yuan Kuok Nee

Ecosystem Enabling Engineer:

Shin Wei Lim

Supermicro: Toby McClean **Principal Solutions Architect**

Key Contributors

Ecosystem Edge AI System Verification and Qualification Manager:

Edel Curley

1 Introduction

Intel® AI Edge Systems offers a balance between computing and AI acceleration to deliver optimal TCO, scalability, and security. AI Edge systems enable customers to jumpstart development through a hardened system foundation verified by Intel®. AI Edge systems enable the ability to add AI functionality through continuous integration into business applications for better business outcomes and streamlined implementation efforts.

To support the development of these Edge AI systems, Intel® is offering reference design and verified reference blueprints with AI system configurations that are tuned and benchmarked for different AI System types that support Edge AI Workloads. Verified reference blueprints (VRB) include Hardware BOM, Foundation Software configuration (OS, Firmware, Drivers) tested and verified with supported Software stack (software framework, libraries, orchestration management).

This document describes a verified reference blueprint architecture using the 5th Gen Intel® Xeon® Scalable processor family, and Intel® Data Center GPU Flex 140/170 or Intel® ARC™ A380/A750.

Intel ® AI Edge Systems Verified Reference Blueprint is defined in collaboration with end users and our ecosystem partners to demonstrate the value of the solution for AI Inference use cases. The solution leverages the hardened hardware, firmware, and software to allow customers to integrate on top of this known, good foundation.

Intel® AI Edge System Verified Reference Blueprint helps end users simplify design choices by bundling hardware and software pieces together while making the high performance more predictable. Some of the key benefits of the Reference Configuration based on the 5th Generation Intel® Xeon® Scalable Processor Family processor and Intel Data Center GPU Flex 140/170 or Intel® ARC™A380/A750:

- High core counts and per-core performance
- Compact, power-efficient system-on-chip platform
- Streamlined path to cloud-native operations
- Accelerated AI inference using Intel® AMX and Intel® DL Boost
- Multiple discrete GPU support to accelerate for AI inference workload
- The X^e kernel of Intel® GPUs integrates Extended Vector Engine (XVE) and Extended Matrix Engine (XMX), which accelerate AI workflow and provide powerful and real-time computing power support for AI inference at the edge
- Accelerated encryption and compression
- Platform-level security enhancements

Table of Contents

Figures

Tables

Document Revision History

Doc ID	Revision Number	Description	Date
849097	1.1	Updated Product Image	April 2025
849097	1.0	Initial release	March 2025

2 Design Compliance Requirements

This chapter focuses on the design requirements for Intel ® AI Edge Systems Verified Reference Blueprint with Supermicro SYS-111AD-WRN2 and Supermicro SYS-E403-13E-FRN2T.

2.1 Hardware Requirements

The checklists in this chapter are a guide for the platform tested as part of Intel ® AI Edge Systems Verified Reference Blueprint with Supermicro SYS-111AD-WRN2 and Supermicro SYS-E403-13E-FRN2T. The hardware specifications are detailed below.

Ingredient	Supermicro SYS-111AD-WRN2	Supermicro SYS-E403-13E-FRN2T
Processor	Intel® 14th Generation Core™ i9-14900E Processor 8 P-Cores, 16 E-Cores, 65 W or equivalent SKU	Intel® Xeon® Gold 6538N Processor at 2.1GHz, 32C/64T, 205W or higher number SKU
Total Memory	128 GB DDR5 4800 MT/s	Option 1: DRAM only configuration: 256 GB (16x 16 GB DDR5, 4800 MHz) Option 2: DRAM only configuration: 512 GB (32x 16 GB DDR5, 4800 MHz)
Storage (Boot Drive)	480 GB or equivalent boot drive	480 GB or equivalent boot drive
Storage (Capacity)	1 TB or equivalent boot drive	Minimum 1 TB or equivalent drive
Graphics	Intel® Arc™ A380 Intel® Arc™ A750	1 x Flex 170 or 2 x Flex 170 3 x Flex 140 2 x Arc™ A750
LAN on Motherboard (LOM)	1 Gbps I219-LM for Operation, Administration and Management (OAM)	10 Gbps or 25 Gbps port for video streaming 1/10 Gbps port for Management Network Interface Controller (NIC)
Product Image		

Table 1 Intel ® AI Edge Systems Verified Reference Blueprint with Supermicro SYS-111AD-WRN2 and Supermicro SYS-E403-13E-FRN2T Configuration

2.2 BIOS Settings

To meet the performance requirements for an Intel ® AI Edge Systems Verified Reference Blueprint with Supermicro SYS-111AD-WRN2 and Supermicro SYS-E403-13E-FRN2T, Intel® recommends using the BIOS settings to enable processor P-State and C-State with Intel® Turbo Boost Technology ("turbo mode") enabled. Hyperthreading is recommended to provide higher thread density. For this solution, Intel® recommends using the NFVI profile BIOS settings for on-demand Performance with power consideration.

Setting	Value
Hardware Prefetcher	Enabled
Intel® (VMX) Virtualization Technology	Enabled
Hyper-Threading	Enabled
Intel® Speed Shift Technology	Enabled
Turbo Mode	Enabled
C-States	Enabled
Enhanced C-States	Enabled
C-State Auto Demotion	C1
C-State Un-Demotion	C1

Setting	Value
MonitorMWait	Enabled
Enforce DDR Memory Frequency POR	POR
Maximum Memory Frequency	Auto
Primary Display	Auto
Internal Graphics	Auto
Graphics Clock Frequency	Max CdClock freq based on Reference Clk
VT-d	Enabled
Re-Size BAR Support	Enabled
SR-IOV Support	Enabled

Table 2 BIOS Settings for Supermicro SYS-111AD-WRN2

Setting	Value
Memory Page Policy	Adaptive
ICCP Pre-Grant Level	AMX
Speed Step (P-States)	Enable
Turbo Mode	Enable
EIST PSD Function	HW_ALL
Hardware P-States	Native Mode with No Legacy Support
Enable Monitor MWAIT	Enable
Enhanced Halt State (C1E)	Enable
Hyperthreading (ALL)	Enable
Hardware Prefetcher	Enable
LLC Prefetch	Enable
Extended APIC	Enable
Intel Virtualization Technology	Enable

Table 3 BIOS Version 2.1 Settings for Supermicro E403

2.3 Solution Architecture

Figure 1 shows the architecture diagram of Intel ® AI Edge Systems Verified Reference Blueprint with Supermicro SYS-111AD-WRN2 and Supermicro SYS-E403-13E-FRN2T. The software stack consists of a single category of AI software: Vision AI. The Vision AI application is containerized using docker.

For the Vision AI use case, we are using the Intel® Automated Self-Checkout application, which measures stream density in terms of the number of supported cameras at the target FPS, accounting for all stages within the processing pipeline. The video data is ingested and pre-processed before each inferencing step. The inference is performed using two models: YOLOv5 and EfficientNet. The YOLOv5 model does object detection, and the EfficientNet model performs Object Classification.

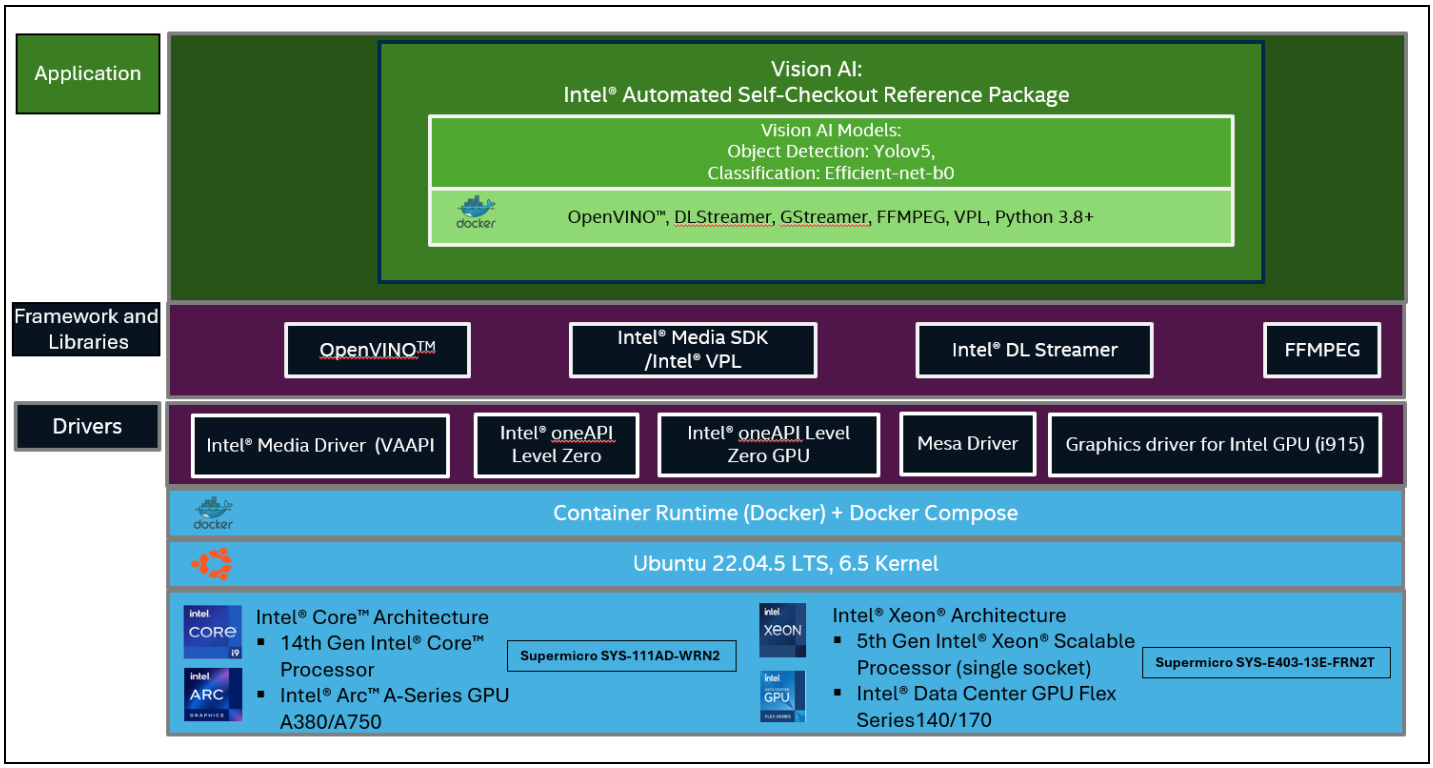


Figure 1 Architecture of the Intel ® AI Edge Systems Verified Reference Blueprint with Supermicro SYS-111AD-WRN2 and Supermicro SYS-E403-13E-FRN2T

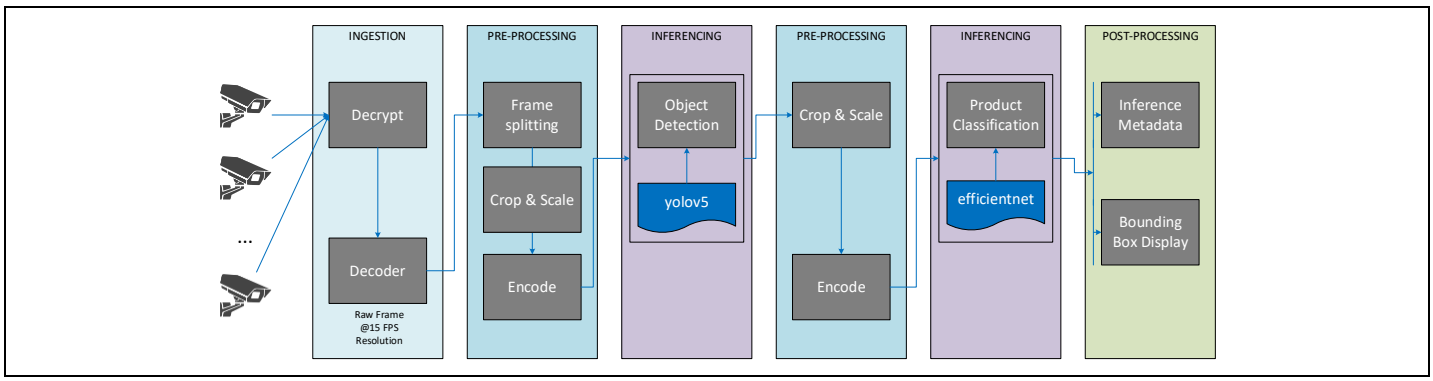


Figure 2 Test Methodology for Vison AI Pipeline

The Intel® Automated Self-Checkout Reference Package (formerly known as Retail Self-Checkout) provides critical components required to build and deploy a self-checkout use case using Intel® hardware, software, and other open-source software. Vision workloads are large and complex and need to go through many stages. For instance, in the pipeline below, the video data is ingested, pre-processed before each inferencing step, inferenced using two models - YOLOv5 and EfficientNet, and post-processed to generate metadata and show 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 input is decoded within the CPU 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 from 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 and Deep Learning Inference Engine from OpenVINO™ Toolkit to provide optimized inference. The DLStreamer accelerates the media analytics pipeline for Vision AI use cases and allows for offloading to the underlying Intel® ARC™ and Intel® Data Center Flex GPUs.

The media analytics pipeline for Vision AI utilizes DLStreamer to perform object classification on the Regions of Interest (ROIs) detected by gvadetect using gvaclassify element and Intermediate Representation (IR) formatted object classification model. The models used for detection are in OpenVINO Intermediate Representation (IR) 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 to reduce 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.

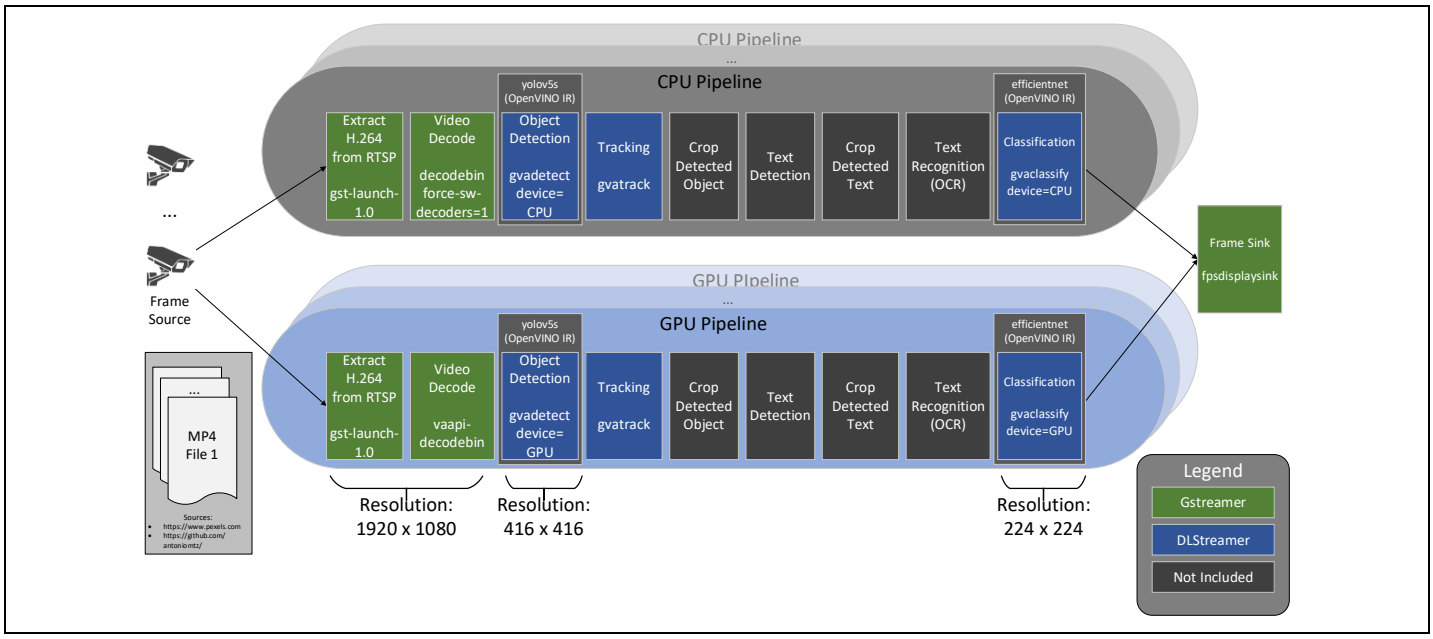


Figure 3 Detailed Test Methodology Vision AI Pipeline

The table below is a guide for assessing the conformance to the software requirements of the Intel ® AI Edge Systems Verified Reference Blueprint to ensure that the platform meets the requirements listed in the table below.

Ingredient	SW Version Details
OS	Ubuntu 22.04.4 Desktop LTS[1]
Kernel	6.5 (in-tree generic)
Microcode	Core: 0x123
	Xeon: 0x21000161
OpenVINO	2024.0.1
Docker Engine	27.1.0
Docker Compose	2.29
Intel® Level Zero for GPU	1.3.29735.27
Intel® Graphics Driver for GPU (i915)	24.3.23
Media Driver VAAPI	2024.1.5
Intel® OneVPL	2023.4.0.0-799
Mesa	23.2.0.20230712.1-2073
OpenCV	4.8.0
DLStreamer	2024.0.1
FFMPEG	2023.3.0
Python	3.8+

Table 4 SW Configuration

2.4 Platform Technology Requirements

This section lists the requirements for Intel's advanced platform technologies.

Enterprise AI requires Intel® AVX (Advance Vector Extensions) or AMX (Intel® Advance Matrix Extensions) to be enabled to reap the benefits of hardware-accelerated convolution.

2.5 Platform Security

For Intel® AI System for the Edge, it is recommended that Intel® Boot Guard Technology to be enabled so that the platform firmware is verified 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 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).

2.6 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 https://github.com/speed47/spectre-meltdown-checker.

[1] Desktop based OS selected as this kernel version includes native in-tree i915 driver support for Intel® Xe based discrete GPUs.

3 Platform Tuning and GPU Driver Setup

3.1 Additional Linux Packages Installation

3.1.1 Install Docker

Follow the instructions at https://docs.docker.com/engine/install/Ubuntu*/ to install Docker Engine on Ubuntu*.

3.1.2 Install Intel® ARC™ GPU Drivers

Refer to the following for instructions on installing the Intel® Client GPU driver: https://dgpu-docs.intel.com/driver/client/overview.html#installing-client-gpus-on-ubuntu-desktop-22-04-lts. Refer to Table 4 for a list of the installed software versions.

3.1.3 Install Intel® Data Center GPU Drivers

In case the end user is installing a server OS, then the following instructions will need to be followed. For a desktop OS refer to Section 3.1.2.

Refer to the following for instructions on installing the Intel® Data Center GPU driver: https://dgpu-docs.intel.com/driver/installation.html#ubuntu. Refer to Table 4 for a list of the installed software versions.

4 Performance Verification

This chapter aims to verify the performance metrics for the Intel ® AI Edge Systems Verified Reference Blueprint to ensure that there is no anomaly seen. Refer to the information in this chapter to ensure that the performance baseline for the platform is as expected.

The Supermicro SYS111-AD-WRN2 and Supermicro E403 solutions were tested on August 06, 2024, with the following hardware and software configurations listed in Table 4.

4.1 Memory Latency Checker (MLC)

The Memory Latency Checker can be downloaded from https://www.Intel®.com/content/www/us/en/developer/articles/tool/Intel®r-memory-latency-checker.html. Download the latest version, unzip the tarball package, go into the Linux* folder, and execute `./mlc`. Table 5 and Table 5 Memory Latency Checker below should be used as a reference for verifying the validity of the system setup.

Key Performance Metric	Supermicro SYS-111AD-WRN2	Supermicro E403
Idle Latency (ns)	123.8	150.3
Memory Bandwidths between nodes within the system (using read-only traffic type) (MB/s)	53577	260425

Table 5 Memory Latency Checker

Peak Injection Memory Bandwidth (1 MB/sec) using all threads	Supermicro SYS-111AD-WRN2	Supermicro E403
All Reads	52574	255504
3:1 Reads-Writes	50359	211857
2:1 Reads-Writes	50319	202797
1:1 Reads-Writes	50145	186870
STREAM-Triad	50231	206753
Loaded Latencies using Read-only traffic type with Delay=0 (ns)	464.78	183.11
L2-L2 HIT latency (ns)	47.0	73.6
L2-L2 HITM latency (ns)	47.3	74.7

Table 6 Peak Injection Memory Bandwidth (1 MB/sec) Using All Threads

Note: If the latency performance and memory bandwidth performance are outside the range, please verify the validity of the Platform components, BIOS settings, kernel power performance profile used, and other software components.

4.2 Vision AI Performance

Intel ® AI Edge Systems Verified Reference Blueprint with Supermicro SYS-111AD-WRN2 and Supermicro SYS-E403-13E-FRN2T configuration, the platform CPU with AMX should be able to process up to 23 streams at full HD 14.95FPS with HEVC codec, and up to 26 streams when equipped with a single Intel® Data Center Flex 170. The input video stream used for the vision AI tests has 3 objects.

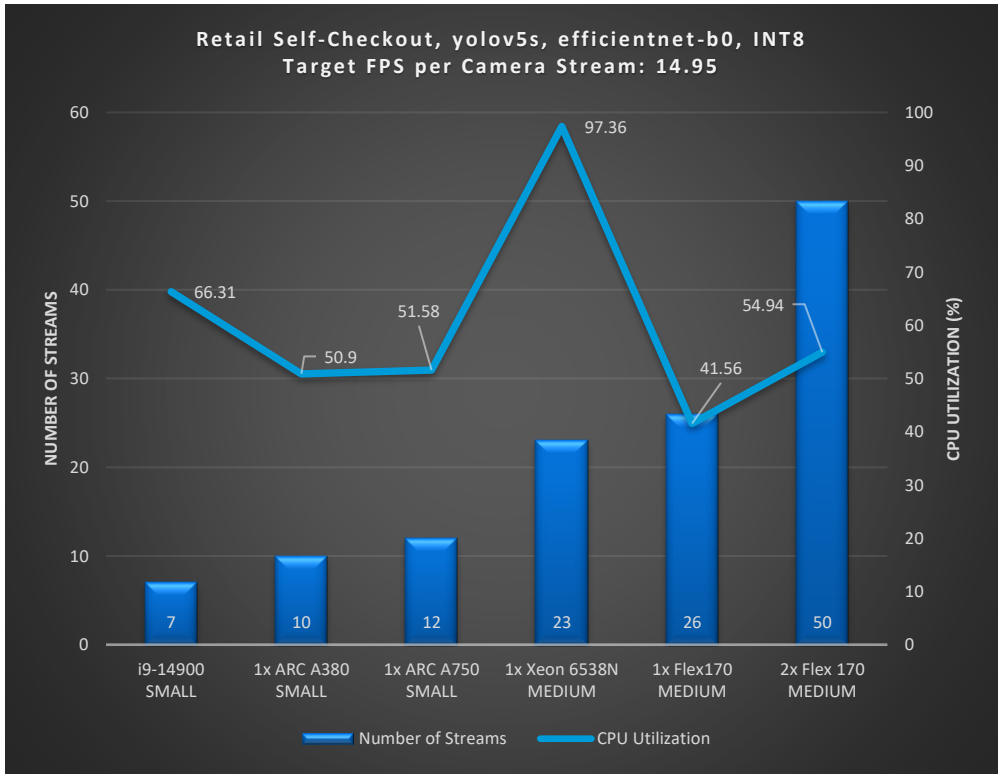


Figure 4 Vision AI Performance Graph on Intel CPU and GPU

Figure 4 presents the maximum number of supported IP camera streams at a target of 14.95 FPS for each of the benchmarked configurations. This chart illustrates the range of various potential configurations, including running exclusively on Intel® Core or Intel® Xeon® Scalable Processors and running on single or multiple Intel® discrete GPUs. In addition, the chart depicts the amount of remaining CPU utilization headroom available for running other workloads on either Intel® Core or Intel® Xeon Scalable Processors.

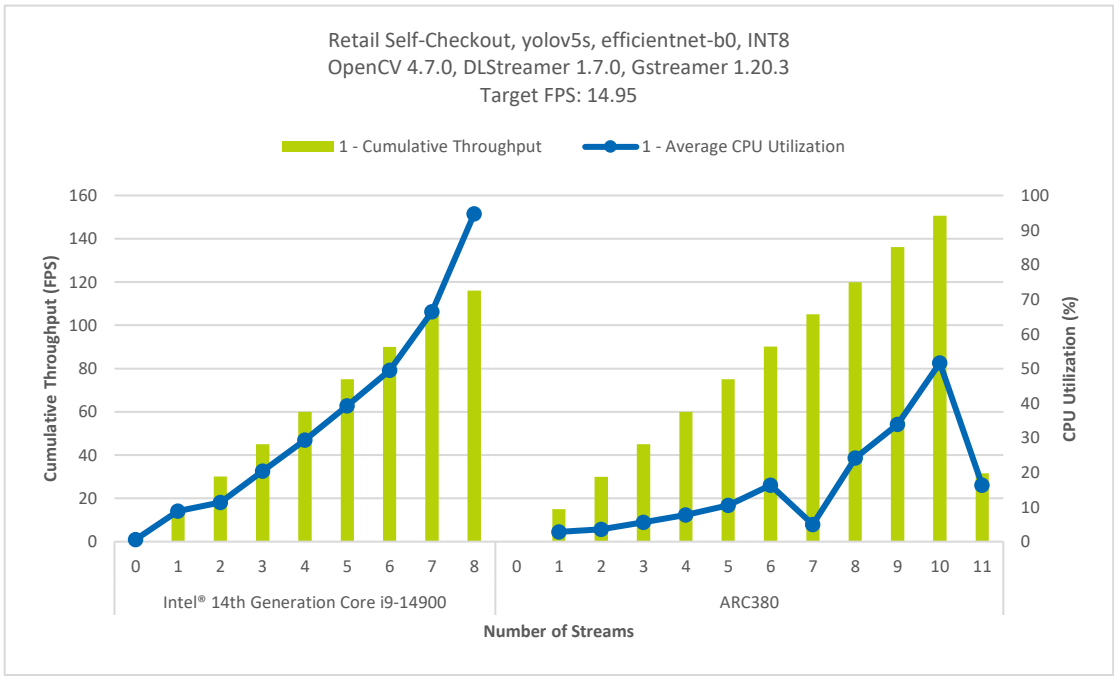


Figure 5 Vision AI Supermicro SYS-111AD-WRN2 Performance Separately on either Intel® Core® or Intel® ARC™

Figure 5 displays the detailed results comparing Intel® Core i9-14900 against Intel® ARC™ A380. In this case, Intel® Core supports up to 7 IP camera streams, while Intel® ARC™ A380 supports up to 10 camera streams. Note in this case, at 11 camera streams the GPU memory is exhausted on Intel® ARC™ A380.

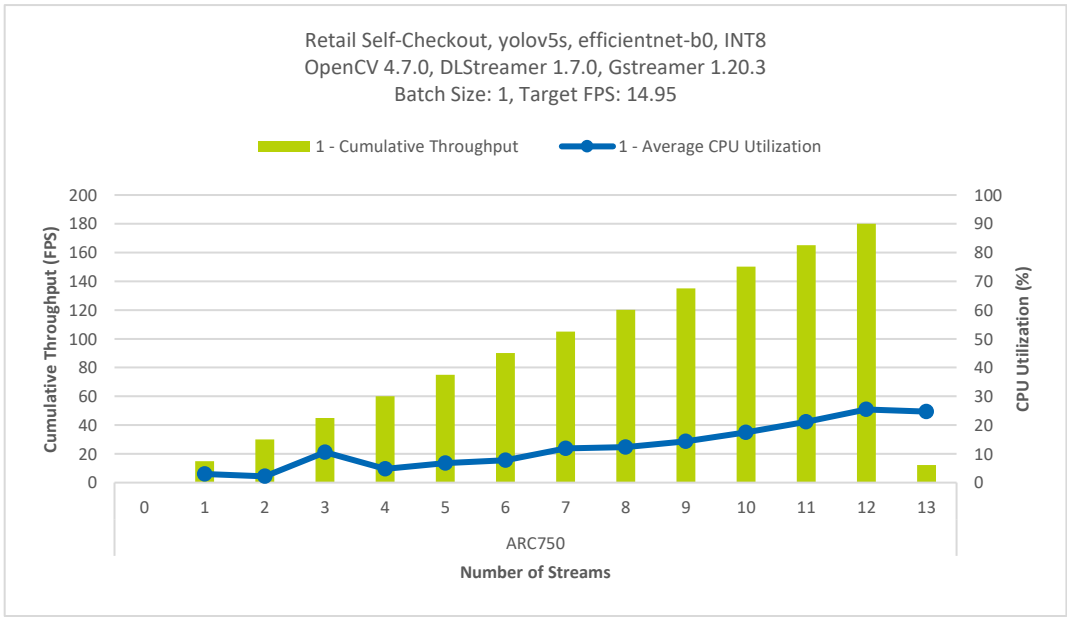


Figure 6 Vision AI Supermicro SYS-111AD-WRN2 Performance on Intel® ARC™

Figure 6 displays the detailed results for Intel® ARC™ A750. In this case, Intel® ARC™ A750 supports up to 12 camera streams. Note in this case, at 13 camera streams the GPU memory is exhausted on Intel® ARC™ A750.

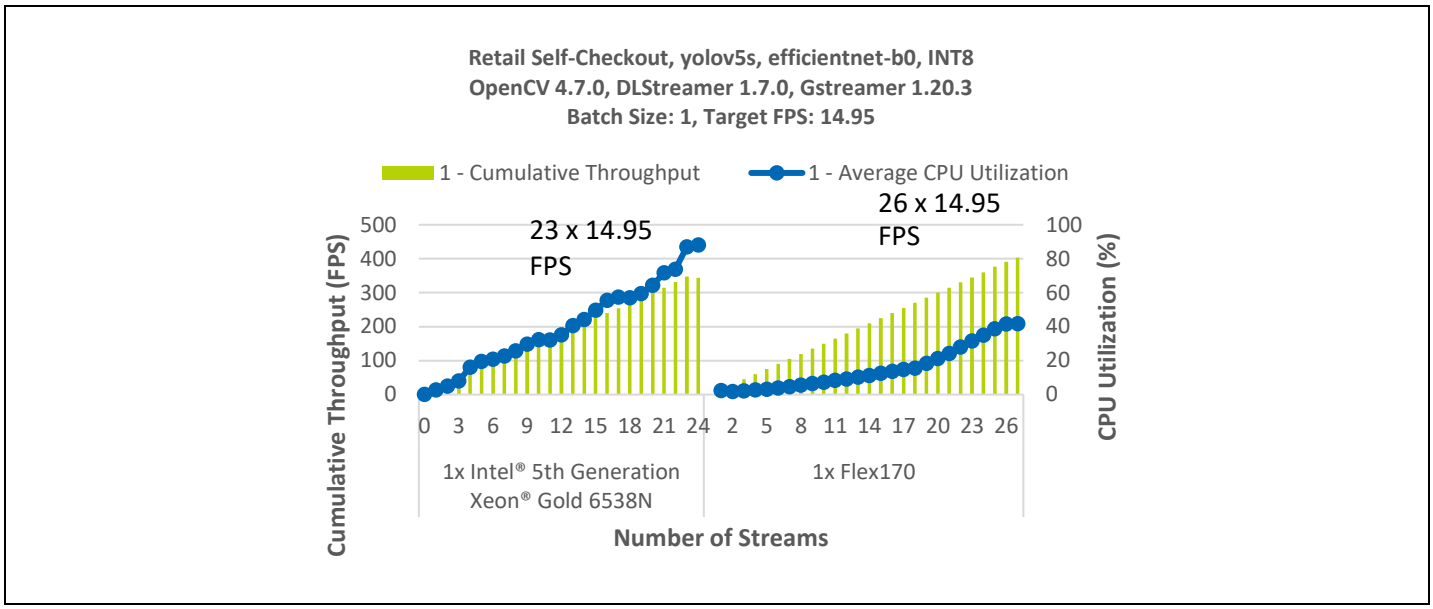


Figure 7 Vision AI Supermicro SYS-E403-13E-FRN2T Performance on Xeon® 6538N and Intel® Flex 170

Figure 7 displays the detailed results comparing 5th Generation Intel® Xeon® Scalable Processors against 1x Intel® Data Center Flex 170 GPU. In this case, 5th Generation Intel® Xeon® Scalable Processors support up to 23 IP camera streams, while 1x Intel® Data Center Flex 170 GPU up to 26 camera streams. Note in this case, at 27 camera streams the GPU memory is exhausted on the Intel® Data Center Flex 170 GPU.

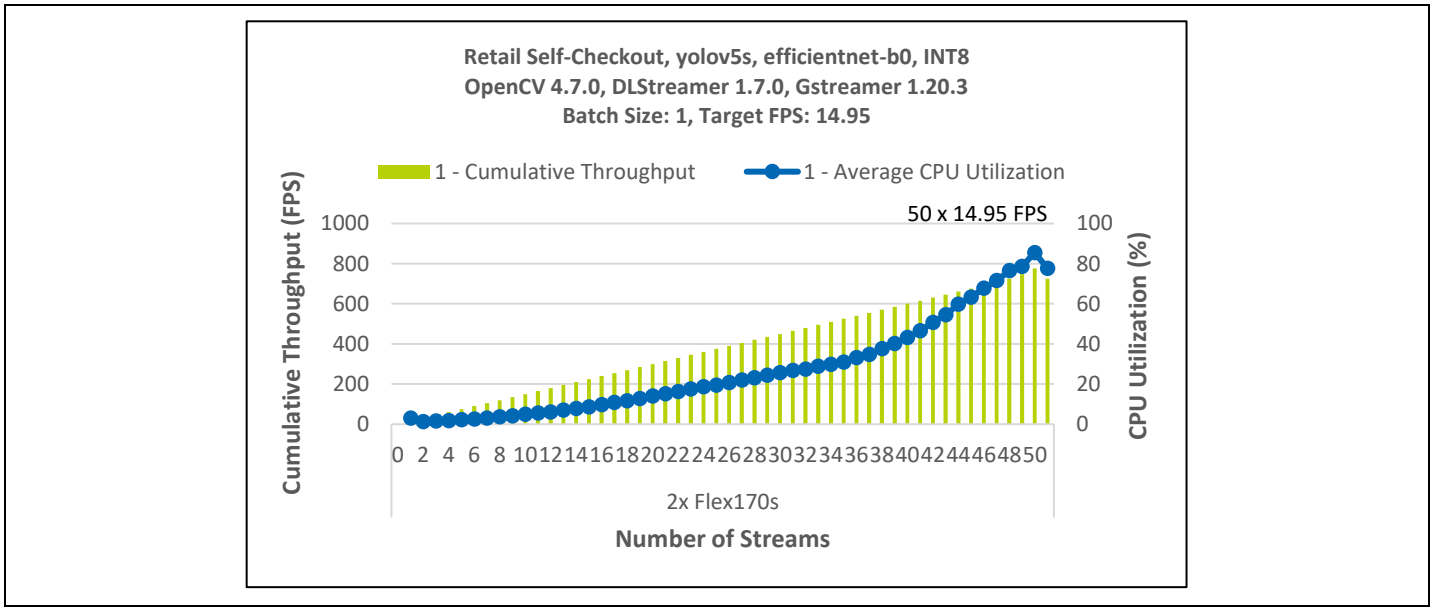


Figure 8 Vision AI Supermicro SYS-E403-13E-FRN2T Performance with Multiple Intel® Flex 170

Figure 8 displays the detailed results for 2x Intel® Data Center Flex 170 GPUs. In this case, 2x Intel® Data Center Flex 170 GPUs support up to 50 camera streams. Note in this case, at 51 camera streams the GPU memory is exhausted on the 2x Intel® Data Center Flex 170 GPUs.

5 Summary

The Intel ® AI Edge Systems Verified Reference Blueprint with Supermicro SYS-111AD-WRN2 and Supermicro SYS-E403-13E-FRN2T defined on Intel® Core i9-14900 with Intel® ARC™ on Supermicro SYS-111AD-WRN2 along with single socket 5th Gen Intel® Xeon® Scalable processors with multiple Intel® Data Center Flex GPUs on Supermicro E403 addresses the capabilities for AI Inference offering the value propositions detailed within the tables below.

Configuration	Number of IP Camera Streams
Intel® Core i9-14900	7
1x Intel® Arc™ A380	10
1x Intel® Arc™ A750	12

Table 7 Vision AI Use Case with Supermicro SYS-111AD-WRN2

Configuration	Number of IP Camera Streams
1x Intel® Xeon® 6538N	23
1x Intel® Data Center Flex 170 GPU	26
2x Intel® Data Center Flex 170 GPUs	50

Table 8 Vision AI Use Case with Supermicro SYS-E403-13E-FRN2T

This configuration, combined with architectural improvements, feature enhancements, and integrated Accelerators with high memory and IO bandwidth, provides a significant performance and scalability advantage in support of today's AI workload.

The Intel® Core and Intel® Xeon Scalable Processor platforms are optimized for AI intensive workloads coupled with Intel® ARC™ and Intel® Data Center Flex GPUs.

1 Notices & Disclaimers

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. © 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.

2 Configuration

Test by Intel as of 6th August 2024
See Hardware Configuration - Table 1