Intel® ESQ for AI Edge Systems Report Manual

Application Note

March 2026

Contents

Figures

Tables

Revision History

Date	Revision	Description
April 2026	2026.1.0	• First Release

§

1.0 Introduction

Intel® Edge System Qualification (ESQ) is a self-service validation framework developed by Intel to qualify AI Edge platforms against defined performance, reliability, compatibility, and interoperability requirements. It is the standardized mechanism used to qualify systems for the AI Edge ecosystem, enabling reduced developer friction, seamless hardware–software integration, and accelerated market adoption.

More on Intel® ESQ:

- ESQ for AI Edge System : https://builders.intel.com/ecosystem-engagement/solution-hub/systems/edge-systems-qualification/ai-edge-systems
- ESQ Framework in GitHub : https://github.com/open-edge-platform/edge-system-qualification
- Edge AI Catalog : https://builders.intel.com/ecosystem-engagement/solution-hub/edge-ai-catalog/partner-showcase

This document provides guidance on test methodologies and result interpretation based on the generated report from ESQ from which removes guess work for edge device positioning and recommendations for the intended audience and vertical use-cases.

1.1 Terminology

Table 1. Terminology

Term	Description
Intel® ESQ	Intel® Edge System Qualification
GPU / iGPU	Graphics Processing Unit / iGPU refers to internal GPU device
NPU	Neural Processing Unit
CPU	Central Processing Unit
FPS/fps	Frames per second
NVR	Network Video Recorder

Table 2. Reference Documents

Document	Document No./Location
Intel® ESQ for AI Edge System Release Note	Link
Intel® ESQ (Edge System Positioning) Source Code	Link
Edge Compass	RDC#

1.2 Benefits

By understanding how each system performs and behaves under AI workloads through ESQ, Intel and its ecosystem gain clear, data-driven insights into system capabilities and readiness. ESQ removes ambiguity from system evaluation, enabling consistent qualification across platforms while establishing a common language for performance, reliability, and interoperability. This clarity helps ecosystem partners confidently position their solutions, articulate customer value, and scale offerings aligned to real-world AI Edge requirements.

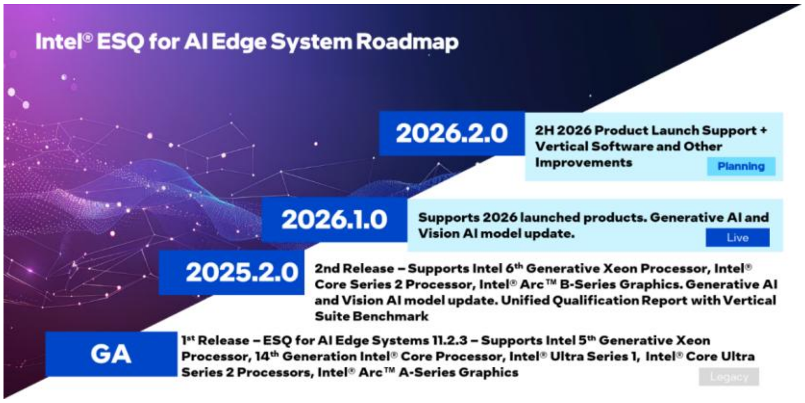
- Ecosystem partners no longer need to guess how to position their systems for Vision AI or Generative AI use cases
- Clear qualification results enable partners to communicate the return on investment and scalability potential to end customers
- Standardized reporting reduces developer friction by aligning hardware and software expectations upfront
- Consistent system behavior improves interoperability across the AI Edge ecosystem
- Qualified systems accelerate customer adoption and shorten time to market

§

2.0 Intel® ESQ for AI Edge Systems Qualification

Intel® Edge System Qualification (ESQ) is a self-service validation framework developed by Intel to ensure that hardware platforms meet strict performance, reliability, and compatibility standards for running AI workloads at the edge. It is designed for system builders, OEMs, and partners who want to certify their edge systems for Vision AI and Generative AI use cases, such as object detection, classification, and large language model inference.

ESQ Software Product Roadmap:



Intel® ESQ for AI Edge System Roadmap

2026.2.0 — 2H 2026 Product Launch Support + Vertical Software and Other Improvements — Planning

2026.1.0 — Supports 2026 launched products. Generative AI and Vision AI model update. — Live

2025.2.0 — 2nd Release – Supports Intel 6th Generative Xeon Processor, Intel® Core Series 2 Processor, Intel® Arc™ B-Series Graphics. Generative AI and Vision AI model update. Unified Qualification Report with Vertical Suite Benchmark

GA — 1st Release – ESQ for AI Edge Systems 11.2.3 – Supports Intel 5th Generative Xeon Processor, 14th Generation Intel® Core Processor, Intel® Ultra Series 1, Intel® Core Ultra Series 2 Processors, Intel® Arc™ A-Series Graphics — Legacy

2.1 How Intel® (ESQ) Works

2.1.1 Step 1 – Provision the System :

System manufacturers prepare and configure their target edge platform with the intended hardware, operating system, drivers, and software stack to reflect the real deployment environment. This ensures qualification results accurately represent system behavior for AI workloads at the edge.

2.1.2 Step 2 – Download ESQ from GitHub

The ESQ framework is downloaded from Intel's public GitHub repository, providing access to the latest validated qualification profiles, AI workloads, and test configurations required for AI Edge system evaluation.

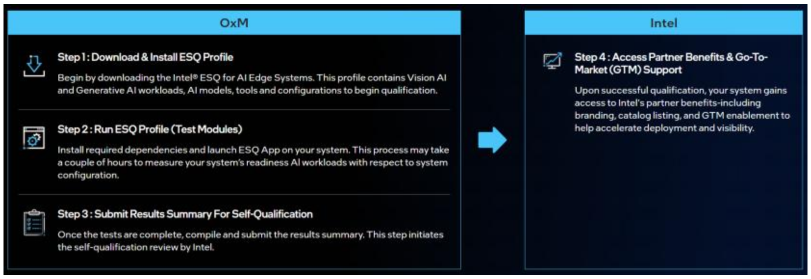
2.1.3 Step 3 – Run ESQ Qualification

ESQ is executed on the provisioned system to run qualification and applicable AI workload tests. These tests assess system readiness for Vision AI and Generative AI use cases, producing a standardized ESQ report that captures configuration details, test outcomes, and performance indicators.

2.1.4 Step 4 – Submit Results to Intel

The generated ESQ report is submitted to Intel for review as part of the self-qualification process. Upon successful qualification, systems may become eligible for ecosystem benefits such as catalog visibility, Intel endorsement, and go-to-market enablement—helping accelerate adoption and partner engagement.

.



Step 1 : Download & Install ESQ Profile

Begin by downloading the Intel® ESQ for AI Edge Systems. This profile contains Vision AI and Generative AI workloads, AI models, tools and configurations to begin qualification.

Step 2 : Run ESQ Profile (Test Modules)

Install required dependencies and launch ESQ App on your system. This process may take a couple of hours to measure your system's readiness AI workloads with respect to system configuration.

Step 3 : Submit Results Summary For Self-Qualification

Once the tests are complete, compile and submit the results summary. This step initiates the self-qualification review by Intel.

Step 4 : Access Partner Benefits & Go-To-Market (GTM) Support

Upon successful qualification, your system gains access to Intel's partner benefits-including branding, catalog listing, and GTM enablement to help accelerate deployment and visibility.

2.2 Edge System Qualification Criteria

Intel® ESQ uses simple qualification acceptance criteria for variety of SKUs and system configurations developed by hardware manufacturers.

By referencing Edge Compass (5Q Roadmap) document, we apply the same AI performance swimlanes and validated system passing criteria as follows:

2.2.1 Edge AI Performance Swimlanes and System Requirements

Category	Processors	Specifications	Storage	Discrete GPU Options
Scalable Performance Graphics Media	Xeon-Based: Intel® Xeon® 6 Processors, Intel® Xeon® 600 W Processors, 5th Gen Intel® Xeon® Scalable Processors Core-Based: Intel® Core Ultra Series 2 or Intel® Core™ Series 2 Processors	Xeon-Based: SKU: Dual and Single Socket Memory: 128 GB DDR5 (Dual Channel) or more Core-Based: Minimum System Memory: 64GB DDR5 (Dual Channel) or more	1TB	Intel® Arc™ B-Series Graphics or Intel® Arc™ Pro B-Series Graphics
Scalable Performance	Intel® Xeon® 6 Processors, 5th Gen Intel® Xeon® Scalable Processors, Intel® Xeon® 600 W Processors	SKU: Dual and Single Socket Minimum System Memory: 128 GB DDR5 (Dual Channel) or more	1TB	-
Efficiency Optimized	Intel® Core™ Ultra Processor Series 2	Graphics: iGPU with 7 Xe-Cores or more SKU: Single Socket Minimum System Memory: 32 GB DDR5 (Dual Channel) or more	512 GB	-
Mainstream	Intel® Core™ Series 2 Processors	Minimum System Memory: 32 GB DDR5 (Dual Channel) or more	512 GB	-
Entry	Intel® Processor for Desktop, Intel® Processor X-series, Intel® Processor N-series	Minimum System Memory: 16 GB DDR5 (Dual Channel) or more	512 GB	-

Note: Based on 2025 and 2026 CPU / Platform launch.

2.2.2 AI Model Configuration and System Passing Criteria

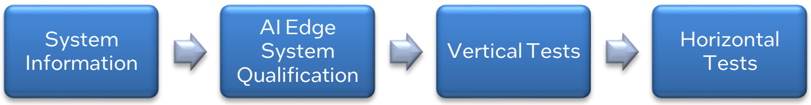
As of 2026.1.0 Intel® ESQ will be using the following Generative AI and Vision AI Models and respective passing criteria for edge system qualification.

AI Edge Swimlane	Generative AI Models	Passing for Gen AI	Vision AI Models	Passing for Vision AI
Scalable Performance Graphics and Media	Qwen3*-32B INT4	10 Tokens /s	multi-stream 1080p15 H.264 YOLO11n INT8 &EfficientNet-B0 INT8	=> 40 streams
Scalable Performance	DeepSeek*-R1-Distill-Qwen-14B INT4	10 Tokens /s	multi-stream 1080p15 H.264 YOLO11n INT8 &EfficientNet-B0 INT8	=> 30 streams
Efficiency Optimized	DeepSeek*-R1-Distill-Qwen-7B INT4	10 Tokens /s	multi-stream 1080p15 H.264 YOLO11n INT8 &EfficientNet-B0 INT8	=> 25 streams
Mainstream	Phi-4-mini-reasoning Phi4-3.8B INT4	10 Tokens /s	multi-stream 1080p15 H.264 YOLO11n INT8 &EfficientNet-B0 INT8	=> 12 streams
Entry	DeepSeek*-R1-Distill-Qwen-1.5B INT4	10 Tokens /s	multi-stream 1080p15 H.264 YOLO11n INT8 &EfficientNet-B0 INT8	=> 4 streams

Note: This is the models and passing criteria set as of 2026.1.0 release.

2.3 ESQ Test and Report Generation Methodology

By default, when ESQ run command is initiated on target system, the test script will run each test as shown below sequentially unless skip options are explicitly applied.

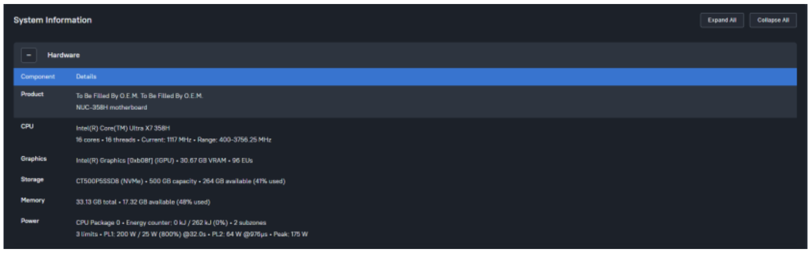


System Information → AI Edge System Qualification → Vertical Tests → Horizontal Tests

Note: For more information to run specific tests, go to : https://open-edge-platform.github.io/edge-system-qualification/main/getting-started/suites/

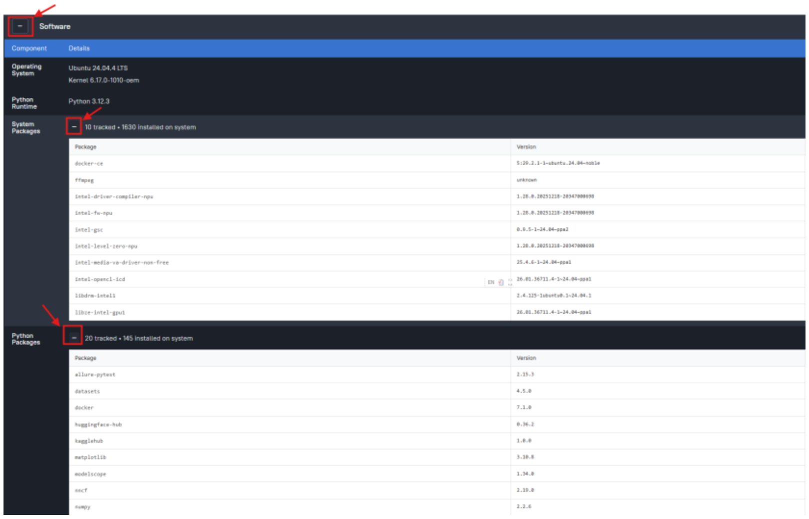
2.3.1 System Information

Your report begins with a complete breakdown of the specific configuration tested. This includes the **Hardware** and **Software** specifications that serve as the baseline for all qualification metrics.



Component	Details
Product	To Be Filled By O.E.M. To Be Filled By O.E.M. NUC-358H motherboard
CPU	Intel(R) Core(TM) Ultra X7 358H 16 cores • 16 threads • Current: 1117 MHz • Range: 400-3756.25 MHz
Graphics	Intel(R) Graphics [0xb08f] (iGPU) • 30.67 GB VRAM • 96 EUs
Storage	CT500P5SSD8 (NVMe) • 500 GB capacity • 264 GB available (41% used)
Memory	33.13 GB total • 17.32 GB available (48% used)
Power	CPU Package 0 • Energy counter: 0 kJ / 262 kJ (0%) • 2 subzones 3 limits • PL1: 200 W / 25 W (800%) @32.0s • PL2: 64 W @976μs • Peak: 175 W

At the software section, the user can view details of System and Python packages that ESQ uses during script runtime.



Component	Details
Operating System	Ubuntu 24.04.4 LTS Kernel 6.17.0-1010-oem
Python Runtime	Python 3.12.3
System Packages	− 10 tracked • 1630 installed on system

Package	Version
docker-ce	5:29.2.1-1~ubuntu.24.04~noble
ffmpeg	unknown
intel-driver-compiler-npu	1.28.0.20251218-20347000698
intel-fw-npu	1.28.0.20251218-20347000698
intel-gsc	0.9.5-1~24.04~ppa2
intel-level-zero-npu	1.28.0.20251218-20347000698
intel-media-va-driver-non-free	25.4.6-1~24.04~ppa1
intel-opencl-icd	26.01.36711.4-1~24.04~ppa1
libdrm-intel1	2.4.125-1ubuntu0.1~24.04.1
libze-intel-gpu1	26.01.36711.4-1~24.04~ppa1

Component	Details
Python Packages	− 20 tracked • 145 installed on system

Package	Version
allure-pytest	2.15.3
datasets	4.5.0
docker	7.1.0
huggingface-hub	0.36.2
kagglehub	1.0.0
matplotlib	3.10.8
modelscope	1.34.0
nncf	2.19.0
numpy	2.2.6

2.3.2 AI Edge System Qualification

This test is mandatory to qualify Intel based edge system as an Intel® AI Edge qualified system and positioning in the right performance swimlanes as discussed earlier. When initiated, the script will run the following tests.

2.3.2.1 Generative AI Test

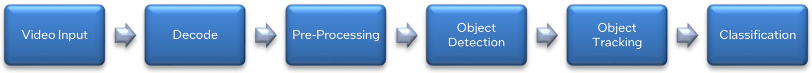
The purpose of this test is to measure the system's inferencing throughput for generative AI models on Intel CPU, GPU and NPU Device.

Details on Gen AI models and passing metrics used in this test are found here

2.3.2.2 Vision AI Test

The purpose of this test is to measure the highest number of 'streams of vision ai pipeline' a system can spawn should the workload be run concurrently on CPU, iGPU or dGPU or NPU device simultaneously.

A single stream of Vision AI pipeline consists of both detection and classification workloads as shown below.



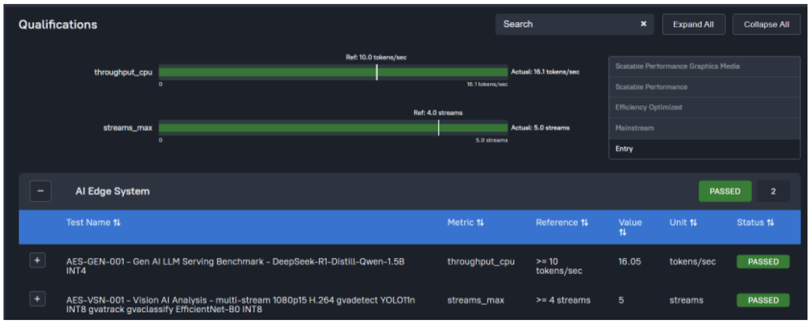
- **Input & Decode:** 1080p15 H.264 video.
- **Pre-processing:** Prepares frames for inference.
- **Object Detection:** YOLO11n INT8 running at ~14.95 fps.
- **Tracking:** Zero-term object tracking.
- **Classification:** EfficientNet-B0 INT8 running at ~14.95 fps.

Complete pipeline: Media Decode (1080p15 H.264) + Pre-processing + YOLO11n INT8 (object detection) @ 14.95fps + Zero-term tracking + EfficientNet-B0 INT8 (classification) @ 14.95fps

2.3.3 Example of AI Edge System Qualification Passing Report.

Here are examples of passing qualification report for each swimlanes

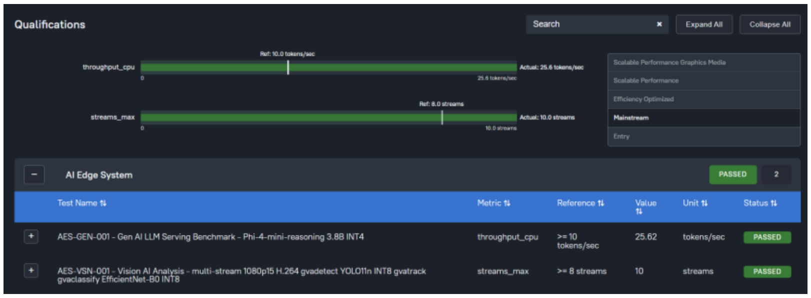
2.3.3.1 Entry System Qualification Passing Report



	Test Name ⇅	Metric ⇅	Reference ⇅	Value ⇅	Unit ⇅	Status ⇅
+	AES-GEN-001 - Gen AI LLM Serving Benchmark - DeepSeek-R1-Distill-Qwen-1.5B INT4	throughput_cpu	>= 10 tokens/sec	16.05	tokens/sec	PASSED
+	AES-VSN-001 - Vision AI Analysis - multi-stream 1080p15 H.264 gvadetect YOLO11n INT8 gvatrack gvaclassify EfficientNet-B0 INT8	streams_max	>= 4 streams	5	streams	PASSED

Note: This is the models and passing criteria set as of 2026.1.0 release.

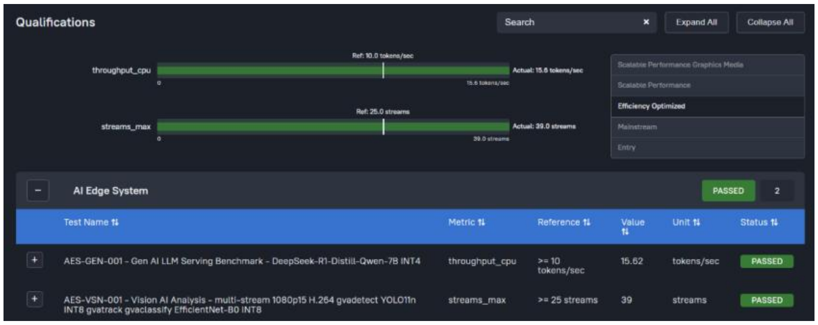
2.3.3.2 Mainstream System Qualification Passing Report



	Test Name ⇅	Metric ⇅	Reference ⇅	Value ⇅	Unit ⇅	Status ⇅
+	AES-GEN-001 - Gen AI LLM Serving Benchmark - Phi-4-mini-reasoning 3.8B INT4	throughput_cpu	>= 10 tokens/sec	25.62	tokens/sec	PASSED
+	AES-VSN-001 - Vision AI Analysis - multi-stream 1080p15 H.264 gvadetect YOLO11n INT8 gvatrack gvaclassify EfficientNet-B0 INT8	streams_max	>= 8 streams	10	streams	PASSED

Note: This is the models and passing criteria set as of 2026.1.0 release.

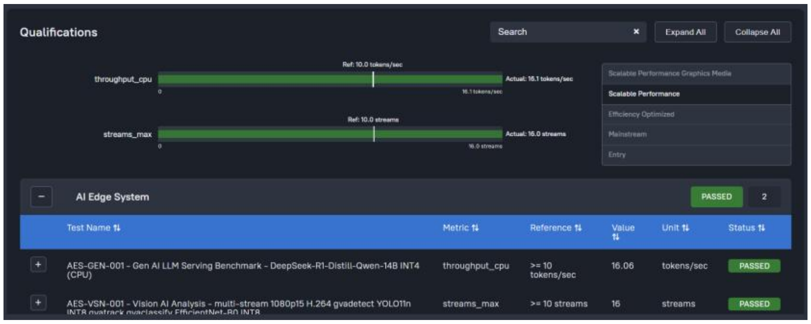
2.3.3.3 Efficiency Optimized System Qualification Passing Report



	Test Name ⇅	Metric ⇅	Reference ⇅	Value ⇅	Unit ⇅	Status ⇅
+	AES-GEN-001 - Gen AI LLM Serving Benchmark - DeepSeek-R1-Distill-Qwen-7B INT4	throughput_cpu	>= 10 tokens/sec	15.62	tokens/sec	PASSED
+	AES-VSN-001 - Vision AI Analysis - multi-stream 1080p15 H.264 gvadetect YOLO11n INT8 gvatrack gvaclassify EfficientNet-B0 INT8	streams_max	>= 25 streams	39	streams	PASSED

Note: This is the models and passing criteria set as of 2026.1.0 release.

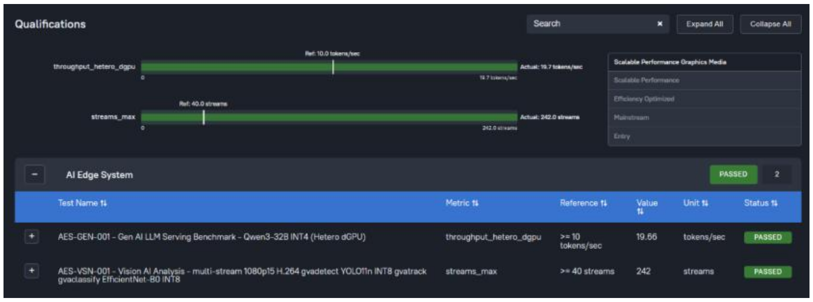
2.3.3.4 Scalable Performance System Qualification Passing Report



	Test Name ⇅	Metric ⇅	Reference ⇅	Value ⇅	Unit ⇅	Status ⇅
+	AES-GEN-001 - Gen AI LLM Serving Benchmark - DeepSeek-R1-Distill-Qwen-14B INT4 (CPU)	throughput_cpu	>= 10 tokens/sec	16.06	tokens/sec	PASSED
+	AES-VSN-001 - Vision AI Analysis - multi-stream 1080p15 H.264 gvadetect YOLO11n INT8 gvatrack gvaclassify EfficientNet-B0 INT8	streams_max	>= 10 streams	16	streams	PASSED

Note: This is the models and passing criteria set as of 2026.1.0 release.

2.3.3.5 Scalable Performance Graphics Media Qualification System Passing Report



	Test Name ⇅	Metric ⇅	Reference ⇅	Value ⇅	Unit ⇅	Status ⇅
+	AES-GEN-001 - Gen AI LLM Serving Benchmark - Qwen3-32B INT4 (Hetero dGPU)	throughput_hetero_dgpu	>= 10 tokens/sec	19.66	tokens/sec	PASSED
+	AES-VSN-001 - Vision AI Analysis - multi-stream 1080p15 H.264 gvadetect YOLO11n INT8 gvatrack gvaclassify EfficientNet-B0 INT8	streams_max	>= 40 streams	242	streams	PASSED

Note: These are the models, results and passing criteria set as of 2026.1.0 release.

§

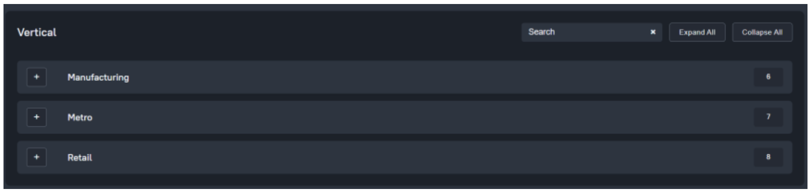
3.0 Vertical & Horizontal Tests

Upon successful completion of AI Edge System qualification tests, the ESQ will continue to run the remaining test cases.

- Vertical - Referencing sample workloads from use-cases retail, manufacturing and smart cities and transportation
- Horizontal – Auxiliary test-cases that helps for ecosystem partners to understand Gen AI or Vision AI performance on individual target device (CPU, iGPU, dGPU or NPU) and system related metrics not limited to memory, GPU performance, or system power.

3.1 Vertical Tests

When ESQ command is run without any skip override command, the ESQ report will have a vertical section that can be expanded for each vertical as shown below.



+	Manufacturing	6
+	Metro	7
+	Retail	8

Note: Vertical Tests as of 2026.1.0 release.

The following sub-sections will discuss workloads for vertical and tests that are covered in detail.

3.1.1 Manufacturing

Manufacturing test includes two sample use cases: pallet defect detection and weld anomaly detection. Pallet defect detection often relies on object detection to locate both the pallet and any visible defects. A lightweight model like YOLOX-Tiny at FP32 precision is well-suited for this task because it balances accuracy with speed and does not require aggressive quantization that might reduce reliability in cluttered warehouse environments. Since pallets and surrounding equipment move relatively slowly from frame to frame, especially during unloading at the receiving dock, the application can tolerate a lower FPS.

Weld anomaly detection, on the other hand, uses classification to identify specific defect types in the welding process. EfficientNet-B0 balances accuracy and speed, especially when deployed at FP16, making it a strong candidate for high-frequency inspection where the weld evolves quickly in each frame. Because the process is continuous and fast, higher FPS is better to ensure no manufacturing defects or anomalies are missed.

In deployment, factories use industrial-grade RGB or greyscale cameras—sometimes thermal or multispectral—depending on the inspection needs. Vision AI workloads for both use cases can run on CPUs for flexibility, GPUs for high-throughput inference, NPUs for power-efficient edge deployment or a combination, allowing manufacturers to tailor performance to the constraints of their production environments.

3.1.1.1 Pallet Defect Detection

Pallet Defect Detection provides automated quality control with AI-driven vision systems. It enables real-time pallet condition monitoring by running inference workflows across multiple AI models. It connects multiple video streams from warehouse cameras to AI-powered pipelines, all operating efficiently on a single industrial PC. This solution enhances logistics efficiency and inventory management by detecting defects before they impact operations. It is a cloud-native application composed of microservices, using pre-trained deep learning models for video analysis. This sample application offers the following:

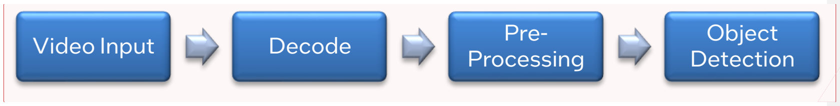
- High-speed data exchange with low-latency compute.
- AI-assisted defect detection in real-time as pallets are received at the warehouse.
- Data processing at the edge for data privacy and efficient use of bandwidth.
- Interconnected warehouse delivery analytics for quick and informed decision making and tracking.

Goal: Locate pallets and identify visible defects in real-time.

Model: YOLOX-Tiny (FP32 precision). FP32 is chosen to balance accuracy and speed without aggressive quantization that could reduce reliability.

Performance Profile: Tolerates lower FPS as pallets move slowly.

Pipeline: 480p30 Decode → Pre-processing → YOLOX-TINY FP32 → Sink.



Commented [AS1]: Its only object detection it seems. I don't see any tracking and classification happening. The diagram should showcase the actual.

Complete Pipeline: Media Decode (480p30 H.264) + Pre-processing + YOLOX-TINY FP32 (object detection) @ 29.5fps + Sink

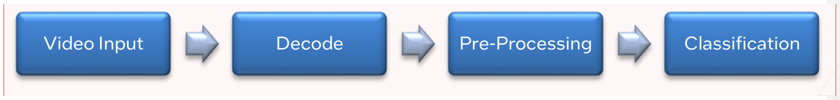
3.1.1.2 Weld Porosity Classification

This pipeline demonstrates how AI-driven analytics enable edge devices to monitor weld quality. The sample app detects anomalous weld patterns and alerts operators for timely intervention, ensuring proactive maintenance, safety, and operational efficiency. No more failures and unplanned downtime.

Goal: Detect specific defect types in continuous welding processes.

Model: EfficientNet-B0 (FP16 precision).

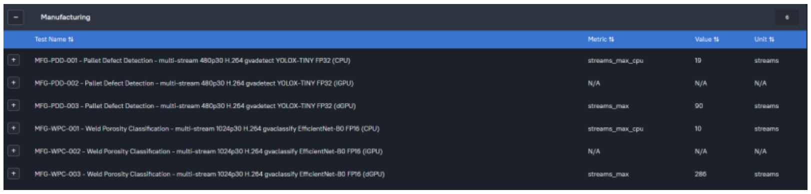
Performance Profile: Requires higher FPS to ensure fast-moving anomalies are not missed.



Commented [AS2]: Same comment as above, check this diagram. Description mentions only classification, diagram shows detection tracking and classification.

Complete Pipeline: Media Decode (1024p30 H.264) + Pre-processing + EfficientNet-B0 FP16 (full-frame classification) @ 29.5fps + Sink

3.1.1.3 Example of Manufacturing Results



Manufacturing

Test Name	Metric	Value	Unit
MFG-PDD-001 - Pallet Defect Detection - multi-stream 480p30 H.264 gvadetect YOLOX-TINY FP32 (CPU)	streams_max_cpu	19	streams
MFG-PDD-002 - Pallet Defect Detection - multi-stream 480p30 H.264 gvadetect YOLOX-TINY FP32 (iGPU)	N/A	N/A	N/A
MFG-PDD-003 - Pallet Defect Detection - multi-stream 480p30 H.264 gvadetect YOLOX-TINY FP32 (dGPU)	streams_max	90	streams
MFG-WPC-001 - Weld Porosity Classification - multi-stream 1024p30 H.264 gvaclassify EfficientNet-B0 FP16 (CPU)	streams_max_cpu	10	streams
MFG-WPC-002 - Weld Porosity Classification - multi-stream 1024p30 H.264 gvaclassify EfficientNet-B0 FP16 (iGPU)	N/A	N/A	N/A
MFG-WPC-003 - Weld Porosity Classification - multi-stream 1024p30 H.264 gvaclassify EfficientNet-B0 FP16 (dGPU)	streams_max	286	streams

Note: Results as of 2026.1.0 release.

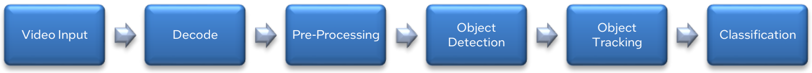
3.1.2 Retail

For retail, reference pipelines used are automated self-checkout and loss prevention pipelines. More details are available on the Intel Developer Focused Webpage.

3.1.2.1 Automated Self-Checkout

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. This reference implementation provides a pre-configured automated self-checkout pipeline that is optimized for Intel® hardware.
Function: Uses a pre-configured pipeline optimized for Intel hardware to manage checkout processes.

Workflow: 1080p15 Decode → YOLO11n INT8 (Detection) → Tracking → EfficientNet-B0 INT8 (Classification).



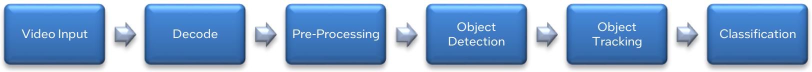
Complete Pipeline:
Media Decode (1080p15 H.264) + Pre-processing + YOLO11n INT8 (object detection) 14.95fps + Zero-term tracking + EfficientNet-B0 INT8 (classification) @ 14.95fps (with batch size 1)

3.1.2.2 Loss Prevention

The Intel® Loss Prevention Reference Package is designed to help with this. It provides the essential components needed to develop and deploy a loss prevention solution using Intel® hardware, software, and open-source tools. This reference implementation includes a pre-configured pipeline that's optimized for Intel® hardware, simplifying the setup of an effective computer vision-based loss prevention system for retailers.

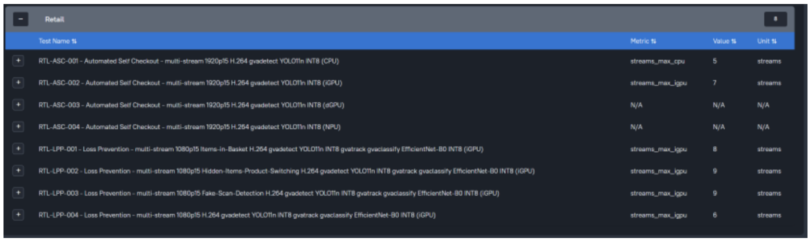
Function: Detects anomalies like "sweet hearting" or missed scans using a similar pipeline structure to self-checkout.

Workflow: Identical to the Self-Checkout pipeline (YOLO11n + EfficientNet-B0).



Complete Pipeline: Media Decode (1080p15 H.264) + Pre-processing + YOLO11n INT8 (object detection) 14.95fps + Zero-term tracking + EfficientNet-B0 INT8 (classification) @ 14.95fps (with batch size 1)

3.1.2.3 Example of Retail Results



Retail

	Test Name ⇅	Metric ⇅	Value ⇅	Unit ⇅
+	RTL-ASC-001 - Automated Self Checkout - multi-stream 1920p15 H.264 gvadetect YOLO11n INT8 (CPU)	streams_max_cpu	5	streams
+	RTL-ASC-002 - Automated Self Checkout - multi-stream 1920p15 H.264 gvadetect YOLO11n INT8 (iGPU)	streams_max_igpu	7	streams
+	RTL-ASC-003 - Automated Self Checkout - multi-stream 1920p15 H.264 gvadetect YOLO11n INT8 (dGPU)	N/A	N/A	N/A
+	RTL-ASC-004 - Automated Self Checkout - multi-stream 1920p15 H.264 gvadetect YOLO11n INT8 (NPU)	N/A	N/A	N/A
+	RTL-LPP-001 - Loss Prevention - multi-stream 1080p15 Items-in-Basket H.264 gvadetect YOLO11n INT8 gvatrack gvaclassify EfficientNet-B0 INT8 (iGPU)	streams_max_igpu	8	streams
+	RTL-LPP-002 - Loss Prevention - multi-stream 1080p15 Hidden-Items-Product-Switching H.264 gvadetect YOLO11n INT8 gvatrack gvaclassify EfficientNet-B0 INT8 (iGPU)	streams_max_igpu	9	streams
+	RTL-LPP-003 - Loss Prevention - multi-stream 1080p15 Fake-Scan-Detection H.264 gvadetect YOLO11n INT8 gvatrack gvaclassify EfficientNet-B0 INT8 (iGPU)	streams_max_igpu	9	streams
+	RTL-LPP-004 - Loss Prevention - multi-stream 1080p15 H.264 gvadetect YOLO11n INT8 gvatrack gvaclassify EfficientNet-B0 INT8 (iGPU)	streams_max_igpu	6	streams

3.1.3 Metro

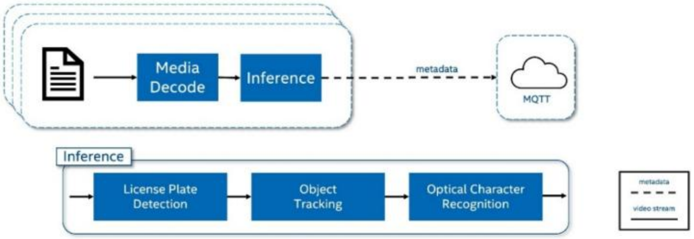
The Metro Benchmark establishes a standardized baseline for visual analytics, typically falling within 20-30% of a device's maximum potential.

3.1.3.1 Proxy Pipeline 1: License Plate Recognition (LPR)

The License Plate Recognition (LPR) Proxy Pipeline is engineered to closely simulate the end-to-end operational workflow of a modern license plate recognition system, providing a rigorous and realistic benchmark for edge AI platforms. This pipeline begins by ingesting video streams from multiple cameras, reflecting real-world deployment scenarios such as parking management, access control, or traffic monitoring. It employs a specialized object detection model to accurately localize license plates within each video frame. Once license plates are detected, the pipeline applies an optical character recognition (OCR) model to extract alphanumeric characters from the localized regions. This two-stage process ensures high accuracy in both detection and text recognition.

The benchmark evaluates the pipeline throughput as the key performance indicators. **Function:** Ingests video, detects plates, and applies OCR for character recognition.

KPI: Pipeline throughput.

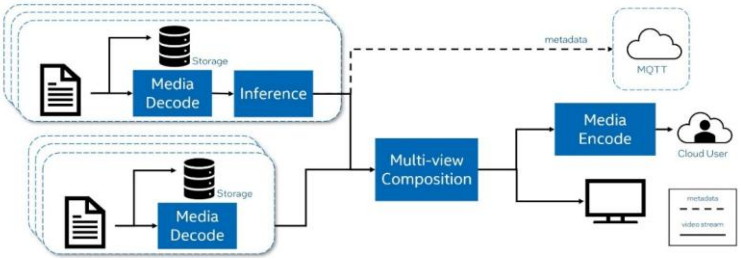


3.1.3.2 Proxy Pipeline 2: Smart NVR

The Smart NVR proxy pipeline is designed to closely emulate the comprehensive media and video analytics workload typical of a modern Smart NVR (Network Video Recorder) system. This pipeline encompasses several critical functions: video ingestion from multiple sources, storage, and the composition of up to *n* simultaneous video streams into a 4x4, 5x5, or 6x6 multi-view display for real-time monitoring on a locally attached screen. In addition to these core media operations, the pipeline subjects a subset of the video channels to advanced AI-driven video analytics, enabling automated detection, classification, and event recognition. This holistic approach provides a rigorous, real-world assessment of both the system's media handling capabilities and its AI inferencing performance, offering valuable insights for benchmarking, system optimization, and downstream marketing differentiation.

The benchmark measures the maximum number of AI channels that the system can support for a variety of inference models and configuration.
Function: Simulates a Network Video Recorder by combining video ingestion, storage, and multi-view composition (e.g., 4x4 or 6x6 grids).

KPI: Maximum number of supported AI channels.



3.1.3.3 Proxy Pipeline 3: Headed Visual AI Pipeline

The Headed Visual AI Benchmark evaluates real-world edge AI performance for multi-stream video analytics with visual display output. It simulates a realistic workload where multiple video streams are simultaneously processed with AI inference (object detection and classification) while composing and displaying results on physical monitors. This benchmark is designed to assess system capabilities for surveillance, digital signage, smart retail, and other visual AI applications that require both AI processing and real-time display.

The benchmark tests different complexity levels based on AI models and platform capabilities:

3.1.3.4 Light AI Workload (yolov5s-416)

Single object detection model (YOLOv5s at 416x416 resolution)

- Entry-level AI processing
- Suitable for basic detection tasks

3.1.3.5 Medium AI Workload (yolov5m-416)

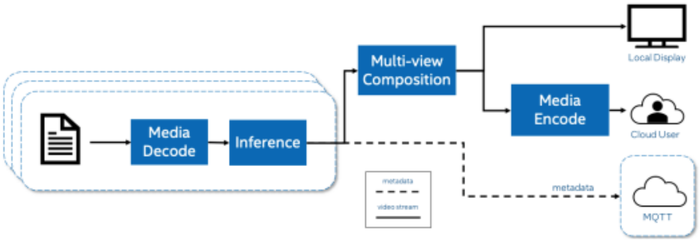
- Medium-sized object detection model (YOLOv5m at 416x416 resolution)
- Increased model complexity and accuracy

3.1.3.6 Heavy AI Workload (yolov5m-416 + efficientnet-b0)

- Combined detection and classification pipeline
- Object detection followed by classification on detected ROIs
- Represents complex multi-stage AI pipelines
- On Intel® Core Ultra platforms, classification leverages NPU as co-processor

Function: It offers a comprehensive evaluation of real-world visual AI performance with display requirements, making it an ideal tool for qualifying edge systems for deployment in video analytics applications.

KPI: Primary performance metric indicating how many simultaneous video streams can be processed while maintaining the target FPS.



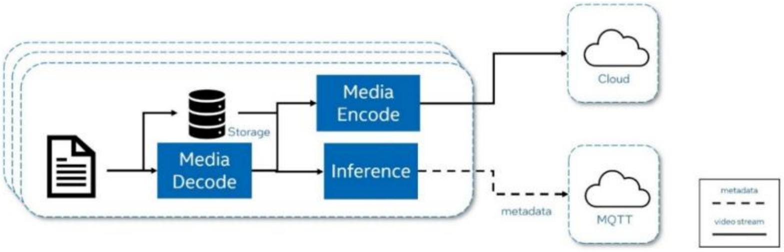
3.1.3.7 Proxy Pipeline 4: VSaaS Gateway

The VSaaS Gateway Proxy Pipeline is designed to approximate the end-to-end workload of a typical Video Surveillance as a Service (VSaaS) gateway. This pipeline manages the local storage of ingested video streams from multiple cameras, applies real-time AI inference to analyze and extract actionable insights from the video data, and performs transcoding to optimize video files for efficient transmission to cloud-based storage or analytics platforms. By replicating these core functions—video ingestion, storage management, AI-powered analytics, and adaptive transcoding—the pipeline provides a comprehensive benchmark for evaluating system performance, scalability, and reliability in real-world VSaaS deployments. This approach enables stakeholders to assess how well a platform can handle demanding surveillance workloads, ensuring robust operation and seamless integration with cloud services.

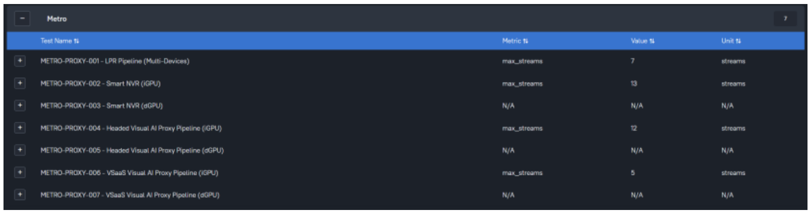
- The benchmark measures the maximum number of channels that the system can support for a variety of inference models and configuration

  Function: Manages local storage, AI inference, and transcoding for cloud transmission.

  KPI: Maximum number of supported channels.



3.1.3.8 Example of Metro (Smart Cities and Transportation) Results



Metro 7

	Test Name ⇅	Metric ⇅	Value ⇅	Unit ⇅
+	METRO-PROXY-001 - LPR Pipeline (Multi-Devices)	max_streams	7	streams
+	METRO-PROXY-002 - Smart NVR (iGPU)	max_streams	13	streams
+	METRO-PROXY-003 - Smart NVR (dGPU)	N/A	N/A	N/A
+	METRO-PROXY-004 - Headed Visual AI Proxy Pipeline (iGPU)	max_streams	12	streams
+	METRO-PROXY-005 - Headed Visual AI Proxy Pipeline (dGPU)	N/A	N/A	N/A
+	METRO-PROXY-006 - VSaaS Visual AI Proxy Pipeline (iGPU)	max_streams	5	streams
+	METRO-PROXY-007 - VSaaS Visual AI Proxy Pipeline (dGPU)	N/A	N/A	N/A

Note: Results as of 2026.1.0 release.

3.2 Horizontal Tests

This section details the horizontal test results. Horizontal testing evaluates the raw capability of the system across few fundamental domains: Generative AI, Computer Vision, and GPU System Memory. These tests provide a baseline of performance that applies to broad categories of workloads

3.2.1 Gen AI Test

Comprehensive Gen AI tests to evaluate the throughput and efficiency of running large language models (LLMs) on various hardware (CPU, iGPU, dGPU, NPU) with INT4 precision for inference
Throughput: Look for the **tokens/sec** metric. Higher is better, indicating the system can generate text faster.

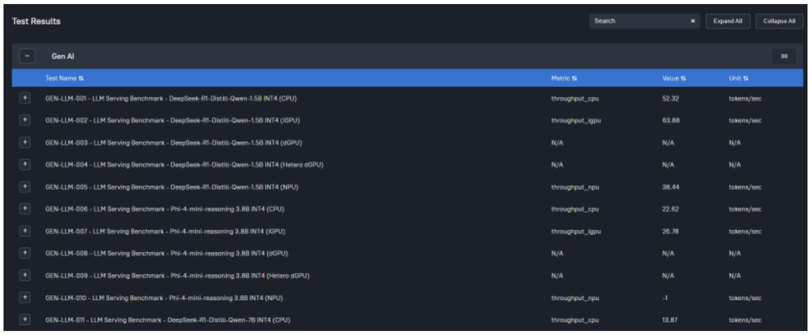
Device Comparison: Use these results to determine which hardware (NPU vs. dGPU vs. CPU) is most efficient for your specific model size (e.g., 7B vs. 70B).

See complete Gen AI Models used 2026.1.0 release here

3.2.1.1 How to Read Results:

- Throughput (tokens/sec): Higher is better; shows how many tokens can be processed per second.
- Device comparison: Use results to see which hardware is most efficient for your target model size.
- Precision: All models use INT4 for inference, balancing speed and accuracy.

3.2.1.2 Horizontal Gen AI Results



	Test Name ⇅	Metric ⇅	Value ⇅	Unit ⇅
+	GEN-LLM-001 - LLM Serving Benchmark - DeepSeek-R1-Distill-Qwen-1.5B INT4 (CPU)	throughput_cpu	52.32	tokens/sec
+	GEN-LLM-002 - LLM Serving Benchmark - DeepSeek-R1-Distill-Qwen-1.5B INT4 (iGPU)	throughput_igpu	63.88	tokens/sec
+	GEN-LLM-003 - LLM Serving Benchmark - DeepSeek-R1-Distill-Qwen-1.5B INT4 (dGPU)	N/A	N/A	N/A
+	GEN-LLM-004 - LLM Serving Benchmark - DeepSeek-R1-Distill-Qwen-1.5B INT4 (Hetero dGPU)	N/A	N/A	N/A
+	GEN-LLM-005 - LLM Serving Benchmark - DeepSeek-R1-Distill-Qwen-1.5B INT4 (NPU)	throughput_npu	38.44	tokens/sec
+	GEN-LLM-006 - LLM Serving Benchmark - Phi-4-mini-reasoning 3.8B INT4 (CPU)	throughput_cpu	22.62	tokens/sec
+	GEN-LLM-007 - LLM Serving Benchmark - Phi-4-mini-reasoning 3.8B INT4 (iGPU)	throughput_igpu	26.78	tokens/sec
+	GEN-LLM-008 - LLM Serving Benchmark - Phi-4-mini-reasoning 3.8B INT4 (dGPU)	N/A	N/A	N/A
+	GEN-LLM-009 - LLM Serving Benchmark - Phi-4-mini-reasoning 3.8B INT4 (Hetero dGPU)	N/A	N/A	N/A
+	GEN-LLM-010 - LLM Serving Benchmark - Phi-4-mini-reasoning 3.8B INT4 (NPU)	throughput_npu	-1	tokens/sec
+	GEN-LLM-011 - LLM Serving Benchmark - DeepSeek-R1-Distill-Qwen-7B INT4 (CPU)	throughput_cpu	13.87	tokens/sec

Note: Horizontal Gen AI as of 2026.1.0 release.

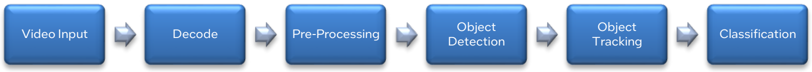
3.2.2 Vision AI Test

Comprehensive Vision AI test to understand the system's ability to handle complex, multi-stream video analytics pipelines, verify how many streams that the actual frame rate met the requirement - 14.95 FPS

3.2.2.1 Multi-Stream Pipelines with Multiple AI Stages :

Workload: Multi-stream 1080p30 H.265 Decoding → Object Detection (**YOLO11n INT8**) → Tracking → Classification (**ResNet50 INT8**).

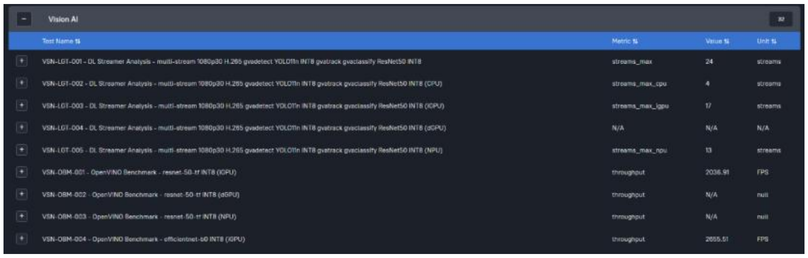
3.2.2.2 Vision VRB Profile:



Workload*:* Multi-stream 1080p15 H.264 Decoding → Object Detection (**YOLO11n INT8**) → Tracking → Classification (**EfficientNet-B0 INT8**).

Stream Max: Represents the maximum number of video streams (from cameras, files, or other media inputs) that the entire system can concurrently process while maintaining the specified target frame rate during DL Streamer pipeline execution. Reflects the combined throughput capability of all available compute resources (CPU, iGPU, dGPU, NPU).

Stream _Max_CPU/iGPU/NPU/dGPU : This metric shows the maximum number of simultaneous video cameras the system can process at the target frame rate, with results reported for CPU, iGPU, dGPU, and NPU individually.



− Vision AI 32

	Test Name ⇅	Metric ⇅	Value ⇅	Unit ⇅
+	VSN-LGT-001 - DL Streamer Analysis - multi-stream 1080p30 H.265 gvadetect YOLO11n INT8 gvatrack gvaclassify ResNet50 INT8	streams_max	24	streams
+	VSN-LGT-002 - DL Streamer Analysis - multi-stream 1080p30 H.265 gvadetect YOLO11n INT8 gvatrack gvaclassify ResNet50 INT8 (CPU)	streams_max_cpu	4	streams
+	VSN-LGT-003 - DL Streamer Analysis - multi-stream 1080p30 H.265 gvadetect YOLO11n INT8 gvatrack gvaclassify ResNet50 INT8 (iGPU)	streams_max_igpu	17	streams
+	VSN-LGT-004 - DL Streamer Analysis - multi-stream 1080p30 H.265 gvadetect YOLO11n INT8 gvatrack gvaclassify ResNet50 INT8 (dGPU)	N/A	N/A	N/A
+	VSN-LGT-005 - DL Streamer Analysis - multi-stream 1080p30 H.265 gvadetect YOLO11n INT8 gvatrack gvaclassify ResNet50 INT8 (NPU)	streams_max_npu	13	streams
+	VSN-OBM-001 - OpenVINO Benchmark - resnet-50-tf INT8 (iGPU)	throughput	2036.91	FPS
+	VSN-OBM-002 - OpenVINO Benchmark - resnet-50-tf INT8 (dGPU)	throughput	N/A	null
+	VSN-OBM-003 - OpenVINO Benchmark - resnet-50-tf INT8 (NPU)	throughput	N/A	null
+	VSN-OBM-004 - OpenVINO Benchmark - efficientnet-b0 INT8 (iGPU)	throughput	2655.51	FPS

Note: Horizontal Vision AI tests as of 2026.1.0 release.

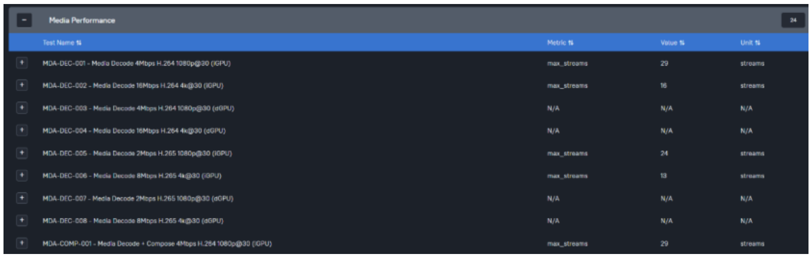
3.2.3 Media Performance

As we noted throughout this document, media decode function plays an important role in preparing the frame prior to inferencing. This test provides comprehensive media decode profiles at various input resolutions and frame rates.

3.2.3.1 How to Read Results:

- Higher number of streams is better; shows how many streams can be spawned at simultaneously at a given time.

3.2.3.2 Media Performance Result



− Media Performance 24

	Test Name ⇅	Metric ⇅	Value ⇅	Unit ⇅
+	MDA-DEC-001 - Media Decode 4Mbps H.264 1080p@30 (iGPU)	max_streams	29	streams
+	MDA-DEC-002 - Media Decode 16Mbps H.264 4k@30 (iGPU)	max_streams	16	streams
+	MDA-DEC-003 - Media Decode 4Mbps H.264 1080p@30 (dGPU)	N/A	N/A	N/A
+	MDA-DEC-004 - Media Decode 16Mbps H.264 4k@30 (dGPU)	N/A	N/A	N/A
+	MDA-DEC-005 - Media Decode 2Mbps H.265 1080p@30 (iGPU)	max_streams	24	streams
+	MDA-DEC-006 - Media Decode 8Mbps H.265 4k@30 (iGPU)	max_streams	13	streams
+	MDA-DEC-007 - Media Decode 2Mbps H.265 1080p@30 (dGPU)	N/A	N/A	N/A
+	MDA-DEC-008 - Media Decode 8Mbps H.265 4k@30 (dGPU)	N/A	N/A	N/A
+	MDA-COMP-001 - Media Decode + Compose 4Mbps H.264 1080p@30 (iGPU)	max_streams	29	streams

Note: Horizontal Media Performance tests as of 2026.1.0 release.

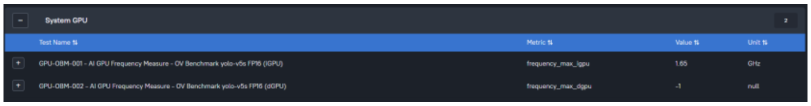
3.2.4 System GPU

Whether its media decode or inferencing vision or Gen AI, an unthrottled GPU frequency and resources is crucial to deliver optimal edge AI performance. This section covers if GPU is running at intended frequency during Vision AI workload.

3.2.4.1 How to Read Results:

- Compare GPU frequency to product datasheets.

3.2.4.2 System GPU Result



System GPU 2

	Test Name ⇅	Metric ⇅	Value ⇅	Unit ⇅
+	GPU-OBM-001 - AI GPU Frequency Measure - OV Benchmark yolo-v5s FP16 (IGPU)	frequency_max_igpu	1.65	GHz
+	GPU-OBM-002 - AI GPU Frequency Measure - OV Benchmark yolo-v5s FP16 (dGPU)	frequency_max_dgpu	-1	null

Note: Horizontal System GPU tests as of 2026.1.0 release.

3.2.5 System Memory

System memory benchmarks evaluate the memory bandwidth and latency of the system, which are critical for AI workloads that process large amounts of data. System Memory tests include:

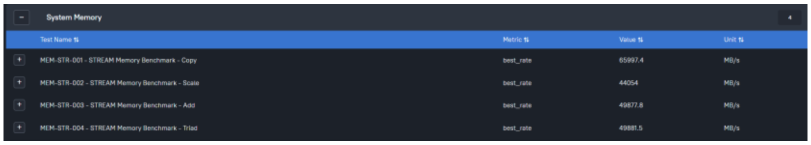
- Copy
- Scale
- Add
- Triad

3.2.5.1 How to Read Results:

Higher bandwidth and lower latency is considered as better. These numbers help you understand if the memory subsystem is a bottleneck for AI workloads.

- Best Rate (MB/s): Indicates the maximum memory bandwidth.
- Average/Min/Max Time (s): Shows consistency and latency.

3.2.5.2 System Memory Result



System Memory 4

	Test Name ⇅	Metric ⇅	Value ⇅	Unit ⇅
+	MEM-STR-001 - STREAM Memory Benchmark - Copy	best_rate	65997.4	MB/s
+	MEM-STR-002 - STREAM Memory Benchmark - Scale	best_rate	44054	MB/s
+	MEM-STR-003 - STREAM Memory Benchmark - Add	best_rate	49877.8	MB/s
+	MEM-STR-004 - STREAM Memory Benchmark - Triad	best_rate	49881.5	MB/s

Horizontal System Memory tests as of 2026.1.0 release.

4.0 Conclusion

Through Intel® Edge System Qualification (ESQ), we are delivering tangible value to the Edge AI ecosystem by enabling transparent, data-driven evaluation of AI inferencing and Vision AI performance. ESQ provides a comprehensive view of how systems behave under real AI workloads, exposing key metrics such as bandwidth, latency, and throughput that help identify bottlenecks and optimize configurations. This results in more reliable, high-performance AI pipelines—critical for demanding use cases such as multi-stream video analytics, large model inference, and production-scale Vision AI deployments.

For OEMs and ODMs, these standardized performance insights also support clearer, more credible discussions with downstream partners, enabling confident system positioning and informed decision-making across the AI infrastructure.

The ESQ Report User Manual guides readers through these benefits, helping partners interpret qualification results, extract actionable insights, and fully realize the value of ESQ in delivering scalable, interoperable, and market-ready Edge AI systems.

§