White Paper

Industry / Solution Focus Area

Intel® AI Edge Systems Verified Reference Blueprint with Dell

Generative AI and Vision AI Performance on Dell PowerEdge R760 with 5th Generation Intel® Xeon® Scalable Processors.

Authors

Ecosystem Edge AI System Architect:

Abhijit Sinha

Cloud Software Architect

Timothy Miskell

Strategic Alliances Director (Dell):

Suresh Ramasamy

1 Introduction

The Intel® AI Edge Systems Verified Reference Blueprint (VRB) with Dell represents an optimized commercial AI system delivered and sold through OEMs in the ecosystem. This is a commercial platform verified-configured, tuned, and benchmarked using Intel's reference AI software application on Intel® hardware to deliver optimal performance for Enterprise applications.

The Intel® AI Edge Systems Verified Reference Blueprint with Dell offers a balance between computing and AI acceleration to deliver optimal TCO, scalability, and security. The Intel® VRB for AI on Dell PowerEdge R760 enables enterprises to jumpstart development through a hardened system foundation verified by Intel® and enables 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 solutions built on top of the Intel® AI Edge Systems VRB with Dell, Intel® is offering reference design and verified reference blueprint with AI system configurations that are tuned and benchmarked for different AI system types that support Enterprise AI use cases. This includes the Verified Reference Blueprint (VRB), including Hardware BOM along with the Foundation Software configuration (OS, Firmware, Drivers) that has been tested and verified with a supported Software stack (software framework, libraries, and orchestration management).

This document describes a verified reference blueprint using the 5th Gen Intel® Xeon® Scalable processor family.

When network operators, service providers, cloud service providers, or enterprise infrastructure companies choose an Intel® AI Edge Systems Verified Reference Blueprint with Dell configuration, they should be able to deploy the AI workload more securely and efficiently than ever before. End users spend less time, effort, and expense evaluating hardware and software options. The Intel® AI Edge Systems VRB with Dell helps end users simplify design choices by bundling hardware and software pieces together while making the high level of performance more predictable.

This VRB for Computer Vision and GEN AI is based on a single-node architecture that provides an environment to execute multiple AI workloads that are commonly deployed at the edge, such as the "Intel® Automated Self-Checkout Reference Package" and "Generative AI".

All Intel® AI Edge Systems VRB with Dell solutions feature a workload-optimized stack tuned to take full advantage of an Intel® Architecture (IA) foundation. To meet the requirements, the OEM/ODM platform must meet a performance threshold that represents a premium customer experience.

The configuration for the VRB for Vision AI and GEN AI is comprised of the following:

The solution is defined with at least a 64-core 5th Generation Intel® Xeon® Scalable processor and high- performance network, with storage and integrated platform acceleration products.

The Bill of Materials (BOM) requirement details for the configurations are provided below in this document.

The VRB is defined in collaboration with enterprise vertical users, service providers, 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.

Table of Contents

Figures

Tables

2 5th Generation Intel® Xeon® Scalable Processors

This VRB provides numerous benefits to ensure end users have excellent performance for their AI Inference applications. Some of the key benefits of the VRB based on the 5th Generation Intel® Xeon® Scalable Processor Family processor include:

- 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
- Accelerated encryption and compression
- Platform-level security enhancements

Figure 1. Dell PowerEdge R760

3 Design Compliance Requirements

This chapter focuses on the design requirements for the Intel® AI Edge Systems VRB with Dell for Computer Vision, and GEN AI.

The checklists in this chapter are a guide for assessing the platform's conformance to the VRB. The hardware requirements are detailed below.

Ingredient	Requirement	Required/ Recommended	Quantity
Processor	Intel® Xeon® Platinum 8592+ Processor at 1.9GHz, 64C/128T, 350W or higher number SKU	Required	2
Memory	16x 32 GB DDR5, 5600 MT/s (512 GB total) or higher	Required	16 (8 per NUMA)
Storage (Boot Drive)	480 GB or equivalent boot drive	Required	1
Storage (Capacity)	1 TB or equivalent drive [recommended Non-Uniform Memory Access (NUMA) aligned]	Recommended	4 (2 per NUMA)
Network	Intel® Ethernet Network Adapter E810-2CQDA2	Required	1
LAN on Motherboard (LOM)	10 GbE or 25 GbE port for Pre-boot Execution Environment (PXE) and Operation, Administration and Management (OAM)	Required	2 (1 per NUMA)
	1/10 GbE port for Management Network Interface Controller (NIC)	Required	1

Table 1. HW BOM for Intel® AI Edge Systems VRB with Dell

To meet the performance requirements for an Intel® AI Edge Systems Verified Reference Blueprint with Dell, Intel® recommends using the BIOS settings for enabling processor P-states and C-states 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 considerations.

Chapter 3 of "BIOS Settings for Intel® Wireline, Cable, Wireless and Converged Access Platform" (Document ID #747130) documents the NFVI profile for BIOS settings.

Setting	Value
Hardware Prefetcher	Enabled
Intel® (VMX) Virtualization Technology	Enabled
Hyper-Threading	Enabled
Intel® Speed Shift Technology (P-States)	Enabled
Turbo Mode	Enabled
C-States	Enabled
Enhanced C-States	Enabled

Table 2. BIOS Settings for Intel® AI Edge Systems VRB with Dell

Additionally, for this specific solution, please make the corresponding additional BIOS changes on top of the NFVI BIOS Profile to yield optimal performance for the solution configurations.

Setting	Value
Sub NUMA Clustering	SNC2
AVX P1	Level 2
AVX ICCP Pre-grant License	Enabled
AVC ICCP Pre-Grant Level	512 Heavy
Memory Page Policy	Adaptive

Table 3. Additional Recommended BIOS Settings

Note: BIOS settings differ from vendor to vendor. Please contact your Intel® Representative for the NFVI BIOS Profile Document ID #747130 or if you have difficulty configuring the exact setting in your system BIOS.

Figure 2 shows the architecture diagram of the VRB. The software stack consists of two categories of AI software:

1. Generative AI
2. Vision AI

Both applications are containerized using docker.

The Generative AI use case leverages large language models (LLMs) using the Intel® Extension of PyTorch (IPEX) framework to perform LLM inference on Intel® Xeon® Scalable Processor-based CPUs.

The Vision AI use case leverages the Intel® Automated Self-Checkout application, which measures the maximum achievable stream density. 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 detects objects, and the EfficientNet model classifies Objects.

The table below is a guide for assessing the conformance to the VRB's software requirements. Ensure that the platform meets the requirements listed in the table below.

Ingredient	SW Version Details
OS	Ubuntu 22.04.4 LTS
Kernel	6.5 (in-tree generic)

Ingredient	SW Version Details
OpenVINO	2024.0.1
Docker Engine	27.1.0
Docker Compose	2.29
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

Table 4. SW BOM Intel® AI Edge Systems VRB with Dell

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

The VRB requires Intel® Virtualization Technology (VT) to be enabled to reap the benefits of hardware virtualization. Either Intel® Boot Guard or Intel® Trusted Execution Technology establishes the firmware verification, allowing for platform static root of trust.

Platform Technologies		Enable/ Disable	Required/ Recommended
Intel® VT	Intel® CPU Virtual Machine Extension (VMX) Support	Enable	Required
	Intel® I/O Virtualization	Enable	Required
Intel® AMX	Intel® Advanced Matrix Extensions	Enable	Required
Intel® Boot Guard	Intel® Boot Guard	Enable	Required
Intel® TXT	Intel® Trusted Execution Technology	Enable	Recommended

Table 5. Virtualization Technology and TXT Settings

For the VRB, it is recommended that Intel® Boot Guard Technology be enabled so that the platform firmware is verified as being suitable during the boot phase.

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

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 .

4 Performance Verification

This chapter aims to verify the VRB's performance metrics to ensure that no anomalies are seen. Refer to the information in this chapter to ensure that the platform's performance baseline is as expected.

The solution was tested on August 30, 2024, with the following hardware and software configurations.

Hardware	Configuration
CPU	2x Intel® Xeon® Platinum EMR 8592+ Processor
Sockets	2
Cores per Socket	64

Hardware	Configuration
LLC Cache	320MB Cache
TDP per CPU	350W
Simultaneous Multithreading (SMT)	Intel® Hyper-Threading Technology Enabled
CPUs	256
CPU Frequency	1.9 GHz base clock speed, 2.9 GHz all-core turbo frequency, 3.9 GHz max turbo frequency,
NUMA Nodes	2
Hyperthreading	Enable
Turbo	Enable
C-State	Enable
Total Memory	16x32GB 512GB, DDR5-5600 MT/s, 1DPC, 8 channels
Hard Drive/ Disk	2x 447.1G INTEL SSDSC2KB48
Network Interface Card/AIC	1x Dual port Intel® Ethernet Network Adapter E810-2CQDA2 2x Ethernet Connection X722 for 1
Network speed	1 GbE
Microcode	0x21000240
OS/Software	Ubuntu 22.04.4 (kernel 6.5.0-44-generic)

Table 6. HW BOM for the System Under Test

5 Generative AI

In the current technological landscape, Generative AI (GenAI) workloads and models have gained widespread attention and popularity. Large Language Models (LLMs) have emerged as the dominant models driving these GenAI applications. The generation task is memory bound due to iterative decode and KV Cache which needs special management to reduce memory overheads. All Gen AI models are downloaded from the Hugging Face repository.

Intel® Extension for PyTorch* provide a lot of specific optimizations for these LLMs with platform features optimizations for performance boost on Intel® hardware. The optimizations take advantage of Intel® Advanced Vector Extensions 512 (Intel® AVX-512) Vector Neural Network Instructions (VNNI) and Intel® Advanced Matrix Extensions (Intel® AMX) on Intel® CPU.

For benchmarking the models are targeted solely to the CPU. In both cases, we leverage the IPEX-LLM framework. For the CPU benchmarks, a larger memory footprint is available unlike in GPU where we are restricted with the GPU VRAM.

To better trade-off the performance and accuracy, different low-precision solutions like weight-only-quantization is also enabled. Additionally, tensor parallel and pipeline parallelism mechanism is also adopted for distributed inference to get lower latency for LLMs.

The Large Language Model (LLM) proxy workload highlights the Generative AI processing capabilities of the VRB specifically with the 8B to 40B parameter model is supported directly on 5th Gen Intel® Xeon® Scalable processors. Specifically, we have tested Llama3-8B, GPT-Neox20B and Falcon40B model with bfloat16, INT8 and INT4 precision.

The weight only quantization method was used for model quantization for converting model from bfloat16 to INT8 and INT4. For faster inference on dual socket CPUs with multiple NUMA regions, we have used auto tensor-parallelism (TP) using DeepSpeed optimization with Sub-NUMA Clustering (SNC) setting.

Ingredient	Software Version Details
Framework /Toolkit	CPU: PyTorch v2.3.100+cpu Deepspeed v0.14.0 Transformers v4.38.1 IPEX-LLM
Topology or ML Algorithm	tiiuae/falcon-40b EleutherAI/gpt-neox-20b meta-llama/Llama-3-8b-hf microsoft/Phi-3-mini-4k-instruct TinyLlama/TinyLlama-1.1B-Chat-v1.0
Libraries	oneDNN v3.4.1 oneCCL v2021.11 torch-ccl v2.3.0+cpu Intel® Neural Compressor v2.4.1
Model Precision	BF16, INT8, INT4

Ingredient	Software Version Details
Quantization methods	weight-only-quantization
Warmup steps	1
Number of Iterations	4
Batch Size	1, 2, 4, 8, 16, 32
Beam Width	1 (greedy search)
Input Token Size	32, 256, 1024, 2048
Output Token Size	1024
Compiler	GCC version 12.3.0
Python	3.10.12
OS	Ubuntu Desktop LTS, Kernel 6.5

Table 7. GenAI Benchmark Parameters

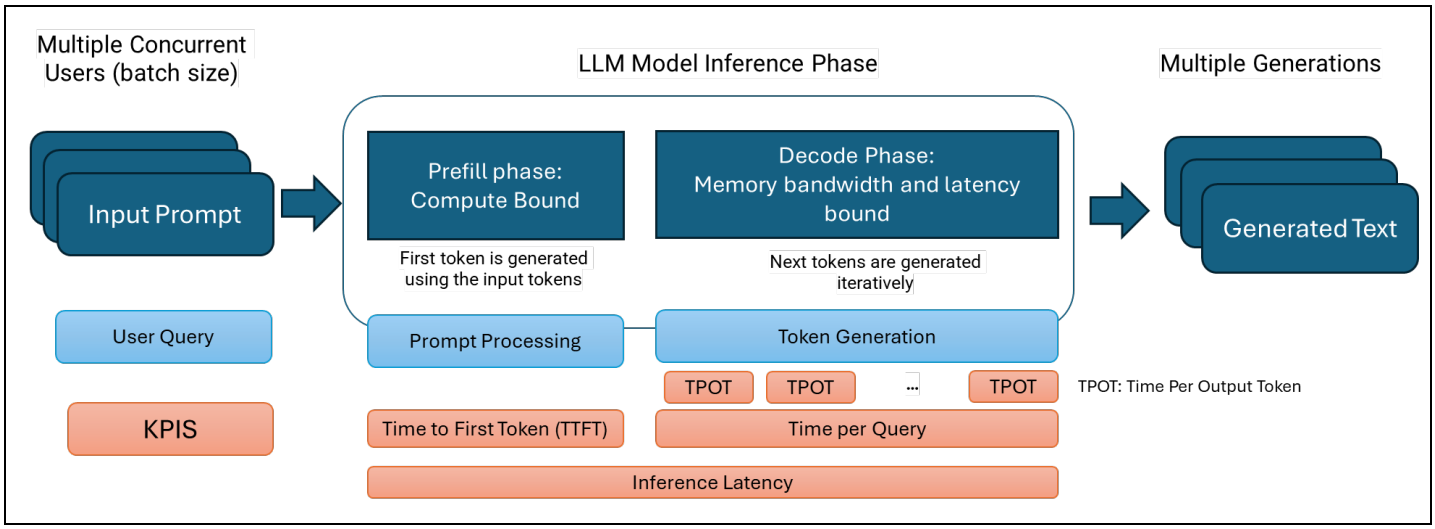


Figure 2. LLM Inference and KPIs

The Intel® AI Edge Systems Verified Reference Blueprint with Dell ensures that the system's results follow the expected results, as shown below, to baseline the platform's performance. The results shown include performance values for the next token latency, the achievable number of tokens per second, and the inference latency.

In terms of KPI measurements for Gen AI, we measure:

- First Token Latency: The time taken to process the input prompt and generate the first token.
- Average Next Token Latency: The time taken to generate each new token for the output sequence.
- Throughput: The average number of tokens generated every second.
- Inference Latency: The time taken from the initial prompt until the response text is generated.
- Time per Query: The time take from the first token until the last token.

6 Generative AI on Dual Socket 5th Gen Intel® Xeon® Scalable Processor

The Large Language Model (LLM) proxy workload highlights the Generative AI processing capabilities of the Verified Reference Blueprint platform, specifically with the 8B to 40B parameter model, which is supported directly on 5th Gen Intel® Xeon® Scalable processors. We have tested the Llama3-8B, GPT-Neox20B, and Falcon40 B models with bfloat16, INT8, and INT4 precision.

The weight-only quantization method is used for model quantization to convert the model from bfloat16 to INT8 and INT4. For faster inference on dual socket CPUs with multiple NUMA regions, we have used auto tensor-parallelism (TP) using DeepSpeed while enabling sub-NUMA clustering (SNC).

The figure shows the Cumulative next tokens per second for the 1024 input and 1024 output tokens with varying batch sizes and model precisions, with an inference time of less than 60 seconds.

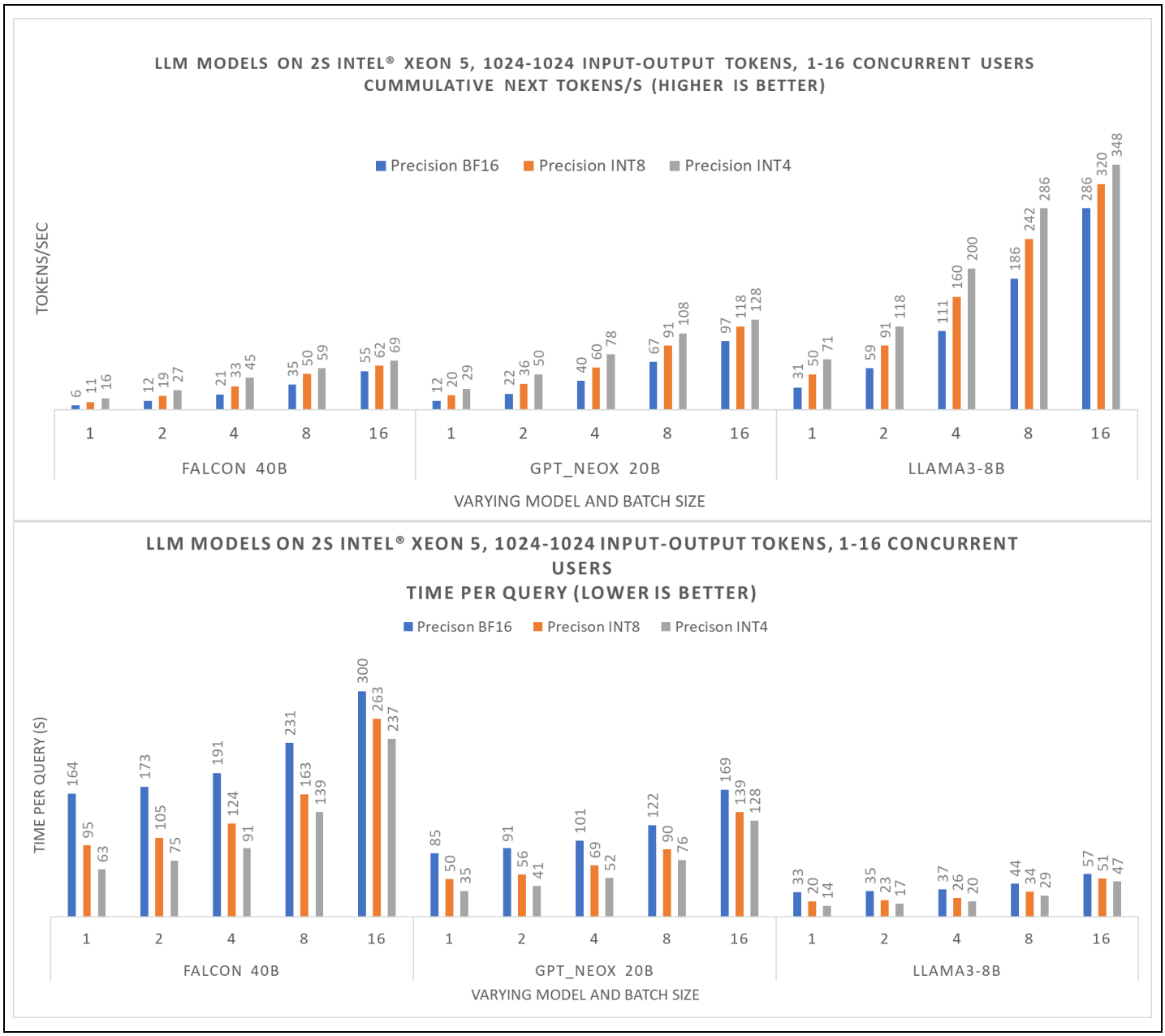


Figure 3. Performance for Multiple LLM Models with Variable Batch Sizes

7 Vision AI on Dual Socket 5th Gen Intel® Xeon® Scalable Processor

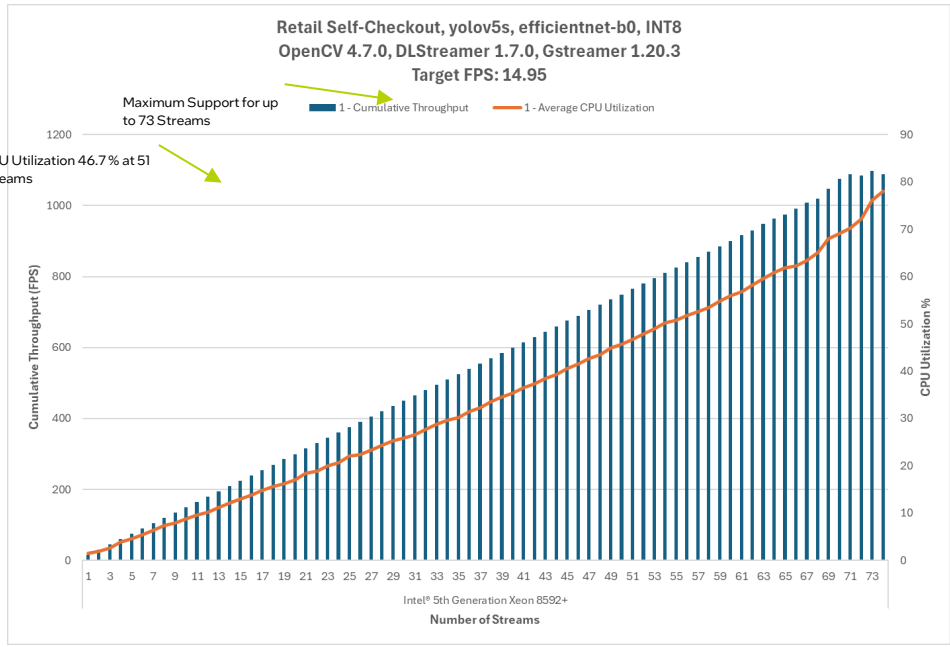
The Intel® Automated Self-Checkout Reference Package provides critical components required to build and deploy a self-checkout use case using Intel® hardware, software, and other open-source components. The Intel® Automated Self-Checkout serves as a proxy workload for Vision AI applications and leverages the YOLOv5 model for performing detection along with the efficientnet-b0 model for performing classification.

Ingredient	Software Version Details
OpenVINO™	2024.0.1
DLStreamer	2024.0.1
FFmpeg	2023.3.0
VPL	2023.4.0.0-799
Python	3.8+

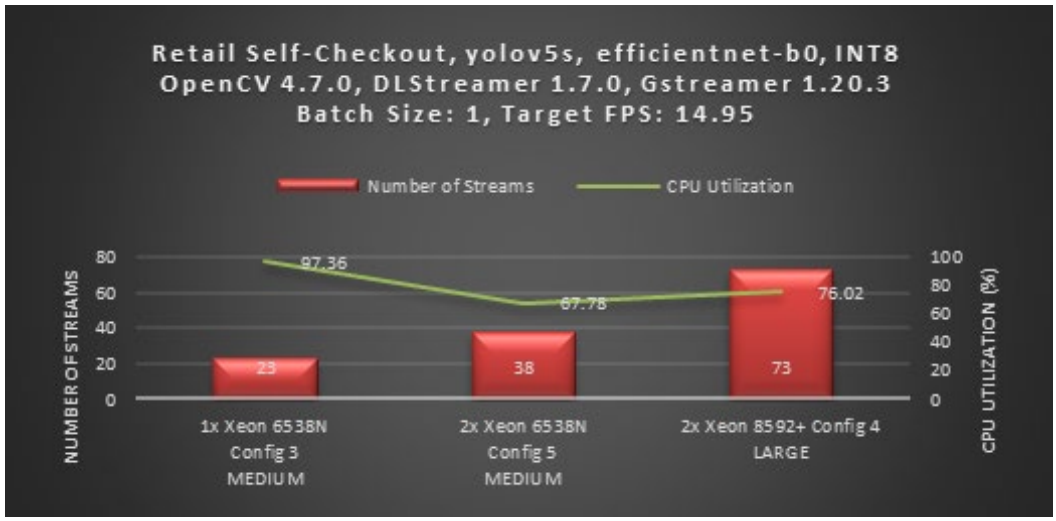
Ingredient	Software Version Details
OS	Ubuntu Desktop LTS, Kernel 6.5 (gcc 11.4.0)

Table 8. Vision AI SW Libraries

The VRB with 5th Generation Intel® Xeon® SP should be able to service up to 73 IP camera streams at 14.95 FPS per stream, for an aggregate of up to 1098 FPS.



The figure below summarizes the Vision AI results across multiple platform configurations.



8 Summary

The Intel® AI Edge Systems Verified Reference Blueprint with Dell defined on dual socket 5th Gen Intel® Xeon® Scalable processors address the capabilities for AI inference offering the following value propositions:

1. For the Generative AI use case

Model	Precision	Config	Batch Size	Tokens/s
Llama3 8B	INT8	2x 8592+	32	670
GPT-NEOX-20B	INT8	2x 8592+	2	39
Falcon 40B	INT4	2x 8592+	1	17

2. For the Vision AI use case:

Configuration	Number of IP Camera Streams
2x Intel® Xeon® 8592+	73

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

These processors are optimized for network, cloud-native, wireline, and wireless core-intensive workloads and are especially suited for AI workloads coupled with Intel® Ethernet E810-Network Controllers.

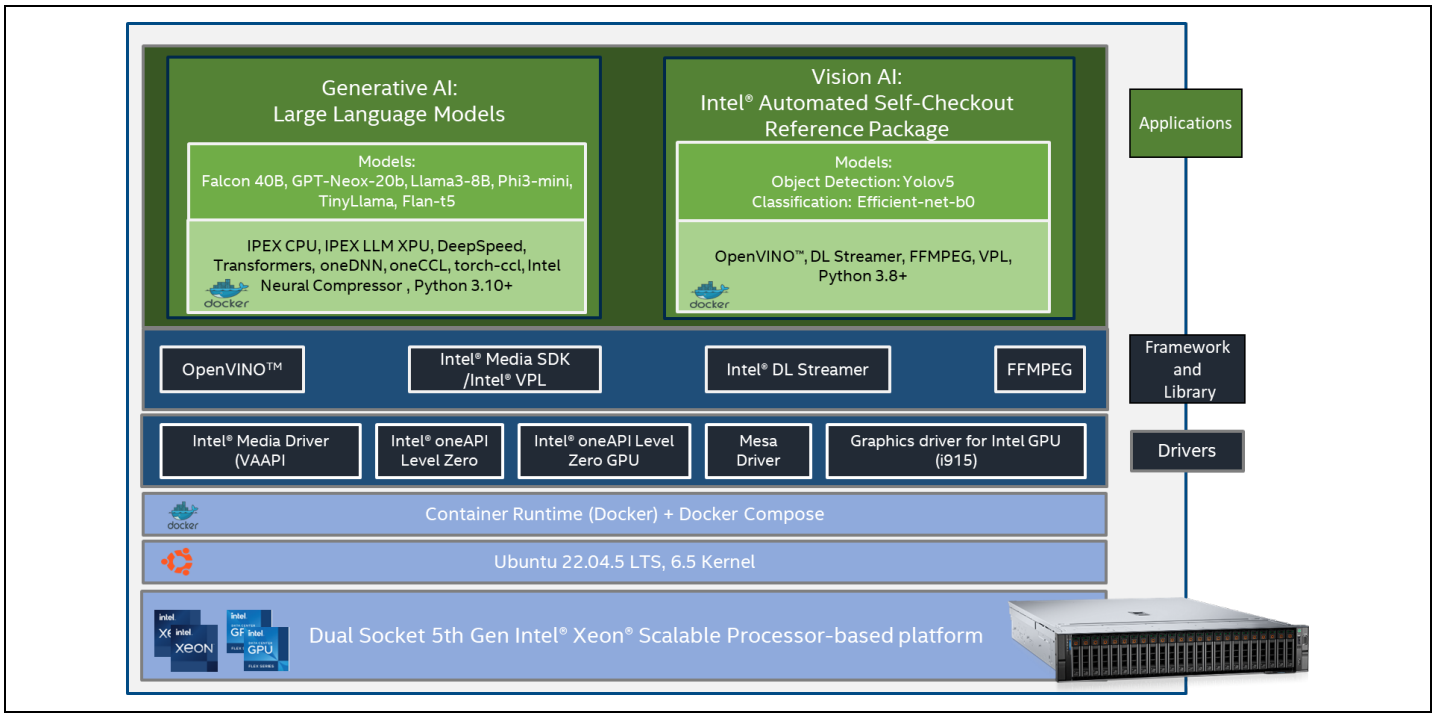


Figure 4. Architecture for the Intel® Verified Reference Blueprint for AI with Dell PowerEdge R760

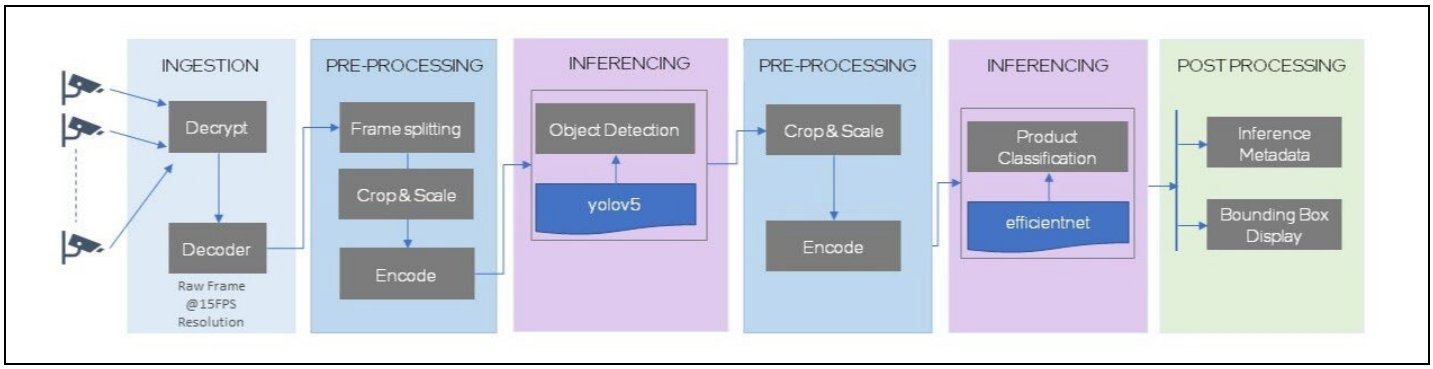


Figure 5. Vision AI Processing Pipeline

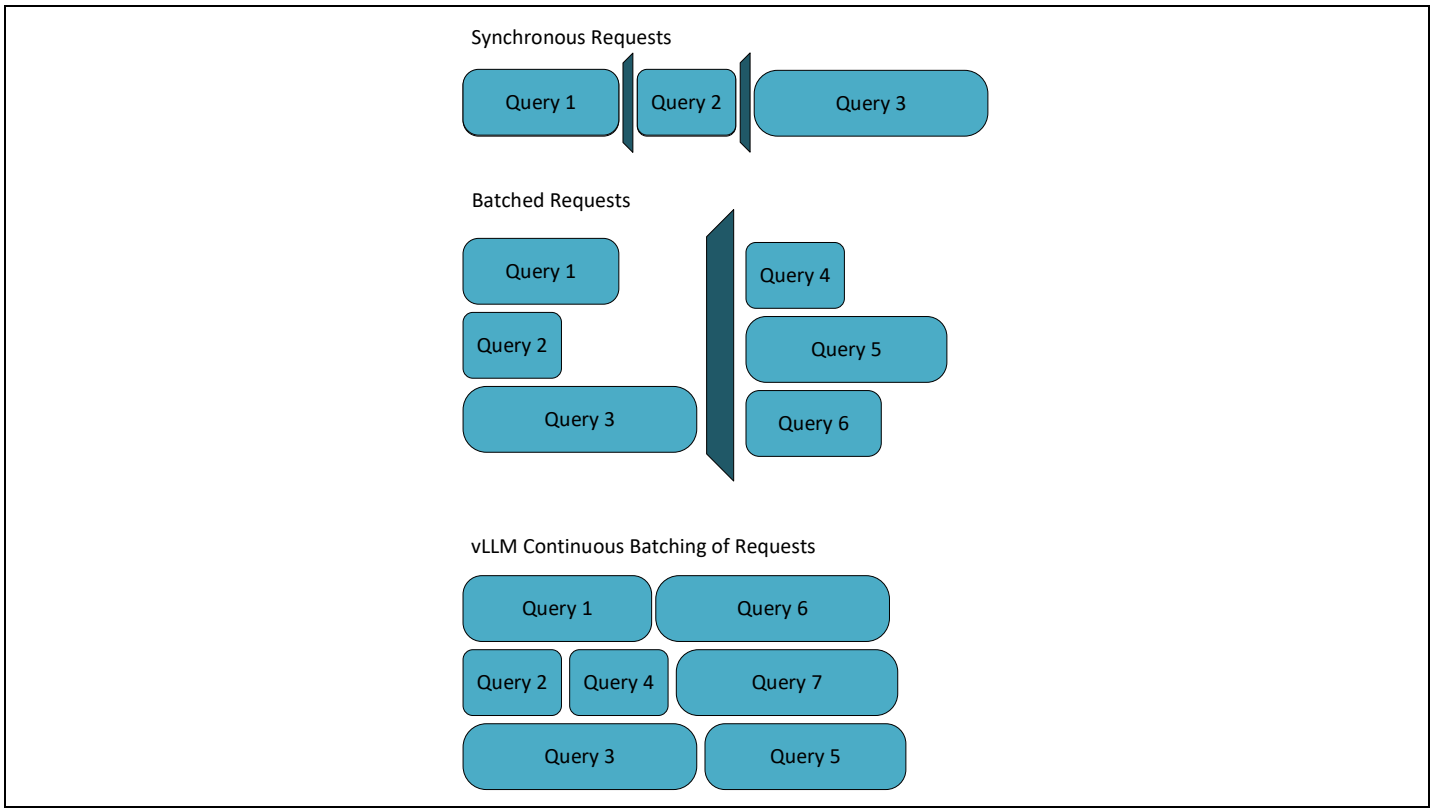


Figure 6. LLM Processing Stages

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 6