Case Study

Retailers, Warehouse Operators

Intel® CPUs are TCO-Efficient Hardware for Retail IoT AI Model

Gimlet Labs' self-checkout AI model, running on Intel® Architecture, targets high-volume retail IoT applications such as coolers and vending machines, where low TCO is important

Convenience stores and other retailers are increasingly adopting self-checkout systems to meet customer expectations for faster shopping experiences.

Self-checkout systems promise the dual benefit of reducing wait times and enhancing customer convenience, which is particularly appealing in fast-paced environments like convenience stores. Retailers are also seeing self-checkout as a way to address rising operational costs. Labor shortages and the cost of employing checkout staff have added to the financial pressure on retail businesses, making self-checkout an attractive alternative.

With the integration of artificial intelligence, machine learning, and computer vision, modern self-checkout systems are more adept at recognizing items, preventing theft, and facilitating various payment methods, from mobile payments to contactless options.

Internet of things (IoT) technology is being combined with AI models to incorporate certain devices into the self-checkout system, such as coolers, vending machines and other devices that are omnipresent at convenience stores, grocery stores, quick serve restaurants and other locations. Because these systems are enclosed, they need a dedicated computer vision (CV) system to track inventory and customer shopping. For some in the industry, there is the perception that the hardware required to power such systems is prohibitively expensive and power hungry.

Gimlet Labs, an Intel® Solution Builders Retail Builders Community member, was invited by Intel to work on a solution prototype for a large beverage manufacturer using optimized AI vision pipelines on low-cost Intel hardware. The beverage manufacturer originally developed a prototype with a discrete GPU, but after a trial, the feedback was that it was not cost effective and consumed too much electricity to be practical in a retail location.

Gimlet Brings AI to the Edge

Gimlet Labs is a developer platform for building, deploying, and monitoring AI pipelines to edge devices using generative AI to simplify the creation of custom models.

As seen in Figure 1, the Gimlet Platform enables the building, deployment, monitoring and improvement of AI workloads in minutes. Gimlet runs as a Docker container and can automatically execute and optimize AI pipelines to a target device, removing the need to manually port models across different platforms.

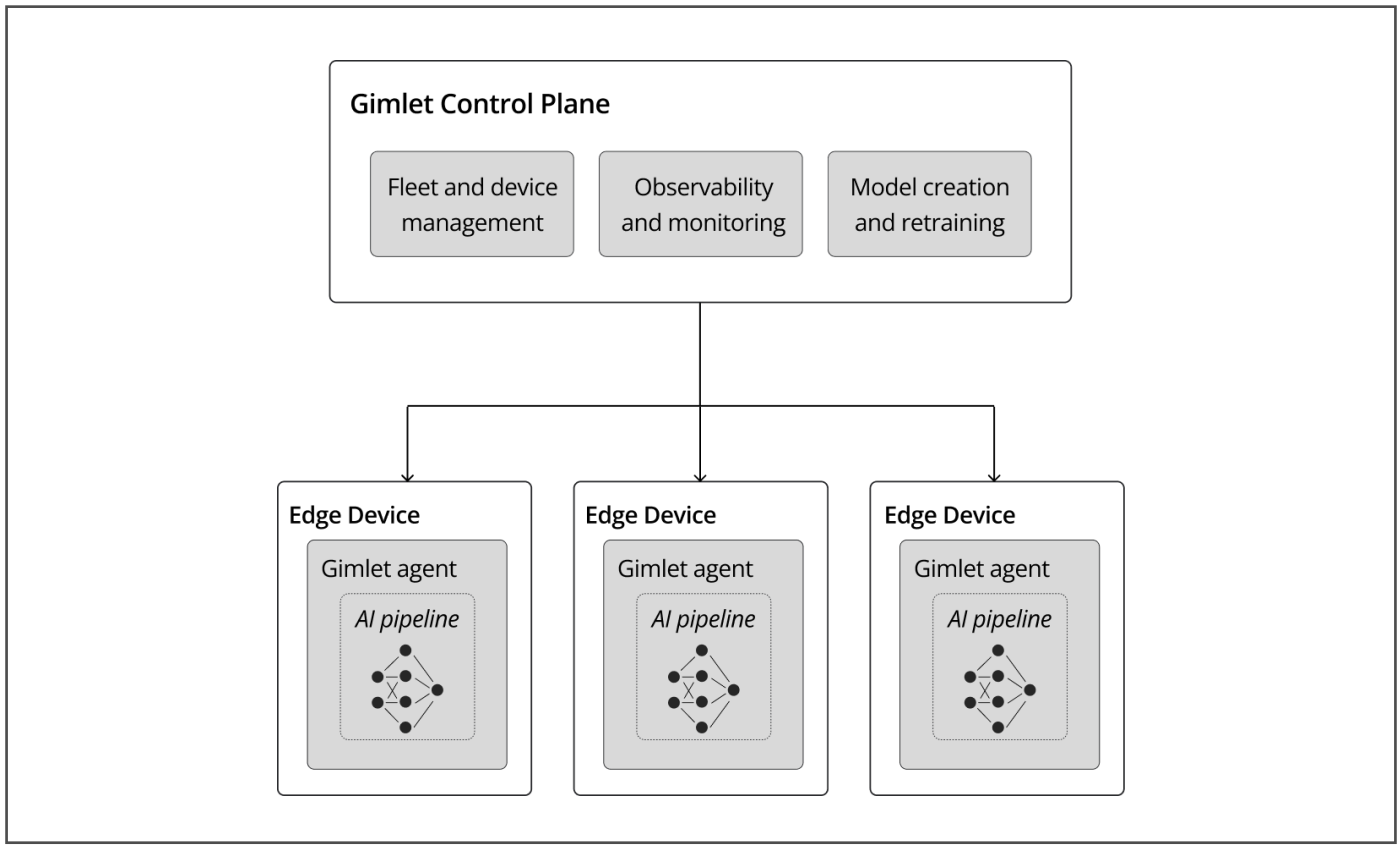

Figure 1. View of how the Gimlet platform is designed to build, deploy, monitor and improve AI workloads.

Key features of the Gimlet Platform include:

- Designed to bring full lifecycle model management to edge AI
- Generates task-specific models using generative AI prompts
- Can also use pre-built models
- Supports a wide range of hardware
- Deploys via a container-based runtime
- Accepts human labeling or automatic self-learning for model improvement

Figure 2 shows how Gimlet can create a task-specific AI model to track shopping carts using a one-sentence prompt.

Testing Real-Time Response

Gimlet conducted performance tests of its optimized AI model in the Gimlet lab using a standard beverage cooler[1]. A USB camera with a wide-angle lens was placed inside the cooler. The camera was located in the upper corner of the cooler, based on the specification of the beverage manufacturer. This placement provided a complete but angled view of the products inside.

The challenge for the AI model is to classify the exact beverage that is being taken by the customer when the camera does not have a direct view. Early versions of the prototype utilized two cameras to catch customers reaching for beverages - one in the top corner of the cooler and the other in the bottom corner. That configuration also was changed because a single-camera system can meet the accuracy objective and is a simpler design.

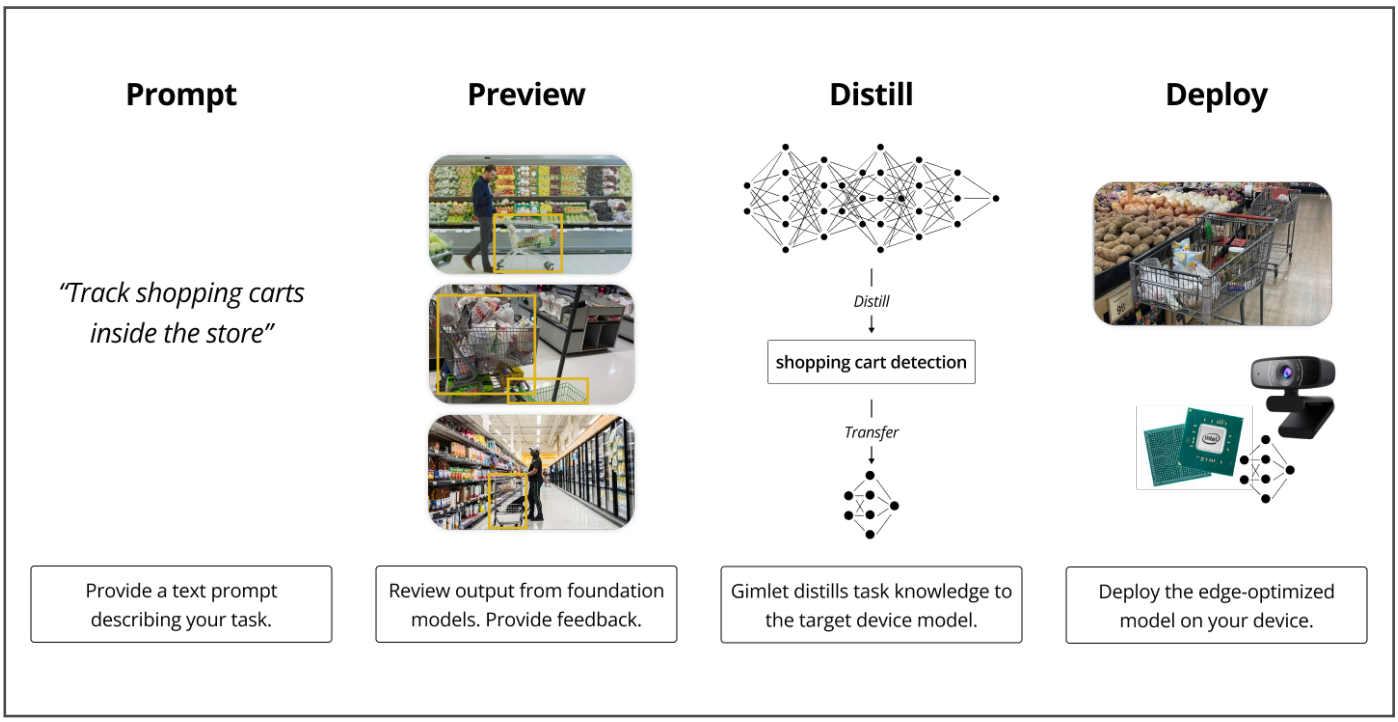

Figure 2. Process for creating a task using the Gimlet Platform.

Gimlet developed the model (see Figure 3) used in the test using a base architecture called YOLO, which is a well-known AI object detector model that the Gimlet team was able to train to be task specific. In addition to YOLO, the Gimlet team added an embedding model that could distinguish between various can products. This multistage pipeline is needed to detect the can, determine where it is in the frame of the camera, recognize the product, and determine whether it was taken, put back, or relocated.

This pipeline needs to be run differently for each CPU. The Gimlet AI agent is able to optimize the execution of that pipeline based on what CPU or GPU is being used. This allows the agent to exploit any accelerators on the devices, such as integrated GPUs that are built into all of the Intel® processors used in this test.

Making the Solution Cost Effective

To test the performance of the prototype model and meet the cost-effectiveness mandate from the customer, the Gimlet team decided to test both newer CPUs that are found in multifunction edge servers and older devices that might already be installed in a store and thus could run the model on unused cores.

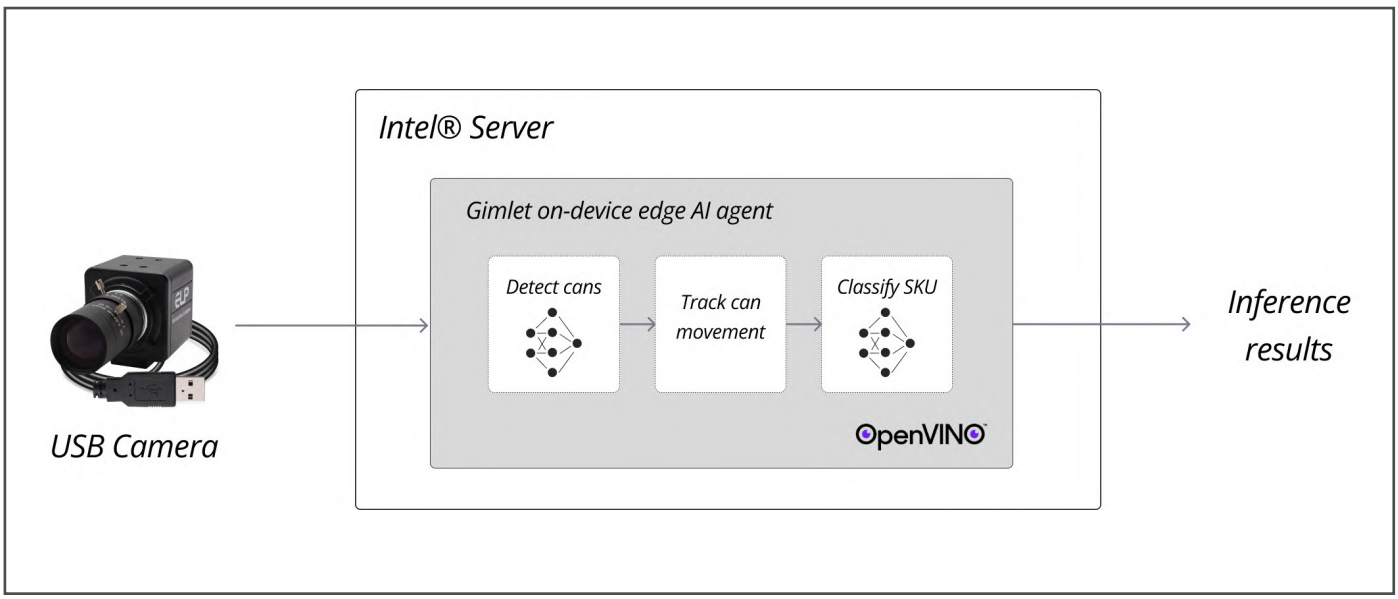

Figure 3. Process for creating a task using the Gimlet Platform.

The Intel® CPUs tested include:

CPU	Launch Date	Features
11th Generation Intel® Core™ i3-1115G4 processor (Configurable base frequency is 1.7 GHz (TDP-down at 12W) or 3.00 GHz (TDP-up at 28 W). Max Turbo frequency is 4.10 GHz.)	Q3 2020	Two core, four thread CPU with Intel® UHD Graphics for 11th Generation Intel® Processors with 48 execution units. Device has a 28W TDP and supports DDR4 memory.
Intel® Celeron® Processor N5105 (Base frequency is 2.00 GHz with burst frequency of 2.9 GHz.)	Q1 2021	Four core, four thread CPU with Intel® UHD Graphics with 24 execution units. Device has a 10W TDP and supports DDR4 memory.
Intel® Core™ Ultra 7 Processor 155H (Performance-core Base Frequency 1.4 GHz and Performance-core Max Turbo Frequency 4.8 GHz.)	Q4 2023	Six Performance-cores, eight Efficient-cores, two low-power Efficient cores for a total of 22 threads; integrated eight-core Intel® Arc™ graphics. Device features a 20W minimum assured TDP and 115W maximum TDP supporting DDR5 memory.

OpenVINO™ Toolkit Accelerates AI Inferencing

An important part of the test set up was the use of OpenVINO Toolkit open source software that accelerates AI inference development. OpenVINO Toolkit also delivers lower latency and higher throughput while maintaining accuracy, reducing model footprint, and optimizing hardware use. In this use case, OpenVINO Toolkit was used on all three Intel CPUs.

The Gimlet Platform leverages OpenVINO Toolkit when targeting Intel architecture devices, and was used on all three CPUs in this case.

Tests Measure Frames Per Second

In the tests, the Gimlet model was run on each Intel CPU with increasing amounts of video sent through the model on each CPU until it hit its maximum frames per second (FPS). This testing showed whether the CPUs could deliver real-time model inference performance.

For a computer vision application, performance of 15-20 FPS is typically considered real time. These represent the performance minimums, and any performance higher than that improves the ability to catch fast moving objects and still correctly classify them.

As seen in Figure 4, all three Intel CPUs demonstrated real-time performance, with the Intel® Core™ Ultra 7 Processor 155H delivering more than three times the 20 FPS industry benchmark for real world performance.

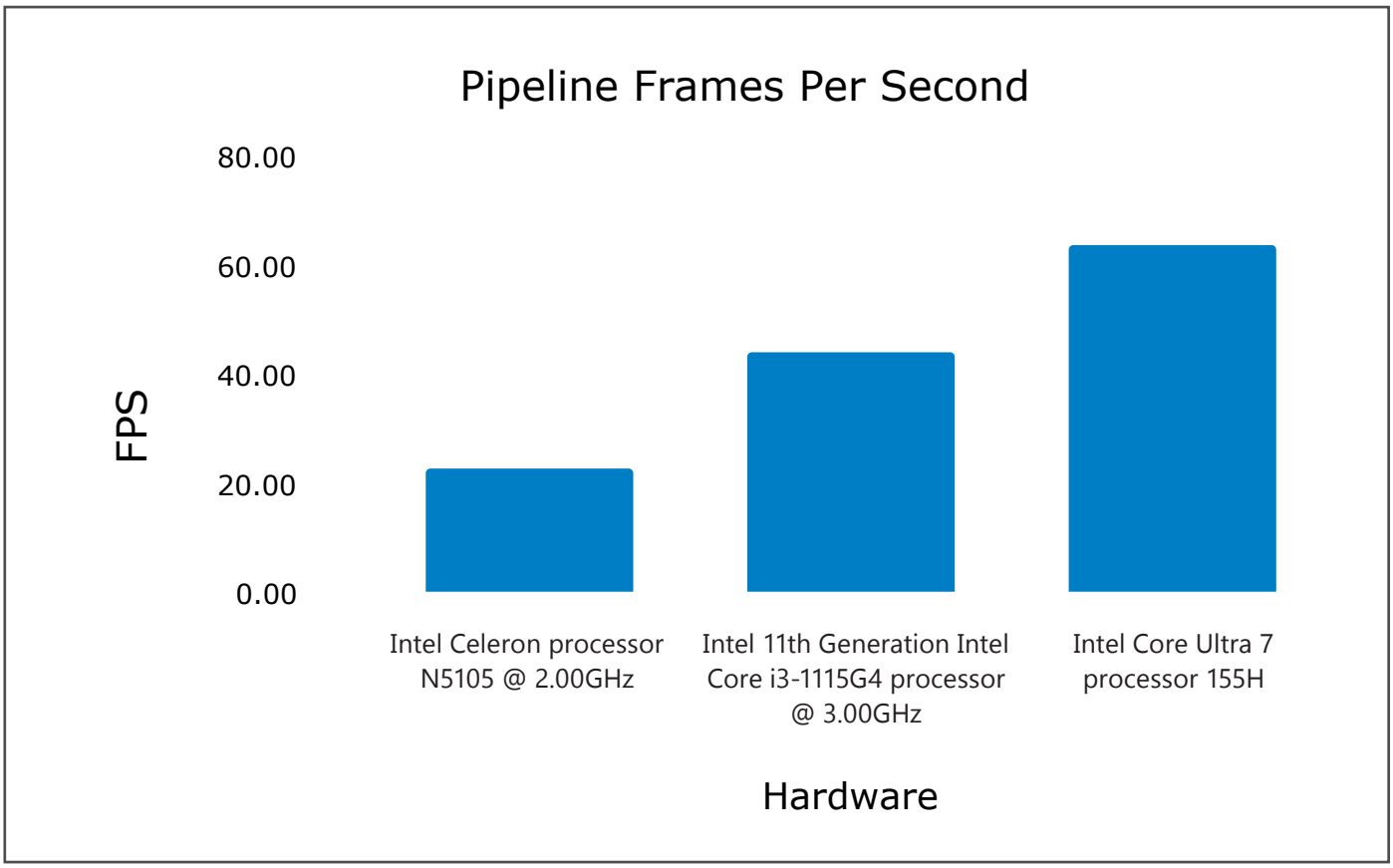

Figure 4. Performance result in frames per second for three Intel CPUs (higher is better).

Hardware	FPS Average	FPS Range
Intel Celeron Processor N5105 @ 2.00GHz	22.83	11.7 - 33.8
11th Generation Intel Core i3-1115G4 Processor @ 3.00GHz	44.02	27.3 - 57.9
Intel Core Ultra 7 Processor 155H	63.8	18.1 - 107.6

Table 1. Average and peak FPS performance for three Intel CPUs.

Other metrics captured during the tests can be seen in Table 1, including the maximum FPS rate and the average FPS rate. The average FPS rate is useful when trying to understand default behavior for the model, and the range is useful for understanding the performance under different workload conditions.

Conclusion

Increased use of customer self-checkout is a growing strategy for retailers to cope with employee and customer dynamics, and finding a solution for commercial beverage coolers is an important part of the overall system. But these systems need to be cost effective. Low power consumption is also an important part of the solution. Gimlet Labs developed a prototype multi-stage AI model that delivers near real time performance on new and old Intel CPUs. All of these CPUs, which featured integrated GPU functionality, delivered near real-time performance with TDPs as low as 10W. The tests show that with the right model, a cost-effective and high-performance solution can be created.

Learn More

Gimlet Labs

Demo of pipeline on Intel® Celeron® Processor N5105

Intel® OpenVINO™ Toolkit

Intel® Industry Solution Builders

Intel® Core™ Processors

Intel® Core™ Ultra Processors

Intel® Arc™ GPUs

intel®

[i] Intel® Celeron® Processor N5105 SUT: 1-node, 1x Intel Celeron N5105 processor with 4 cores and 4 threads. Total DDR4 memory was 8 GB (1 slots/ 8 GB/ 2933 MHz); microcode 0x24000026; Intel® Hyper-Threading Technology – not enabled; Intel® Turbo Boost Technology – not enabled. BIOS version: ATJSLCPX.0037.2022.0715.1547. Software: OS was Ubuntu 20.04.03 LTS; kernel was Linux 5.15.0119--generic. Benchmark/workload software: Custom AI vision pipeline for can recognition. Libraries: OpenVINO™ Toolkit 2024.0.0, Python 3.11.9, PyTorch 2.3.1. Test conducted by Gimlet Labs on Aug. 29, 2024.

11th Generation Intel® Core™ i3-1115G4 Processor SUT: 1-node, 1x 11th Generation Intel Core i3-1115G4 processor with 2 cores and 4 threads. Total DDR4 memory was 4 GB (1 slot/ 4 GB/ 3200 MHz); microcode 0xb8; Intel® Hyper-Threading Technology – enabled; Intel® Turbo Boost Technology – enabled. BIOS version: TNTGL357.0064.2022.0217.1550. Software: OS was Ubuntu 20.04.03 LTS; kernel was Linux 5.15.0119--generic. Benchmark/workload software: Custom AI vision pipeline for can recognition. Libraries: OpenVINO™ Toolkit 2024.0.0, Python 3.11.9, PyTorch 2.3.1. Test conducted by Gimlet Labs on Aug. 29, 2024.

Intel® Core™ Ultra Processor 7 155H SUT: 1-node, 1x Intel Core Ultra processor 7 155H with 16 cores and 22 threads. Total DDR4 memory was 32 GB (2 slots/ 16 GB/ 5600 MHz); microcode 0x1f; Intel® Hyper-Threading Technology – enabled; Intel® Turbo Boost Technology – enabled. BIOS version: M791A010A-XB. Software: OS was Ubuntu 20.04.03 LTS; kernel was Linux 6.8.0-47-generic. Benchmark/workload software: Custom AI vision pipeline for can recognition. Libraries: OpenVINO™ Toolkit 2024.0.0, Python 3.11.9, PyTorch 2.3.1. Test conducted by Gimlet Labs on Aug. 29, 2024.

Notices & Disclaimers

Performance varies by use, configuration and other factors.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for details. No product or component can be absolutely secure.

Intel optimizations, for Intel compilers or other products, may not optimize to the same degree for non-Intel products.

Your costs and results may vary.

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

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

See our complete legal Notices and Disclaimers.

Intel is committed to respecting human rights and avoiding causing or contributing to adverse impacts on human rights. See Intel's Global Human Rights Principles. Intel's products and software are intended only to be used in applications that do not cause or contribute to adverse impacts on human rights.

Intel® Turbo Boost Technology requires a PC with a processor with Intel Turbo Boost Technology capability. Intel Turbo Boost Technology performance varies depending on hardware, software and overall system configuration. Check with your PC manufacturer on whether your system delivers Intel Turbo Boost Technology. For more information, see https://www.intel.com/content/www/us/en/gaming/resources/turbo-boost.html?wapkw=turbo%20boost

0125/CG/DJA/PDF Please Recycle 364011-001US