Solution Brief

Health and Life Sciences Solution Architects / R&D Leads, Software Developers

congatec Adds AI to Ultrasound Systems with Intel®-Based Modules

conga-TC750 is based on COM Express standard and leverages heterogeneous computing capabilities of Intel® Core™ Ultra Processors (Series 2) to deliver a powerful, scalable, and field-upgradable module

The ultrasound market is witnessing a rapid integration of AI technologies, driven by the increasing need for faster, more accurate, and accessible diagnostic tools. Ultrasound systems with AI can enhance image quality by reducing noise and artifacts, leading to clearer and more detailed visualizations, which are crucial for early disease detection and accurate diagnosis.

Other benefits of AI in ultrasound systems include workflow enhancements that reduce scanning time. Clinicians can also transport mobile, point-of-care ultrasound systems directly to the patient – whether that is in a clinic or in their home.

Adding AI to Ultrasound Machines

Building AI into an ultrasound product design means adding specialized AI accelerator chips, such as graphics processing units (GPUs), which consume significant amounts of power and generate heat. Manufacturers have three approaches to building these capabilities into high-volume products (see Figure 1):

- Chips down (customize an off-the-shelf single board computer)
- Use specialized Computer-On-Module Express (COM-E) form factors
- Create a custom board just for that product

The COM-E option is faster and more flexible because it offers a complete AI system that is ready to be designed into the new ultrasound product. Adding AI this way delivers a more future-proof solution with pin-compatible interchangeability and upgradeability.

Congatec is an Intel® Industry Solutions Builders member and has developed a product that satisfies the market need for AI in ultrasound machines and Xray machines with its conga-TC750, a family of five COM Express Compact modules based on the Intel® Core™ Ultra Processor (Series 2) product family.

conga-TC750 delivers AI performance

The conga-TC750 is congatec's latest COM-E for AI-enabled ultrasound systems. The module delivers AI performance up to 99 trillion operations per second (TOPS)[1]. The 95mm x 95mm module supports up to 128 GB RAM with in-band error correction code (ECC) which maintains memory integrity when the system is in use near radiation. The module supports PCI Express Gen 4 for fast input / output and USB Gen 4.

The Computer-on-Module is used everywhere and meets strict industrial standards for long-term, revision-controlled availability.

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

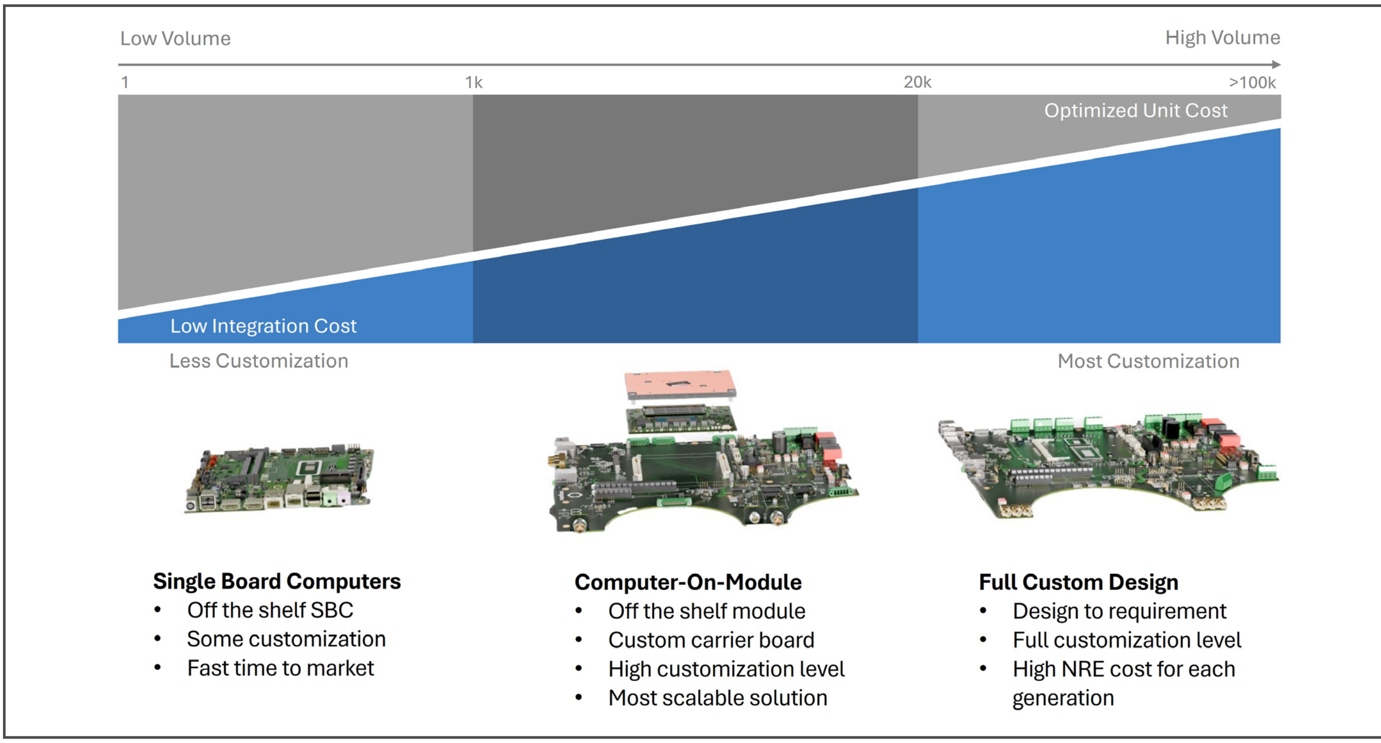

Figure 1. Comparing options for adding specialized AI capabilities to an ultrasound machine.

Medical carrier boards isolate interfaces galvanically and use touch-safe sockets to guard against leakage currents.

Industrial carrier boards favor rugged, lockable connectors rated for wide temperature, shock, and vibration.

aReady Software Stack

The conga-TC750 is available with congatec's application ready (aReady) software stack that simplifies the integration of AI, internet of things (IoT) connectivity, and security functionality. The stack includes high-performance embedded software building blocks for faster customer application readiness.

The aReady software stack includes the following layers:

- Virtualization layer: The built-in virtualization engine is the RTS Real-Time Hypervisor, which is combined with congatec's own Hypervisor-on-Module firmware that consolidates multiple applications on a single physical module for full resource utilization.
- Operating system layer: the conga-TC750 is available with either ctrlX OS that is specialized for industrial applications, or Ubuntu Pro, a Linux OS designed with security features. In addition to these Linux-based systems, Microsoft Windows remains available as an operating system option for specific ultrasound applications, depending on customer requirements.

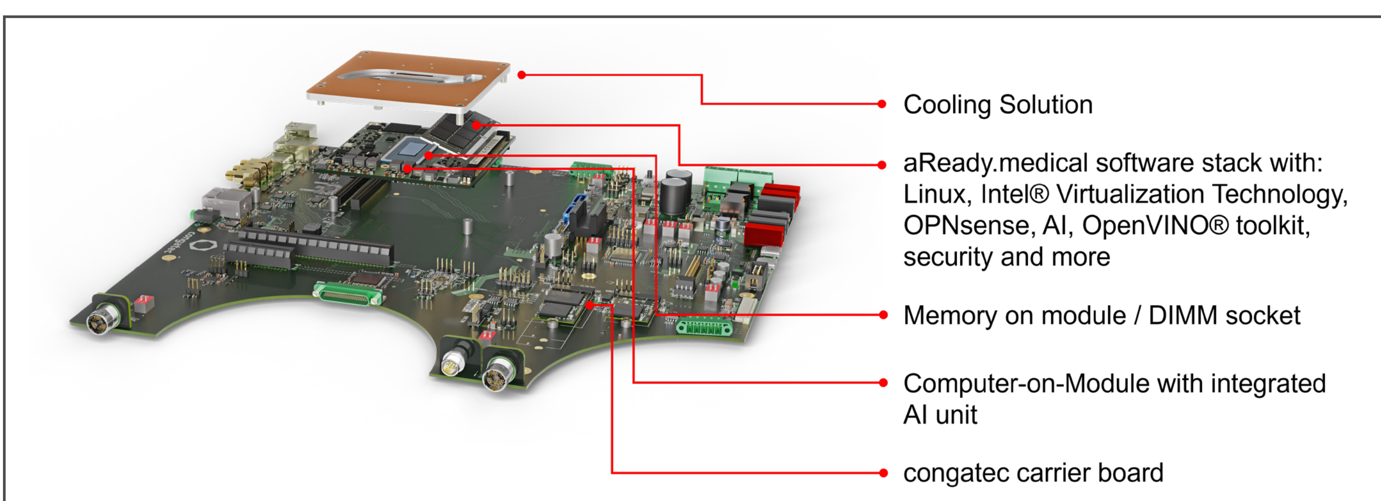

Figure 2. Exploded view of conga-TC750 COM-E module for ultrasound systems.

- Software layer: This layer is composed of a customizable number of functional software building blocks that support customer applications. This layer includes software such as Intel® Distribution of OpenVINO™ Toolkit for AI inferencing. For security, congatec has integrated OPNsense open source firewall and routing software for cyber security and networking. Other software at this layer supports IoT, human machine interface and other applications.

On top of all of these software layers is the customer's application written with this stack in mind.

Modules are Easily Upgradable

The predecessor to the conga-TC750 is the conga-TC700 based on Intel® Core™ Ultra Processors (Series 1). Since this board is also based on the COM-E standard and is pin-compatible with the conga-TC750, product designers can swap out their conga-TC700 and upgrade to the performance and features of the conga-TC750 based on Intel® Core™ Ultra Processors (Series 2).

Intel AI Accelerators Power conga-TC750

The conga-TC750 module gets its performance from the heterogeneous compute power built into the Intel Core Ultra Processors (Series 2) product family, which includes:

- CPU that features up to 16 cores including performance and efficiency cores
- Built-in Intel® Arc Graphics that are powerful enough to replace entry-level discrete GPUs
- Integrated neural processing unit (NPU) for AI inferencing

CPU Delivers Performance and Efficiency

The CPU features up to eight Performance-cores (P-cores) optimized for single-threaded and burst workloads, up to eight Efficiency-cores (E-cores) for multi-threaded tasks, and two Low Power E-cores (LP E-cores) for background processes and sustained efficiency. This approach dynamically allocates resources based on workload demands, optimizing the performance per watt.

Graphics Processor for Compute Intensive Tasks

A key advancement within the Intel Core Ultra processors is the integration of Intel® Arc™ Graphics. This integrated GPU provides substantial GPU performance. In an AI-enabled ultrasound, the GPU is used for a range of compute-intensive tasks that include:

- **Real-time image display**: Converting raw data from piezoelectric transducers into 2D or 3D rendered ultrasound images
- **Image enhancement**: Filtering, smoothing, and contrast adjustment in real time as well as processing color overlays and depth information
- **Rendering and visualization**: Generating 3D representations of the examined tissue and providing interactive visualization for the user (e.g., zooming, rotating the display, etc.)
- **Parallelization of tasks**: Handling large amounts of data coming from ultrasound transducers and preparing it for further processing and displaying on the monitor

By incorporating a high-performance GPU directly onto the processor die, Intel Core Ultra processors minimize power consumption and platform footprint, which is ideal for power-constrained devices like battery-operated ultrasound systems.

Integrated NPU Delivers Sustained Processing

In addition, the Intel Core Ultra processor family has an integrated neural processing unit (NPU) enabling on-die hardware acceleration for AI workloads. For the conga-TC750 the NPU provides advanced functions and automated decision-making including:

- **Automatic pattern recognition**: For detecting and classifying tissue structures (e.g., tumors, lesions, or anomalies) and automatically segmenting organs or blood vessels
- **AI-assisted image optimization**: Enhances image quality using neural networks and providing AI feedback to help optimally position the ultrasound probe
- **Real-time inferences**: Analyzing flow patterns in Doppler mode and predicting diagnostic indicators (e.g., tissue elasticity, blood flow velocity)
- **Assistance systems**: Assisting physicians in diagnosis by highlighting abnormal areas (e.g., detecting a suspicious structure and automatically highlighting it in real time)

Minimize Power Consumption

Reducing power consumption in ultrasound machines is crucial for extending battery life in mobile systems that move from exam room to exam room. The Intel Core Ultra processor's heterogeneous compute architecture consumes less power than discrete GPUs and is designed with optimized low thermal design power (TDP) characteristics. This efficiency is achieved through advanced process technology and intelligent power management strategies. The Intel Core Ultra processor family can also dynamically adjust power consumption based on workload demands, which maximizes performance while minimizing thermal output, leading to cooler and quieter operation.

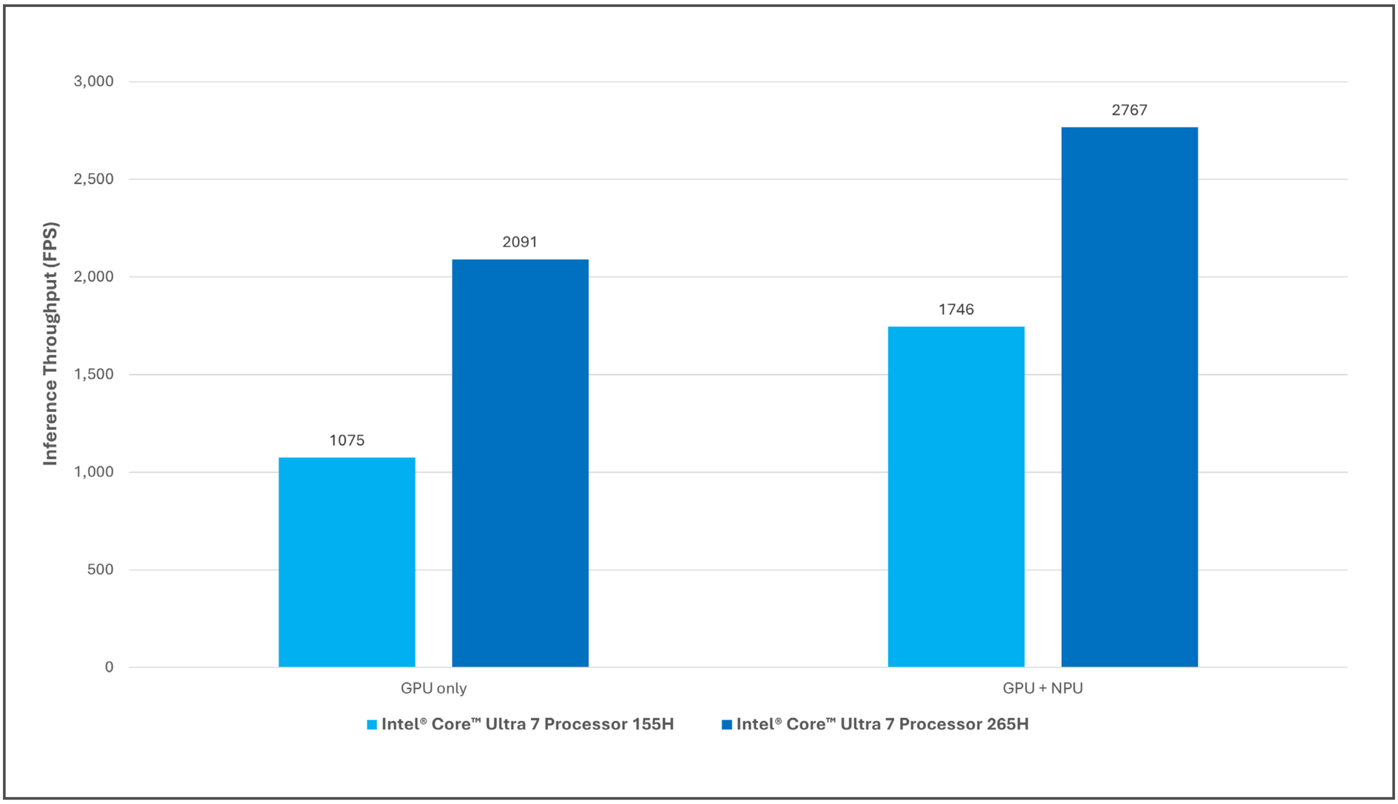

Figure 3. ResNet-50 inference throughput performance on Intel Core Ultra7 Processor 155H and Intel Core Ultra 7 Processor 265H.

Outstanding AI Inferencing Performance

Intel Core Ultra processors (Series 2) have built-in Intel® Arc Graphics as well as built-in NPUs, making it possible to deploy real-time AI applications directly on ultrasound machines. The result is not only lower bill of materials (BOM) costs, but also high inference throughput, assisting with faster diagnoses.

To evaluate inferencing performance, tests were conducted using ResNet-50, an image classification model that serves as a reliable benchmark for AI inferencing workloads in medical imaging. For these tests, batch size 8 was used, highlighting the performance boost achieved for applications requiring higher inference throughput when run on either the GPU alone or on both the GPU and NPU in Intel® Core™ Ultra Processors (Series 1) and Intel® Core™ Ultra Processors (Series 2) (see Figure 3).

In these tests, the systems based on the latest Intel® Core™ Ultra 7 Processor 265H deliver a 95% increase in inference throughput compared to systems using the Intel® Core™ Ultra 7 Processor 155H when running ResNet-50 (224 x 224) solely on the built-in GPU. This performance boost is due to the addition of 4-deep systolic support in the new generation that enhances matrix multiplication capabilities.

Customers can achieve even greater AI inference throughput by utilizing both the GPU and NPU for neural network deployment. As shown in Figure 3, this strategy can result in a 2.57x performance increase on Intel Core Ultra 7 Processor 265H when running ResNet-50 on both the GPU and NPU, compared to using only the GPU in the previous generation[i].

This improvement is significant because it offers customers nearly double the AI inference throughput performance on conga-TC750 compared to the previous conga-TC700 for image classification tasks like ResNet-50 when applications are deployed on the built-in GPU. Customers seeking enhanced performance are encouraged to upgrade by swapping out modules, which also leverages the additional processing capacity provided by the built-in NPU. This upgrade allows customers to run low-power AI inference applications on the NPU, offering flexibility in deploying AI applications.

Intel® Software Ecosystem

Developers can use the Intel® OpenVINO™ toolkit to build, optimize and deploy deep learning inferencing applications across Intel® processors and other hardware platforms (Figure 4). This toolkit is compatible with popular model frameworks like PyTorch, TensorFlow, and ONNX, enabling developers to seamlessly boost inference performance of AI models for different use cases like ultrasound and Xray systems.

With OpenVINO toolkit's unique "write once, deploy anywhere" approach, developers can build applications that can run on various Intel® compute engines, such as CPUs, GPUs, and NPUs, this versatility allows developers to optimize their solutions for different hardware, maximizing the performance of their AI workloads.

As a result, OpenVINO™, an open source toolkit, helps simplify the deployment of AI applications on congatec's aReady Software Stack to conga-TC700 and conga-750 modules. This reduces development costs and speeds up the time to market.

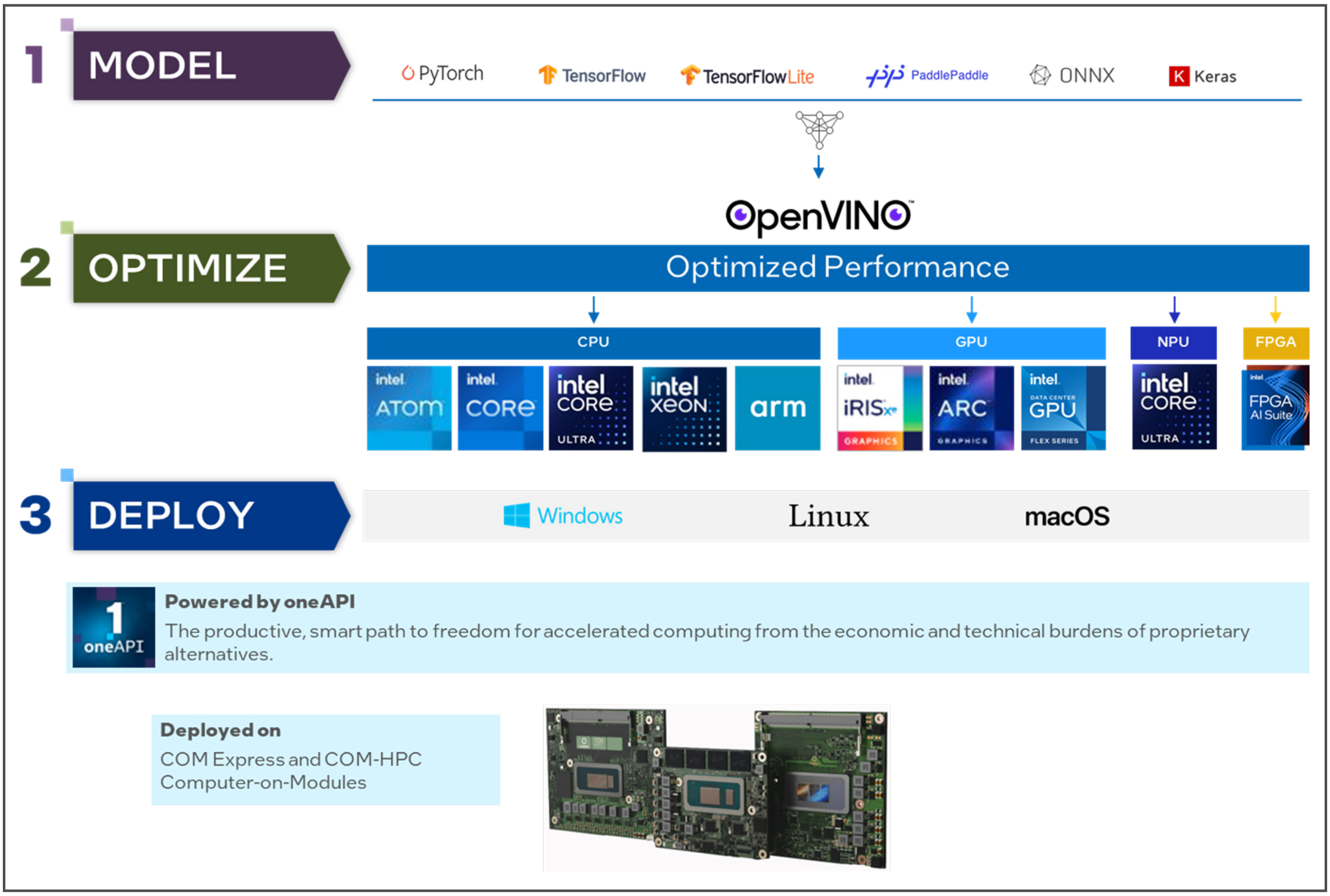

Figure 4. Intel and congatec AI application development pipeline for developers.

AI Applications are Easy to Deploy on the conga-TC750

The code snippet in Figure 5 demonstrates how easy it is to compile a model and run inference on the desired device using the OpenVINO™ Runtime API. In step #2, developers can effortlessly select the desired inference device on the Intel Core Ultra processor (GPU or NPU) by using the device_name parameter when loading and compiling the model with compile_model(). This enables developers to use the same code for running inference across the different inference devices available.

To begin optimizing and deploying AI applications with the OpenVINO toolkit, developers can explore numerous demos available in [OpenVINO™ Notebooks](). These demos span several industries, including retail, smart city, manufacturing, and healthcare.

```
#1 Create OpenVINO Core object instance
core = ov.Core()

#2 Compile model for the target inference device
# GPU
compiled_model = core.compile_model("ov_model.xml", device_name="GPU")
#To use NPU, replace GPU with NPU in the above command.

# Get the input and output layers of the optimized model
input_layer_ir = compiled_model.input(0)
output_layer_ir = compiled_model.output(0)

#3 Read the input image
image = cv2.imread(str(image_filename))
# Process input image
# N,C,H,W = batch size, number of channels, height, width.
N, C, H, W = input_layer_ir.shape
# Resize the image to meet network expected input sizes.
resized_image = cv2.resize(image, (W, H))
# Reshape to the network input shape.
input_image = np.expand_dims(resized_image.transpose(2, 0, 1), 0)

#4 Run inference
result = compiled_model([input_tensor])[output_layer_ir]
```

Figure 5. Ease of model deployment on GPU or NPU.

Conclusion

Ultrasound and Xray machines that use AI inferencing are meeting rising demand for faster, more accurate, and more accessible diagnostic systems. The conga-TC750 is providing the compute power to enable AI capability using cost-effective and high-performance COM-E modules that are easy to design into a system and are pin-compatible with previous models for easy in-field upgrades. Congatec's choice of the Intel Core Ultra processor family for its COM-E modules provides integrated CPU, GPU and NPU for high compute performance in a small board footprint and with low power consumption. As the tests in this paper show, using the Intel Core Ultra 7 265H Processor provides twice the performance over the previous model. And use of the devices enables congatec to benefit from the Intel software ecosystem including commonly used training tools like ONNX and OpenVINO toolkit to optimize trained models. Partnering with Intel is helping congatec to deliver very competitive products that are enabling the next wave of medical diagnostics.

Learn More

congatec home page

congatec medical landing page

COM Express Type 6 Ecosystem

Intel® Core™ Ultra Processors (H & U-SKUs)

Intel® Arc™ GPUs

Intel® Industry Solutions Builders

[i] Intel® Core™ Ultra 7 155H Processor: 1-node, 1x Intel® Core™ Ultra 7 155H Processor with 22 cores. Total DDR5 memory was 32 GB (2 slots/ 16GB/ 4800 MHz); microcode 0X17. Intel® Hyper-Threading Technology - enabled; Intel® Turbo Boost Technology - enabled. BIOS version: 5.3. Application storage is 1TB Kingston OM8SEP41024Q-A0; Network controller: Intel® RTL8125 operating at 2.5 Gbps.

Software: OS was Ubuntu 22.04.4 LTS; kernel was 6.7.10-060710-generic. Benchmark / workload software: OpenVINO™ toolkit 2024.4.0; Models tested: ResNet-50-TF, Yolo-V5S, EfficientNet-B0, Mobilenet-V2. Compiler was GCC 11.4.0; Libraries were DPDK 22.11.0. Other software: OpenCL Compute Runtime Version 24.17.29377.6

Intel® Core™ Ultra 7 265H Processor SUT: 1-node, 1x Intel® Core™ Ultra 7 265H Processor with 16 cores. Total DDR5 memory was 16 GB (2 slots/ 8 GB/ 6400 MHz); microcode 0x110; Intel® Hyper-Threading Technology – not available; Intel® Turbo Boost Technology - enabled. BIOS version: MTLPEMI1.R00.4341.D43.2409200738. Application storage is 1x 465.8G WD BLACK SN770 500GB, 1x 465.8G Samsung SSD 980 500GB.

Software: OS was Ubuntu 2024.04.1 LTS; kernel was 6.11-intel. Benchmark/workload software: OpenVINO™ toolkit 2024.4.0; Models tested: ResNet-50-TF, Yolo-V5S, EfficientNet-B0, Mobilenet-V2. Compiler was GCC 11.4.0; Libraries were DPDK 22.11.0. Other software: OpenCL Compute Runtime Version 24.17.29377.6

Test conducted by Intel on 11/7/24. These tests were not run on congatec modules. These were run by Intel on Intel® hardware. We are including the performance in this paper to indicate that we'd expect similar performance on congatec modules because they use these Intel® processor SKUs.

0625/CG/DJA/PDF Please Recycle 365610-001US