Solution Brief

Medical Imaging, Ultrasound

Intel® Core™ Ultra Processor Powers Samsung Medison's AI Ultrasound

Samsung Medison ultrasound systems run Live HeartAssist* and NerveTrack* AI applications using built-in GPU and NPU on Intel Core Ultra processor, eliminating discrete GPU dependency

Artificial intelligence (AI) offers a wide range of benefits to ultrasound machines, including faster image processing, improvements to clinical workflows, and real-time analytics. However, the high cost and power consumption of discrete GPUs complicate product designs and significantly increase the cost of medical imaging.

Intel is changing this paradigm with its Intel® Core™ Ultra processors that feature integrated GPU and neural processing unit (NPU) accelerators. This capability allows device manufacturers to usher in a wave of entry-level and mainstream ultrasound machines that offer both real-time imaging and real-time AI analytics.

Samsung Medison, one of the world's leading ultrasound manufacturers, is working with Intel to develop two new ultrasound applications – Live HeartAssist and NerveTrack— for their new line of AI-enabled ultrasound machines based on Intel Core Ultra processors.

With the shortage of nurses, real-time AI analytics in ultrasound improves clinical workflow by saving time. Take, for example, the Live HeartAssist application. Traditionally, conducting a fetal heart assessment can take 45 minutes to an hour, but with the Live HeartAssist AI app this same procedure takes only a few minutes.

The heterogeneous computing built into the Intel Core Ultra processors allows ultrasound developers to offload AI tasks to either the NPU or integrated GPU, eliminating the need for a discrete GPU for these inferencing use cases. This breakthrough enables ultrasound developers to integrate advanced AI features into mid and entry-level ultrasound devices, making cutting-edge imaging technology more accessible and cost-effective.

Ultrasound devices are battery-powered, so having built-in low power AI accelerator engines reduces the power draw, lengthening the battery life.

Samsung Medison is AI Ultrasound Leader

Samsung Medison, with headquarters in South Korea, is that country's largest manufacturer of ultrasound diagnostic systems and one of the top ten medical imaging manufacturers worldwide. The company has a history of incorporating advanced technologies and design capabilities – including AI – into its ultrasound systems.

The AI processing performance built into the Intel Core Ultra processors enable the following applications:

Live HeartAssist is an on-device real-time AI application that performs a fetal heart ultrasound on patients in a dramatically reduced timeframe. Without Live HeartAssist, a fetal heart ultrasound can take up to 45 minutes. With Live HeartAssist, practitioners

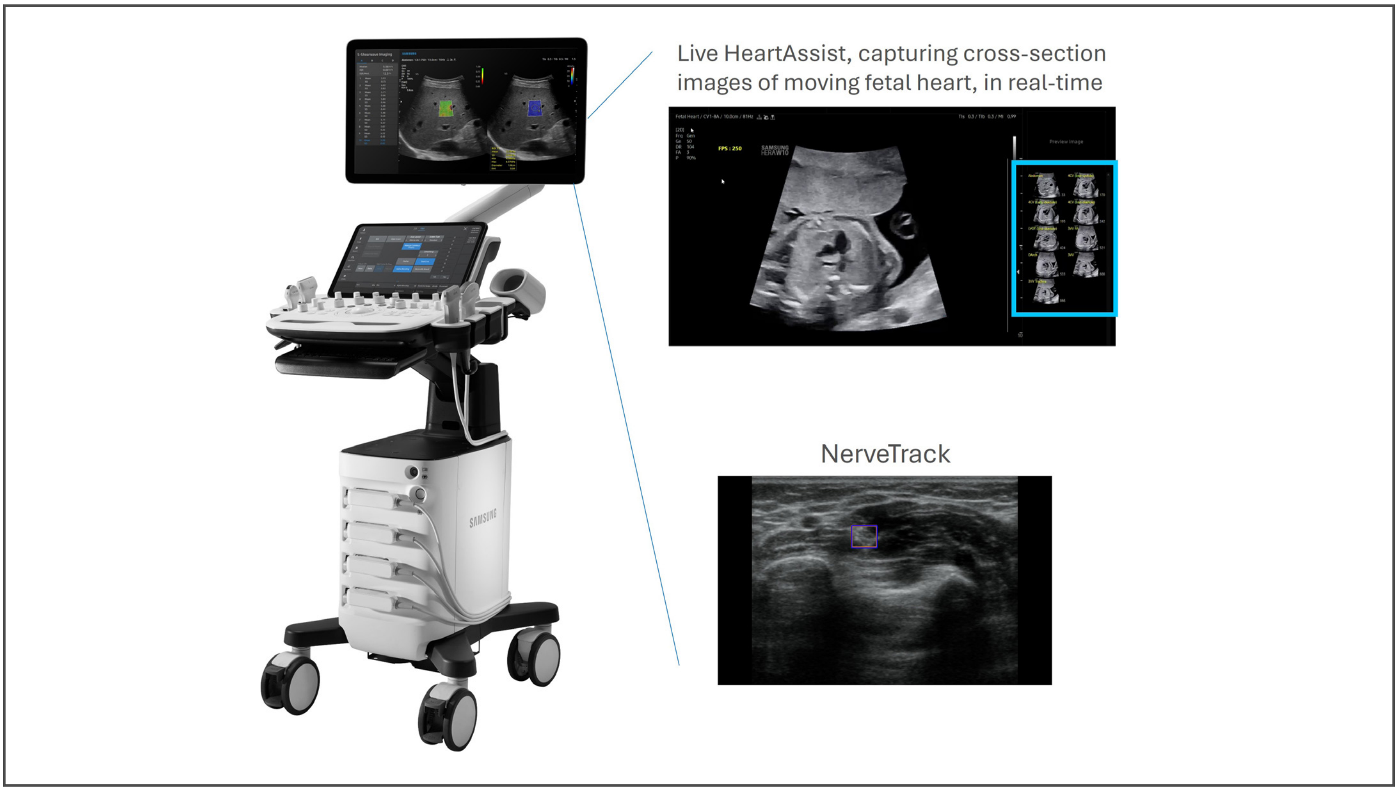

Live HeartAssist, capturing cross-section images of moving fetal heart, in real-time
NerveTrack

Figure 1. Samsung Medison V4 ultrasound is one model that targets use of Intel Core Ultra processors for AI capabilities. Call outs show examples of imaging from Live HeartAssist and NerveTrack.

can see, in minutes, 10 cross sections of the fetus's heart, with the ability to observe some of the fetus's ventricles, ventricular volumes and other heart structures.

NerveTrack is a US Food and Drug Administration-cleared ultrasound feature (see Figure 1) that detects 10 nerves in the body, including the ulnar nerve, greater occipital nerve and interscalene brachial plexus. This feature supports the work of anesthesiologists who are using ultrasound machines for ultrasound-guided regional anesthesia (UGRA) for needle-based interventions. In real time, NerveTrack guides anesthesiologists in administering anesthesia in nerves that can be as small as two millimeters in diameter.

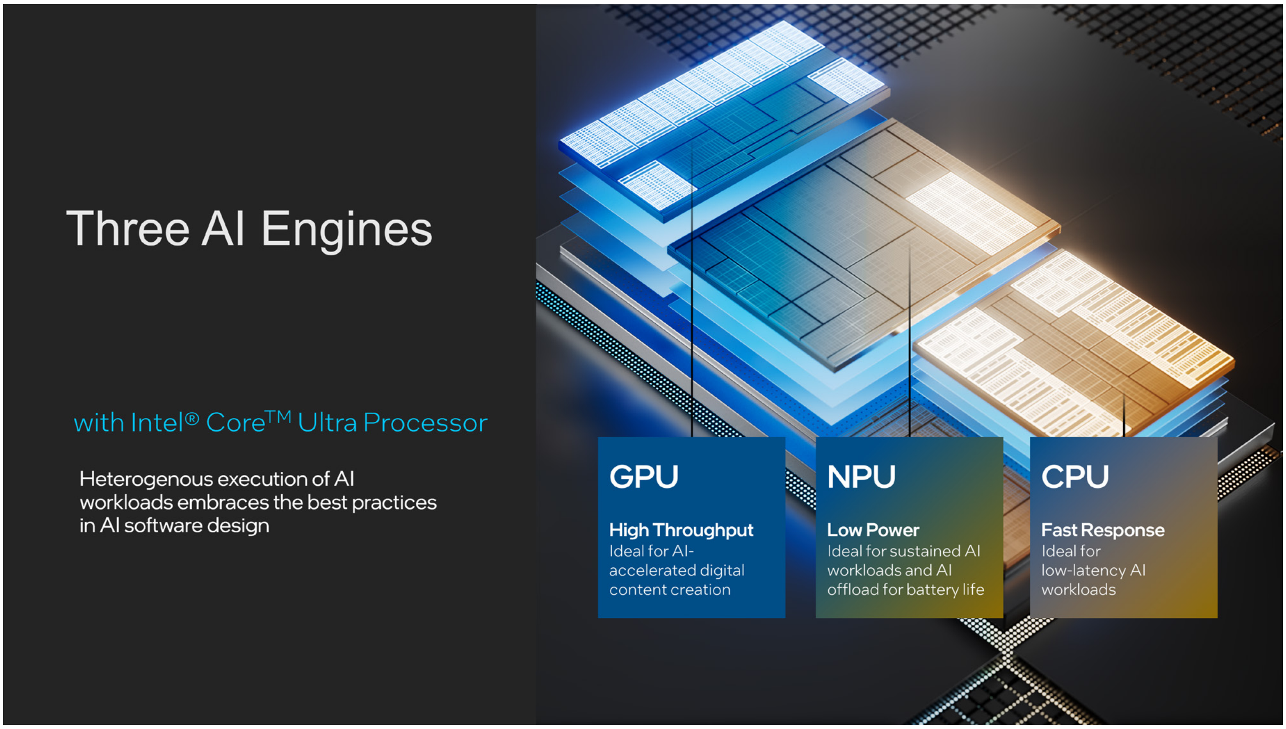
AI Applications Powered by Intel Core Ultra Processors

Samsung Medison chose the Intel® Core™ Ultra processor family to give entry-level and mainstream ultrasounds the performance to run the NerveTrack and Live HeartAssist imaging applications. This processor family is designed for both embedded and AI PC products. It features three compute engines (see Figure 2) in the same System-on-a-Chip (SoC), working together to accelerate AI inferencing. These include:

- **Performance and Efficiency CPU Cores**: The CPU in each Intel Core Ultra processor features between 12 and 16 CPU cores. The CPU features a hybrid architecture combining Performance-cores (P-cores) and single-threaded Efficient-cores (E-cores) on a single processor die, allowing customers to optimize the performance of their products and meet the needs of a variety of use cases. For example, the 16-core devices in the family offers eight E-cores, six P-cores and two low power E-cores and up to 22 threads. Due to the combination of P-cores and E-Cores, the Intel Core Ultra processor family supports thermal design power (TDP) as low as 15W.
- **GPU Offers Parallelized Processing**: The Intel Core Ultra processors feature integrated Intel® Arc™ graphics GPUs that provide 16 vector engines, allowing a high degree of parallelization of AI workloads. The AI performance is enhanced by built in Intel® Deep Learning Boost (Intel® DL Boost) instructions for additional performance.
- **NPU for Sustained AI Throughput**: The Intel Core Ultra processors feature the Intel® AI Boost Neural Processing Unit (NPU) accelerator, delivering up to 11 trillion operations per second (eTOPS). The NPU can be used for fast and low-power inferencing requirements.

All three compute engines are supported by the Intel® Distribution of OpenVINO™ toolkit (2024.0 or later) to accelerate AI inferencing and streamline AI development for computer vision and generative AI applications.

These integrated accelerators were important to Samsung Medison because they allowed the company to eliminate the need for an expensive and high-power external GPU card for AI processing. Live HeartAssist runs completely on the integrated GPU, and NerveTrack can run either completely on the integrated GPU or NPU, freeing up the CPU for other tasks and giving developers the flexibility to run AI apps on different accelerator engines.



Three AI Engines

with Intel® Core™ Ultra Processor

Heterogenous execution of AI workloads embraces the best practices in AI software design

GPU

High Throughput
Ideal for AI-accelerated digital content creation

NPU

Low Power
Ideal for sustained AI workloads and AI offload for battery life

CPU

Fast Response
Ideal for low-latency AI workloads

Figure 2. Intel Core Ultra processors feature three integrated processing engines that can handle AI workloads.

Model Optimization and Deployment with OpenVINO™ Toolkit:

OpenVINO™ toolkit was used to optimize Live HeartAssist and NerveTrack on Intel Core Ultra processors. OpenVINO toolkit is an open-source toolkit for optimizing and deploying deep learning models across heterogenous devices (CPU, GPU, NPU) in Intel® platforms.

It accelerates deep learning inference across various use cases, such as generative AI, video, audio, and language with models from popular frameworks like PyTorch, TensorFlow, TensorFlow Lite (TFLite), ONNX, and PaddlePaddle. The power and simplicity of OpenVINO toolkit is shown in the code snippets below.

Figure 3 shows the installation of OpenVINO toolkit on a Windows system. For more information on the supported hardware, software, and operating systems, refer to this link.

```
# Open the Windows command-line interface as a
non-administrator. Create a Python virtual
environment and activate it

python -m venv openvino_venv

.\openvino_venv\Scripts\activate

# Upgrade pip version

python -m pip install pip –upgrade

# Install openvino and dependencies

pip install openvino==2024.5.0 torch opencv-
python
```

Figure 3. Installation of OpenVINO toolkit.

The code snippet in Figure 4 shows the steps to convert the PyTorch model into intermediate representation format using convert_model() in OpenVINO toolkit. For more information on the model conversion from other frameworks, refer to this link.

```
import openvino as ov

import torch

import cv2

# Load the existing PyTorch model

model = torch.load(“model.pt”)

model.eval()

# Convert model into intermediate representation

ov_model = ov.convert_model(model)

# Save optimized model files (.xml, .bin) on disk

ov.save_model(ov_model, “ov_model.xml”)
```

Figure 4. Conversion of PyTorch model into intermediate representation.

The last step is seen in Figure 5 with the model deployment and inference execution. Included in these commands are those necessary to run the workload on the integrated GPU or the NPU (see blue text).

```
#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”)
# NPU
compiled_model = core.compile_model(“ov_model.
xml”, device_name=“NPU”)
# 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. Model deployment and inference execution. Code snippet shows how easy it is to run the workload on the GPU or NPU.

Ease of Running Code on Desired Compute Engine

The code snippet in Figure 5 illustrates the steps needed to compile the model and run the inference on the desired device using OpenVINO™ Runtime API. Step #2 shows the convenience of selecting the inference device (GPU or NPU) using the device_name parameter while loading and compiling the model with compile_model().

Moreover, the subsequent steps for image preprocessing and inference execution remain the same – independent of the inference device. This feature enables developers to maintain the same version of the source code for running inference across the heterogenous devices (CPU, GPU, NPU).

OpenVINO toolkit support for the Intel Core Ultra processor can enhance AI performance by routing workloads to the right compute engine. This integration helps streamline AI workflows through cross-architectural programming capabilities and automatic compute engine detection. OpenVINO toolkit also offers support and optimizations for popular AI frameworks, including TensorFlow, PyTorch, and ONNX, to help boost performance and simplify development.

> *“The Intel Core Ultra processor has ushered in a new era of innovation in healthcare imaging. Our tests have revealed a remarkable 22% and 25% increase in AI performance throughput for NerveTrack and Live HeartAssist real-time ultrasound imaging applications, respectively, compared to previous generations. This breakthrough, attributed in part to the built-in Intel Arc GPU, allows us to offer advanced AI features in next generation mid- and entry-level ultrasound devices without the need for discrete GPUs, resulting in more accessible and cost-effective, cutting-edge imaging technology. The Intel Core Ultra processor is a game-changer, enhancing patient care and opening new possibilities in the healthcare market.”*
>
> - SungShik Baik, Principal Engineer, Samsung Medison

Better AI Throughput than Entry-Level Discrete GPU

To illustrate the advantage of the Intel Core Ultra processor family, Samsung Medison compared the imaging capabilities of two systems.[1]

Config 1 is a commercial ultrasound based on the 8th Generation Intel® Core™ i5-8400H Processor and an entry-level discrete GPU (full configuration details are in footnote 1).

Config 2 leveraged the Intel® Core™ Ultra 7 Processor 165H with integrated GPU.

Live HeartAssist test results (see Figure 6 table and graph) show 250 frames per second (FPS) throughput for Config 2 compared to 200 FPS for Config 1, showing a 25% increase in AI performance throughput for the Intel Core Ultra 7 Processor 165H.

	CONFIG 1:	CONFIG 2:	% Improvement
	Intel Core i5-8400H Processor + competitor discrete GPU	Intel Core Ultra 7 Processor 165H	
Live HeartAssist	200 FPS	250 FPS	25%
NerveTrack	41 FPS	50 FPS	22%

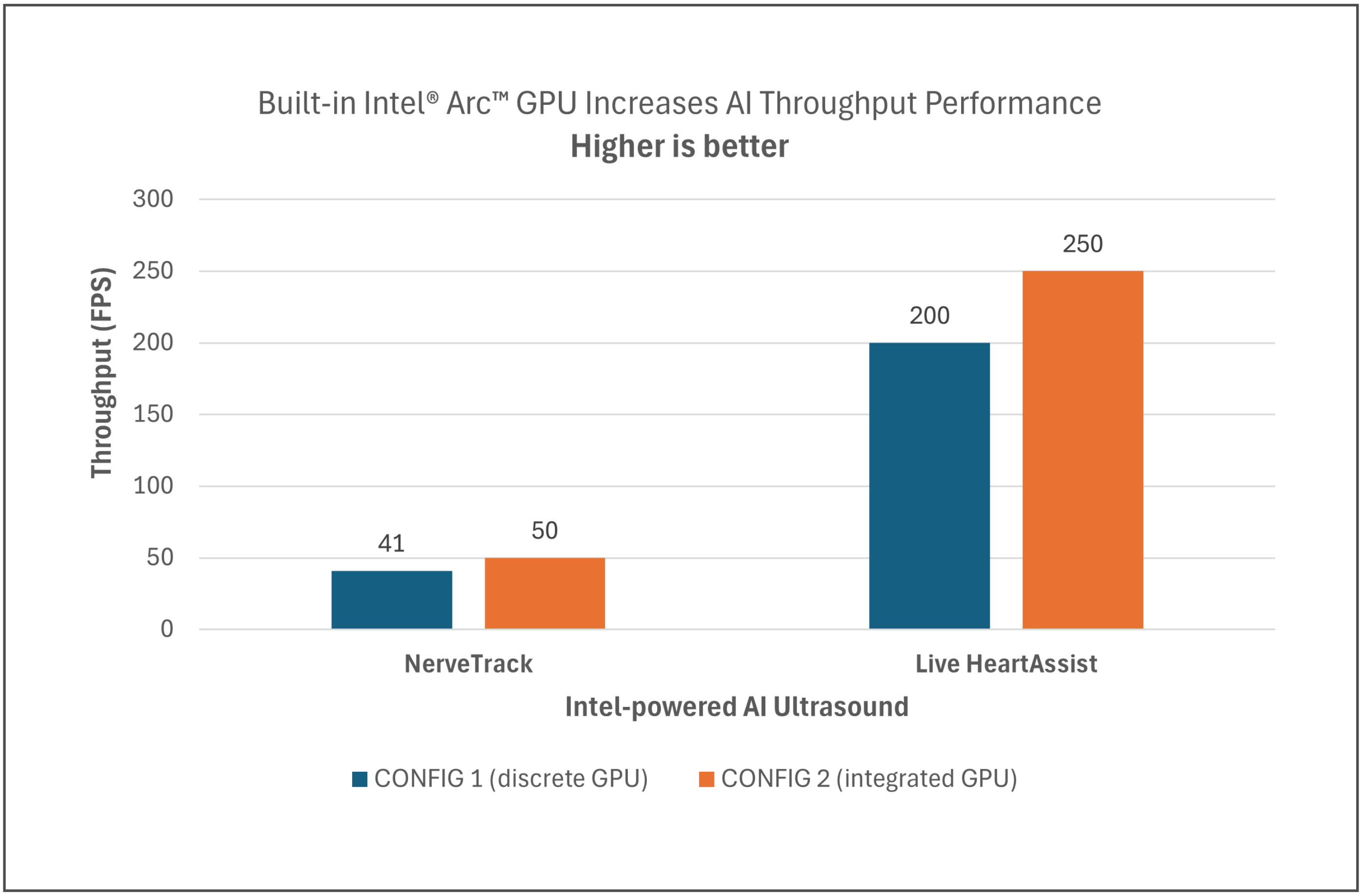


Figure 6. Performance results (in table and graph) show integrated GPU-based ultrasound providing more than 20% imaging improvement over ultrasound with discrete GPU (higher is better).

Conclusion

The use of AI in ultrasound systems is bringing new functionality, on-device and real-time AI analytics, faster imaging, and more efficient workflows to health care providers and their patients. Samsung Medison is a pioneer in this area offering its AI-powered NerveTrack and Live HeartAssist applications for its mid- and entry-level ultrasound systems. These applications need the accelerated processing of a GPU, but discrete GPUs are power hungry and can add extra cost to an ultrasound system. Samsung Medison chose the Intel Core Ultra processor family because these processors feature three integrated AI processing engines and also make use of OpenVINO toolkit, which allows Samsung Medison to easily accelerate and run their solutions on these engines. The combined solution delivers AI performance without the downsides of a dedicated GPU.

Learn More

Samsung Medison Ultrasound Systems

AI Inferencing in Ultrasound with Intel® Core™ Ultra Processors

Intel® Core™ Ultra Processors

Intel® Distribution of OpenVINO™ toolkit

[1] Config 1: 4 core (8 thread) 8th Generation Intel® Core™ i5-8400H Processor + NVIDIA Turing Architecture dGPU. Total DDR4 memory was 16 GB (2 slots/ 8GB/ 2666 MHz); BIOS version: KF08000Q060V117. Boot storage is 512GB M.2 NVMe. Software: OS was Windows 11 23H2; kernel was X64. Workload: NerveTrack. Compiler was C# / OpenCV / OneAPI; Libraries were IPP / MKL / OneAPI. Graphics Driver: Driver date: 2023-09-25, Driver version: 31.0.101.4725. NPU Driver: Driver date: 2024-04-04; Driver version: 32.0.100.2267. Test conducted by Samsung Medison on Nov. 24, 2023.

Config 2: Intel® Core™ Ultra 7 165H Processor with 16 cores. Total DDR5 memory was 16 GB (2 slots/ 8GB/ 5600 MHz); BIOS version: MTLPEMI.R00.3323.D48.2309202233. Boot storage is 512GB M.2 NVMe. Software: OS was Windows 11 23H2; kernel was X64. Workload: NerveTrack. Compiler was C# / OpenCV / OneAPI; Libraries were IPP / MKL / OneAPI. Graphics Driver: Driver date: 2023-09-25, Driver version: 31.0.101.4725.NPU Driver: Driver date: 2024-04-04; Driver version: 32.0.100.2267. Test conducted by Samsung Medison on Nov. 24, 2023.

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.

0125/CG/DJA/PDF Please Recycle 363990-001US