Product Brief

intel ai

Supercharge Applications with AI on an AI PC with OpenVINO™

OpenVINO™ What toolkit is need to power AI applications?

OpenVINO™ toolkit is an open-source software kit for optimizing and deploying AI workloads such as generative AI, computer vision, large language models to deliver cutting-edge applications for the AI PC.

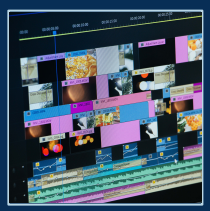
What is the AI PC?

A PC with the latest Intel® Core™ Ultra processor that brings fresh AI experiences in productivity, creativity, and security through a combination of the CPU, GPU, and the all-new NPU.

intel CORE ULTRA

What workloads can be accelerated with AI?

Photo, Video & Music Editing



Consumer & Commercial **AI PC**

Virus/Threat Detection

Content Generation

167Mu by 2027[1]

AI PCs will represent nearly 60% of all PC shipments worldwide, up from 50 million units in 2024

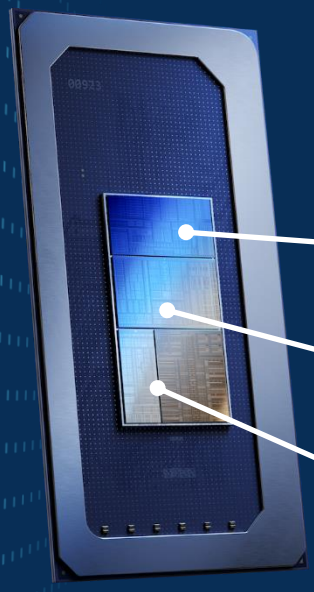
Personal AI Assistance

Collaboration Effects

Revision #827780

Free Download > openvino.ai

Why Intel® Core™ Ultra Processors?



The right balance of power and performance for AI

CPU	**Fast Response** Ideal for lightweight, single-inference, low-latency AI tasks
GPU	**Performance Parallelism & Throughput** Ideal for AI infused in Media/3D/Render pipeline
NPU	**Dedicated Low-Power AI Engine** Ideal for sustained AI and AI offload

How OpenVINO™ accelerates AI applications

Open-Source

Allow for redistribution and commercial use with a permissive open-source Apache* 2.0 license

Performance Optimized

Realize unparalleled performance with a toolkit optimized for Intel hardware

AI, DL Inference

Use across domains to run Generative AI, Computer Vision, Natural Language Processing, Large Language Models and Recommender System inference

Cross-Platform Support

Enhance AI accessibility across Intel® CPU, GPU and NPU

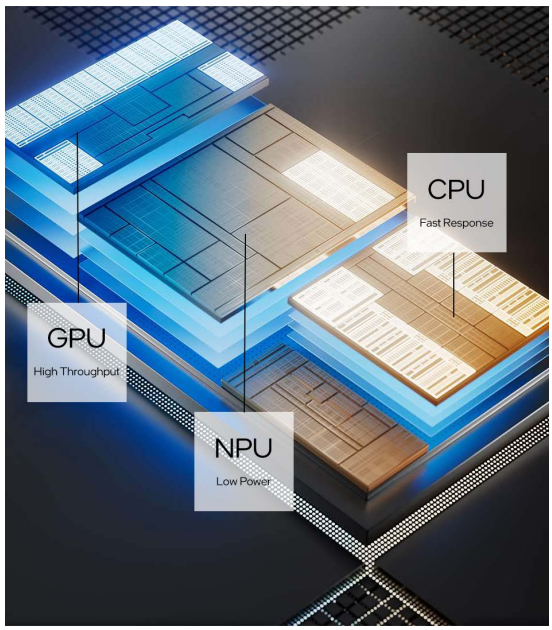
AUTO Device Plug-in

The Automatic Device Selection mode, or AUTO for short, uses a "virtual" device, that selects the accelerator for inference automatically

Detects available accelerator devices

Picks the one best suited for the task

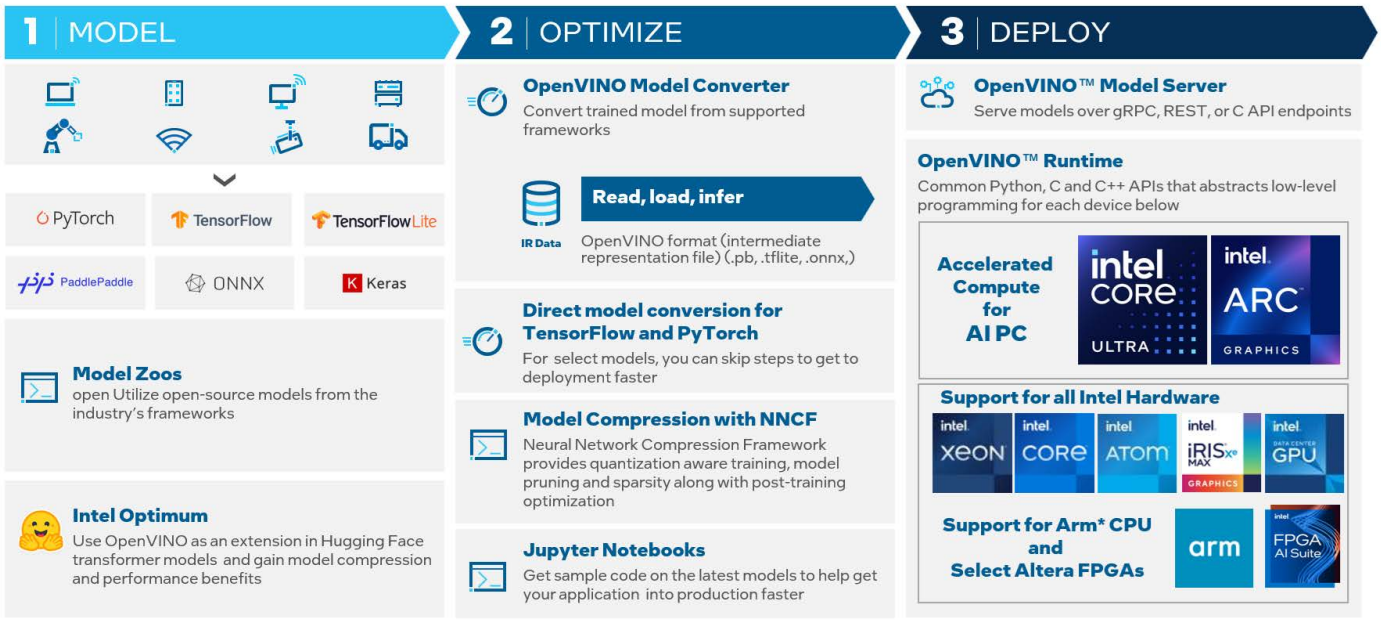
Configures ideal optimization settings



Its easy with a single line of code specifying the device name to AUTO

```
compiled_model = core.compile_model (model=model, device_name="AUTO")
```

Revision #827780



Key Components

Model Conversion API & Tools

Imports trained models from various frameworks (TensorFlow*, PyTorch*, ONNX*, PaddlePaddle*, Keras*, and more) and converts them to a unified intermediate representation file. Two simple API calls, convert_model() and save_model() optimizes and converts models to FP32 or FP16. Also available is the easy to use OpenVINO Converter (OVC) command-line tool providing the same great results.

Why it's important: The OpenVINO Model Converter (OVC) provides the biggest performance boost by conversion to data types that match hardware types (FP32/FP16). Further optimize with NNCF for smaller data types (INT8/INT4).

If your selected model is in one of the OpenVINO supported model formats, you can use it directly, without the need to save as OpenVINO IR. Conversion is performed automatically before inference for maximum convenience.

OpenVINO™ Model Server (OVMS)

A high-performance system for serving models. Implemented in C++ for scalability and optimized for deployment on Intel architectures, the model server uses the same architecture and API as TensorFlow Serving and KServe while applying OpenVINO for inference execution. Inference service is provided via gRPC or REST API, making deploying new algorithms and AI experiments easy.

Why it's important: Model Server hosts models and makes them accessible to software components over standard network protocols.

OpenVINO Runtime

A simple and unified API for inference across multiple compute architectures. It allows heterogeneous execution of layers across hardware targets (CPU, GPU, NPU, and third party ARM* architecture CPUs). The API supports C, C++, Python*, and JavaScript* interfaces, dynamically loading plugins for each hardware type.

The OpenVINO Runtime is deployed inside applications to deliver AI inference acceleration using customer developed models for their applications use cases.

Why it's important: Delivers superior performance for each type without requiring users to implement and maintain multiple code pathways.

Neural Network Compression Framework (NNCF)

Model optimization is an optional offline step of improving the final model performance; from 32-bit, to 16-bit, 8-bit, and 4-bit quantization, pruning, and more.

- Post-training Quantization optimizes the inference of deep learning models by applying the post-training 8-bit integer quantization that does not require model retraining or fine-tuning.
- Training-time Optimization, a suite of advanced methods for training-time model optimization supports methods like Quantization-aware Training, Structured and Unstructured Pruning, etc.
- Weight Compression, an easy-to-use method for Large Language Models' footprint reduction and inference acceleration.

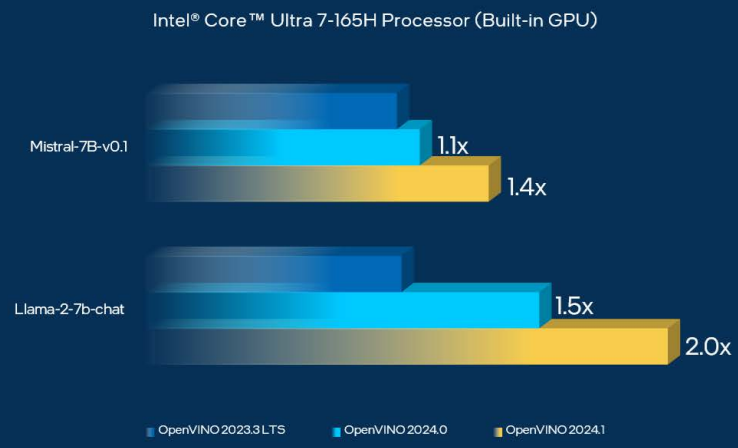
Why it's important: The NNCF reduces neural network sizes for faster training cycles and more-compact models for deep learning inference.

Performance Benchmarks

Accelerate AI PC Development with Optimized GenAI Model Performance

Significant performance gains with every OpenVINO™ release (2023.3 LTS > 2024.0 > 2024.1)

Get the most from AI deployments with up to 2X improvement in OpenVINO performance on GenAI models



Metric: 2nd Token Throughput as Tokens Per Second. Precision: INT4, Batch size: 1, Input: 1024 tokens, Output: 128 tokens, Beam search: 1

Maximize your Stable Diffusion performance on Intel® Core™ Ultra Processors

OpenVINO™ release 2024.1 offers a significant performance boost over 2024.0

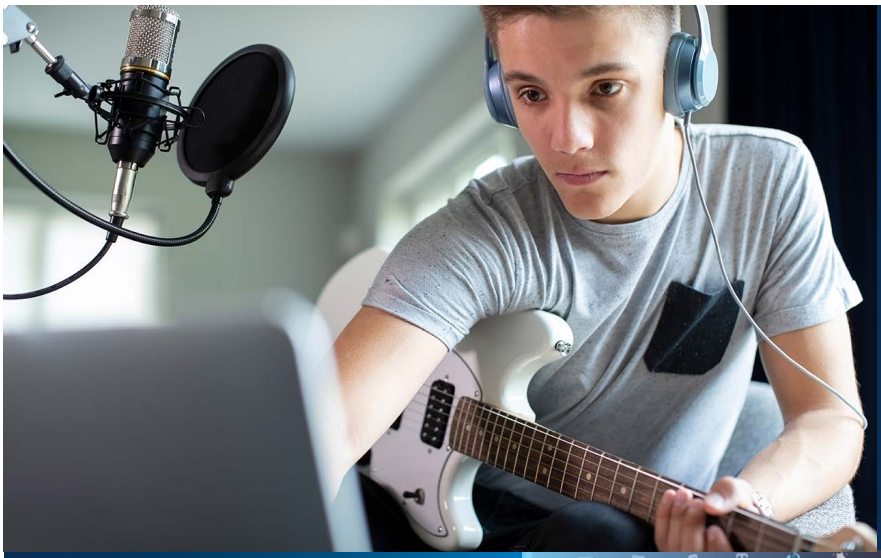
Up to 1.2X performance improvement on Intel® Core™ Ultra Processors with OpenVINO™ toolkit.

Metric: 2nd token throughput as Tokens Per Second. Input tokens: 1024, Batch size: 1, Precision: INT8

System board	Intel Corporation Reef Ridge/Astral Peak CRB
CPU	Intel® Core Ultra 7-165H @1.4 GHz.
Sockets, physical cores/socket	1, 6P+8E+2e
Hyperthreading/turbo setting	Enabled/On
Memory	2x32 GB DDR5 5600MHz
OS	Windows 11
Kernel	10.0.22631 Build 22631
Software	Intel® Distribution of OpenVINO™ Toolkit 2024.0 / 2024.1
BIOS	MTLPEMI1.R00.3471.D56.2403181159
BIOS release date	3/18/2024
BIOS setting	Select optimized default settings, save, and exit
Microcode	1C
Test date	April 2024
Precision and batch size	Int8/Batch 1
Power (TDP)/socket	28W

Workloads
I LLM: Stable-Diffusion-V2-1: (input token length: 1024, steps: 20, image size: 512x512 pixel).
Llama-2-7b-chat, Mistral-7b-v0.1, ChatGLM2-6b: (input token length: 1024, output size: 128 beam: 1)

Success Stories



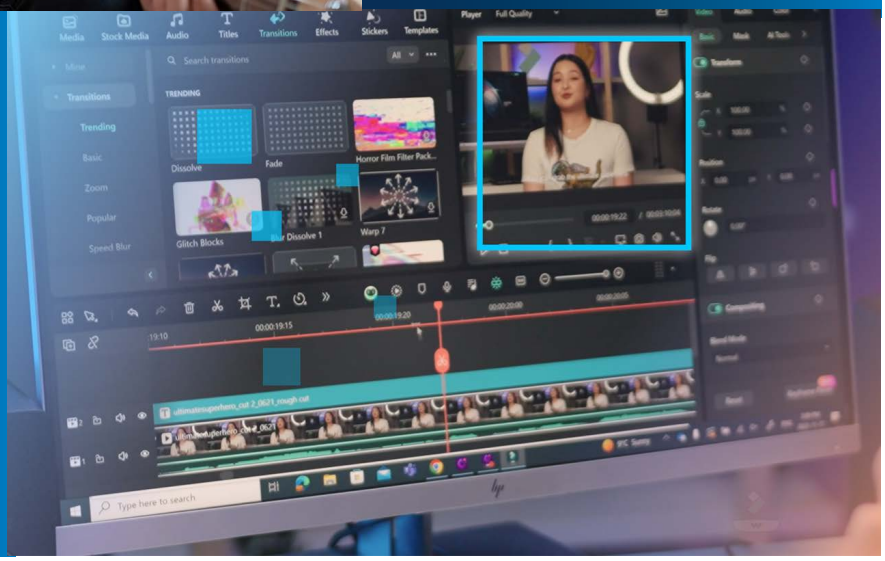
Audacity*

[Audacity harnesses AI for Podcasting and Music Production with Intel]

Wondershare Filmora*

wondershare

[Wondershare Filmora video editor simplifies the video creation process with AI-powered Text-Based Editing, Copilot, and Music Generator]



Resources	Resource Location
OpenVINO™ toolkit webpage	https://openvino.ai
OpenVINO™ toolkit Github*	https://github.com/openvinotoolkit
OpenVINO™ toolkit downloads	https://www.intel.com/content/www/us/en/developer/tools/openvino-toolkit/download.html
AI PC. Transformational Technology	https://intel.com/aipc
Intel® Core™ Ultra™ Processors Family	https://www.intel.com/content/www/us/en/products/details/processors/core-ultra.html

intel.

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.

Revision #827780