Product Brief

Generative AI, Computer Vision, and Large Language Models with the Intel® Distribution of OpenVINO™ Toolkit

# Supercharge Applications with AI from Edge to Cloud

## Develop and Deploy cutting edge AI workloads with Deep Learning Inference

Computer Vision

Generative AI & Large Language Models


Recommender Systems

Natural Language Processing

Intel® Distribution of OpenVINO™ toolkit is an open-source software kit that helps developers and enterprises speed up AI workloads such as generative AI, computer vision, large language models and natural language processing. Streamline deep learning inference development and deployment, and enable easy, heterogeneous execution across Intel® platforms from edge to cloud.

### 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 and 3rd party hardware including ARM*

### 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, NPU, or 3rd party systems and devices through modular and scalable plugin architecture

### Edge to Cloud

Deploy and scale across cloud infrastructure, edge device, and preferred systems with ease

OpenVINO™

## AI the Engine of Opportunity

Demand for AI accelerated solutions is increasing in all markets; enterprise, industrial and consumer alike.

AI Software
Global AI software revenue is projected to grow from USD 10.1 billion in 2018 to **$126.0 billion by 2025**[1]

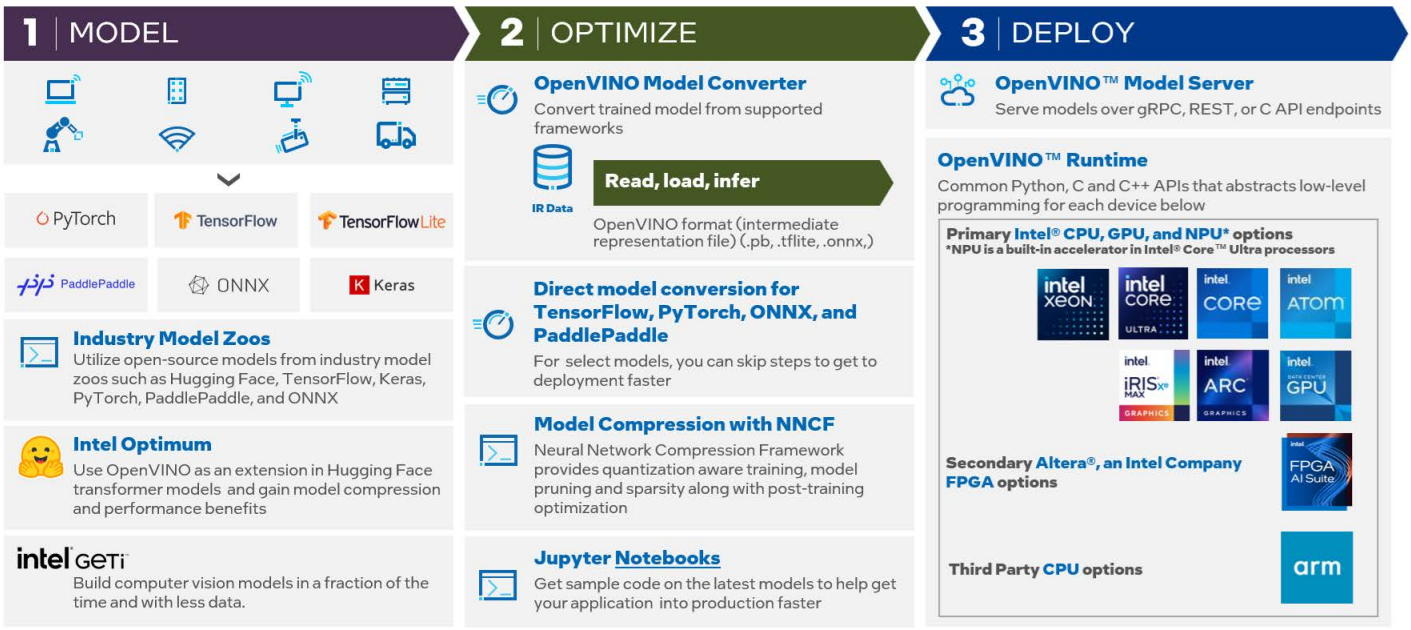
AI as a Service
The global Artificial Intelligence as a Service market was valued at USD 3.91 billion in 2020 and is expected to reach **$43.3 billion by 2026**[2]

AI Hardware
The global artificial intelligence (AI) hardware market was valued at USD 9.8 billion in 2019 and is expected to have a **37.5% growth rate through 2027**[3]

AI in PC
The AI PC will represent nearly 60% of all PC shipments worldwide, growing from 50 million units in 2024 to more than **167 million units by 2027**[4]

**Free Download >**
**software.intel.com/openvino-toolkit**

Revision #671060




## 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++, and Python* 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.

## Key Component Toolkit Add-Ons

### Data Management Framework

Use this add-on to build, transform, and analyze datasets.

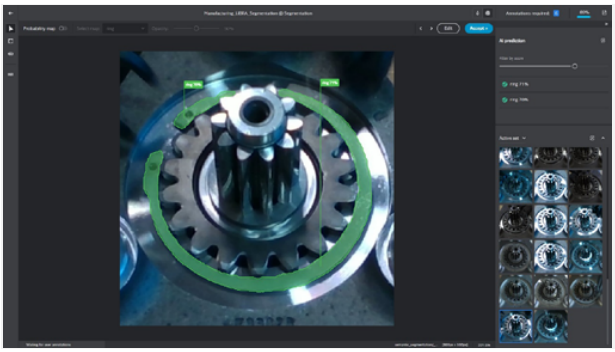
**Why it's important:** A framework and CLI tool to build, transform, and analyze datasets. Add-on provides dataset reading, writing, conversion in any direction, as well as Dataset building, quality checking, comparison, statistics, and model integration.

### Benchmark Tool

Estimate deep learning inference performance on supported devices.

**Why it's important:** The benchmark app allows users to provide high-level "performance hints" for setting latency-focused or throughput-focused inference modes. This hint causes the runtime to automatically adjust runtime parameters, such as the number of processing streams and inference batch size, to prioritize for reduced latency or high throughput.

## Intel® Geti™ Platform



**Accelerate Computer Vision Model Development**

Intel has created the new Intel® Geti™ computer vision platform—a collaborative, intuitive platform where data scientists, developers, and domain experts can work together to train deep learning models for specific computer vision applications. Deploy your custom models with the OpenVINO™ toolkit.
The Intel Geti platform speeds up model development by simplifying labor-intensive tasks and harnessing greater collaboration between teams on one single platform for data labeling, model training, optimization, and retraining. Most importantly, the solution unlocks faster time-to-value for digitalization initiatives with AI.

## OpenVINO™ Toolkit Success Stories

**Audi* automated and enhanced critical quality-control processes in its factories using AI. They are able to reduce human error and ensure all cars are built with even more accuracy and precision.**

**Audi Weld Inspection Hits 100 Percent with Intel**

**Pathr.ai* delivers spatial intelligence to help improve supermarket operations.**

**Pathr.ai**

## Technical Specifications

### CPU

**Supported Hardware**

- Preview: Intel® Xeon® 6 processors
- 1st to 5th generation of Intel® Xeon® Scalable processors
- Intel® Core™ Ultra processor‡
- 6th to 14th generation Intel® Core™ processors
- Intel® Pentium® processor N4200/5, N3350/5, N3450/5 with Intel® HD Graphics
- Intel Atom® processor with Intel® Streaming SIMD Extensions 4.2 (Intel® SSE4.2)
- ARM* and ARM64 CPUs, Apple* M1, M2, and Raspberry Pi*

**Supported Operating System**

- Preview: Ubuntu 24.04, 64 bit
- Ubuntu* 22.04 long-term support (LTS), 64 bit (Kernel 5.15+)‡
- Ubuntu 20.04 LTS, 64 bit (Kernel 5.15+)
- Ubuntu 18.04 LTS with limitations, 64 bit (Kernel 5.4+)
- Windows* 10 and 11‡
- macOS* 10.15 and above, 64 bit
- macOS* 11 and above, ARM64
- Red Hat* Enterprise Linux* 8, 64 bit
- Debian* 9 ARM64 and ARM
- CentOS* 7 64-bit
- ‡ denotes Compatible operating system with Intel Neural Processing Unit

### NPU

**Supported Hardware**

- Intel® Core™ Ultra processor

**Supported Operating System**

- Ubuntu* 22.04 long-term support (LTS), 64 bit (Kernel 5.15+)‡
- Windows* 11‡

### GPU

**Supported Hardware**

Discrete Graphics

- Intel® Data Center GPU Max Series
- Intel® Data Center GPU Flex Series
- Intel® Arc ™ GPU

Integrated Graphics

- Intel® HD Graphics
- Intel® UHD Graphics
- Intel® Iris® Pro Graphics
- Intel® Iris® Xe Graphics
- Intel® Iris® Xe MAX Graphics

**Supported Operating System**

- Preview: Ubuntu 24.04 (64 bit)
- Ubuntu 20.04 LTS (64 bit)
- Ubuntu 22.04 LTS (64 bit)
- Windows 10 (64 bit)
- Windows 11 (64 bit)
- Red Hat Enterprise Linux 8 (64 bit)
- CentOS 7 64-bit

| 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 |
| Intel® Developer Cloud for the Edge | https://devcloud.intel.com/edge |
| Industry Success Stories | https://www.intel.com/content/www/us/en/developer/tools/openvino-toolkit/industry.html |

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.

All monetary figures are in US dollars.

1. Tractica, 2020, Artificial Intelligence Market Forecasts.
2. Mordor Intelligence, 2020, Artificial Intelligence-as-a-Service Market Report.
3. MarketWatch, Artificial Intelligence (AI) Hardware Market 2021 Industry Outlook, Current Status, Supply-Demand, Growth Opportunities, and Top Players Analysis 2030.
4. IDC, 2024, IDC Forecasts Artificial Intelligence PCs.



Revision #671060