Master AI Efficiency

Introduction to OpenVINO™ and AI Inference Optimization

19 Feb 2024

intel.

About our Speakers

Anisha Udayakumar

AI Software Evangelist (India), Intel

Anisha, an AI Software Evangelist at Intel, specializes in the OpenVINO™ toolkit, empowering developers to craft innovative AI solutions. With a background as an Innovation Consultant, she has driven sustainable tech solutions for global clients and continues to inspire the developer community with her insights and expertise.

Problem Statement

AI-Powered Interactive Learning Assistant for Classrooms

Objective

- Develop an AI-powered interactive learning assistant that enhances classroom engagement and supports both students and educators. The assistant should leverage AI to improve learning outcomes in one or more of the following ways:
  - Personalize learning content and feedback based on individual student progress
  - Generate answers, summaries, and study material for self-paced learning
  - Support multimodal interaction via speech, text, and visuals for inclusive engagement

- It should utilize AI and Generative AI (GenAI) models and be optimized with OpenVINO.

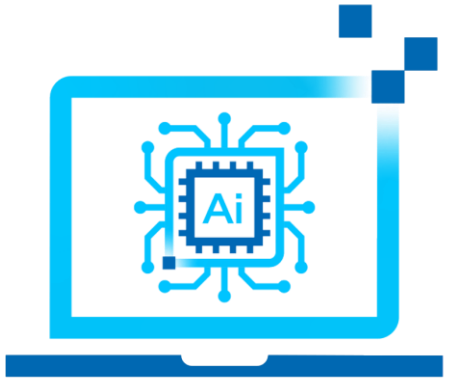
Key Expectations

- Address a specific learning or teaching challenge in classroom environments (e.g., engagement, accessibility, content generation) with an AI-driven approach..
- Use AI/GenAI/LLM models for tasks such as question answering, summarization, lesson planning, or multimodal content creation..
- Optimization with OpenVINO – Convert and optimize models for Intel® CPU, GPU, and NPU, ensuring low latency & high efficiency.
- Demo & AI Inference – Present a functional demo showcasing how the assistant interacts in real time with students/teachers and highlight performance benchmarks.

Different Types of Compute

Cloud

Edge



AI PC

We strike the balance…

Three AI Engines in Intel® Core™ Ultra processor

The right balance of power and performance for building and deploying AI models with the **OpenVINO™** toolkit

NPU

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

CPU

Fast Response
Ideal for low-latency AI workloads

GPU

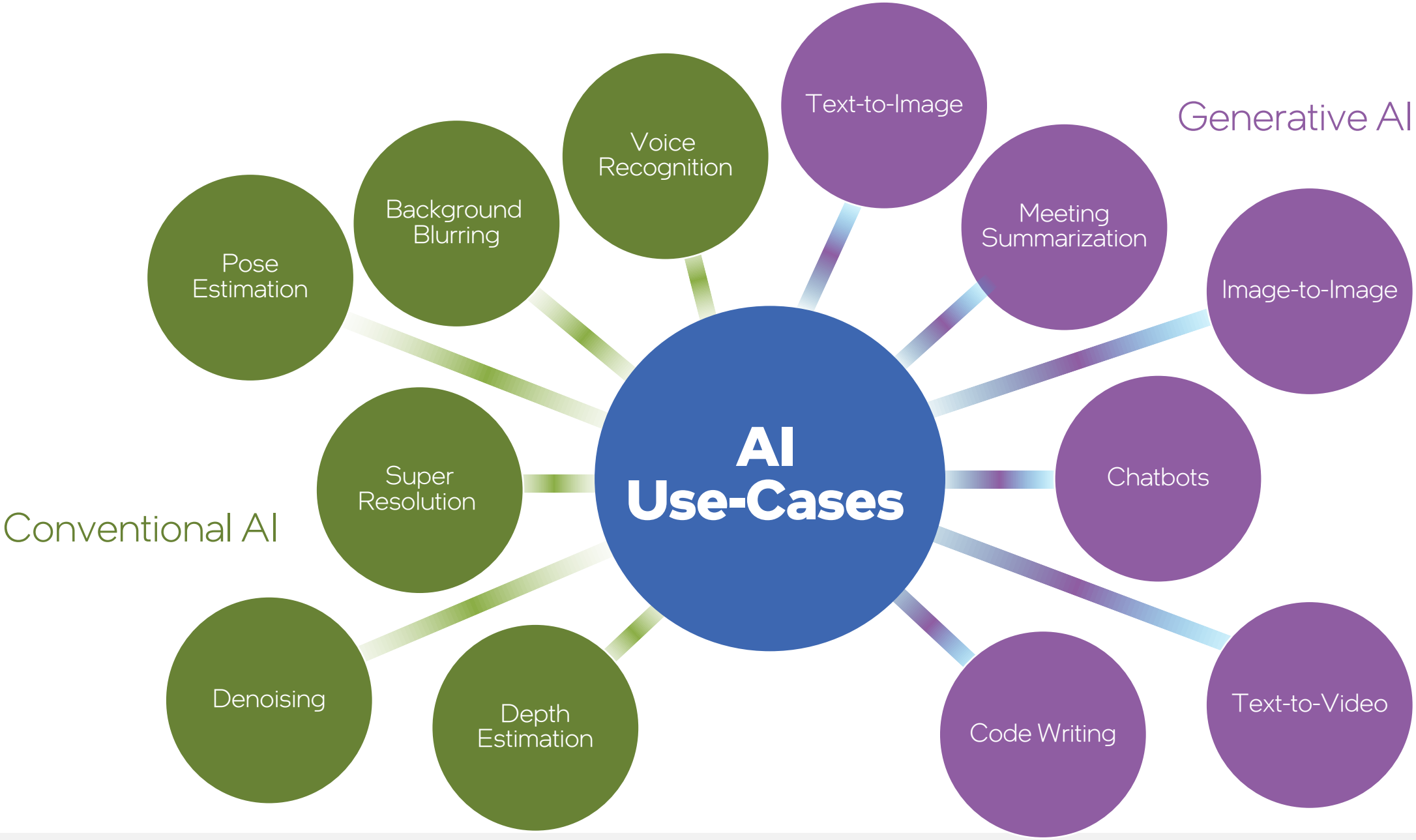
High Throughput
Ideal for AI-accelerated digital content creation and gaming

CPU/NPU/GPU Comparison

NPU

CPU

GPU


AI Use-Cases
Conventional AI
Pose Estimation
Background Blurring
Voice Recognition
Super Resolution
Denoising
Depth Estimation
Generative AI
Text-to-Image
Meeting Summarization
Image-to-Image
Chatbots
Code Writing
Text-to-Video

OpenVINO™

Visual
+
NLP, Audio, LLM, GenAI, Transformers...

and
Neural
Network

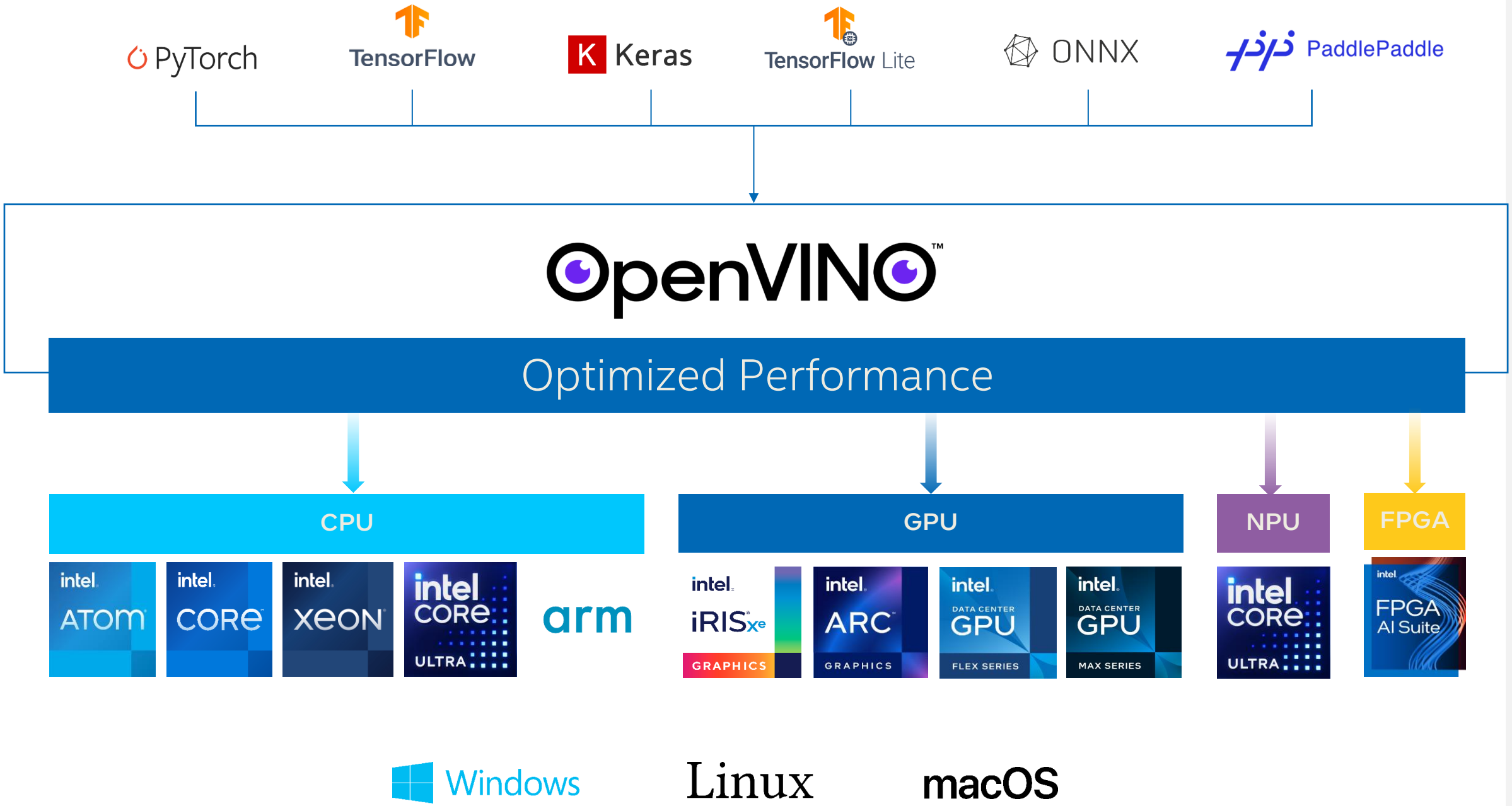
Optimization

Inference

OpenVINO™

DEVELOPER JOURNEY

1 MODEL → 2 OPTIMIZE → 3 DEPLOY


PyTorch
TensorFlow
Keras
TensorFlow Lite
ONNX
PaddlePaddle
OpenVINO™
Optimized Performance
CPU
GPU
NPU
FPGA
intel ATOM
intel CORE
intel XEON
intel CORE ULTRA
arm
intel iRISxe GRAPHICS
intel ARC GRAPHICS
intel DATA CENTER GPU FLEX SERIES
intel DATA CENTER GPU MAX SERIES
intel CORE ULTRA
intel FPGA AI Suite
Windows
Linux
macOS

Installation

ubuntu

Red Hat

docker

ANACONDA

CONDA-FORGE

CONAN
C/C++ Package Manager

npm

www.openvino.ai

Installation

```
pip install openvino
```

ubuntu

Red Hat

docker

ANACONDA

CONDA-FORGE

CONAN
C/C++ Package Manager

npm

www.openvino.ai

Installation

Install OpenVINO™ 2024.5

Package

- **OpenVINO Base Package** — Full inference with basic GenAI
- OpenVINO with GenAI — Full inference + complete GenAI package

Version

- **2024.5** — Recommended
- Nightly Build
- 2023.3 LTS
- 2022.3.2 LTS — Includes NCS2/HDDL support

Operating System

- **Windows**
- macOS
- Linux

Distribution

- **PIP** — Includes NPU plugin, Python API only
- OpenVINO Archives — Includes NPU plugin
- GitHub — Source
- Gitee — Source
- Docker
- Conda
- vcpkg — Source
- Conan
- npm — JavaScript API only

Install

```
# Step 1: Create virtual environment
python -m venv openvino_env
```

```
# Step 2: Activate virtual environment
openvino_env\Scripts\activate
```

```
# Step 3: Upgrade pip to latest version
python -m pip install --upgrade pip
```

```
# Step 4: Download and install the package
pip install openvino==2024.5.0
```

Resources

Installation Instructions Previous Releases System Requirements

Get Started Guide Troubleshooting Guide

Related Tools

Ready to run OpenVINO Notebooks

Hugging Face + Optimum Intel

OpenVINO Tokenizers to streamline tokenizer conversion

NNCF for implementing compression algorithms on models

OVMS for serving models optimized for deployment

Quickstart

```
pip install openvino
```

```python
from openvino import runtime as ov

img = load_img()

core = ov.Core()
# PyTorch, Tensorflow, ONNX, Keras, Tensorflow Lite, Paddlepaddle
model = core.read_model(model="model.pt")
compiled_model = core.compile_model(model=model, device_name="CPU")

output_layer = compiled_model.outputs[0]

result = compiled_model(img)[output_layer]
```

eaidova update on openvinio 2024.5 and nncf 2.14 (#2548)

Preview Code Blame 211 lines (136 loc) · 13.9 KB

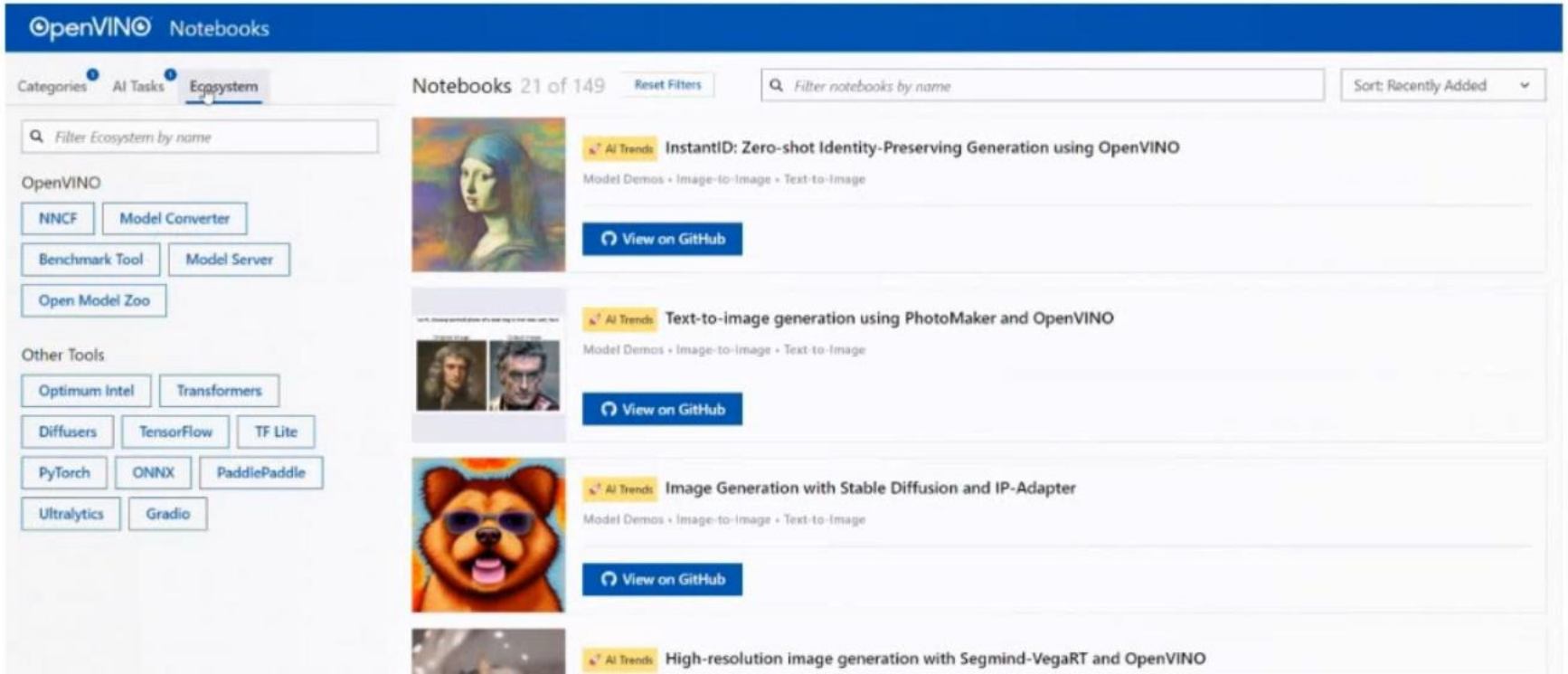
English | 简体中文

OpenVINO™ Notebooks

license Apache 2.0 | treon passing | docker_treon passing

A collection of ready-to-run Jupyter notebooks for learning and experimenting with the OpenVINO™ Toolkit. The notebooks provide an introduction to OpenVINO basics and teach developers how to leverage our API for optimized deep learning inference.

🚀 Checkout interactive GitHub pages application for navigation between OpenVINO™ Notebooks content: OpenVINO™ Notebooks at GitHub Pages



Personal AI Assistant

OpenVINO Kits

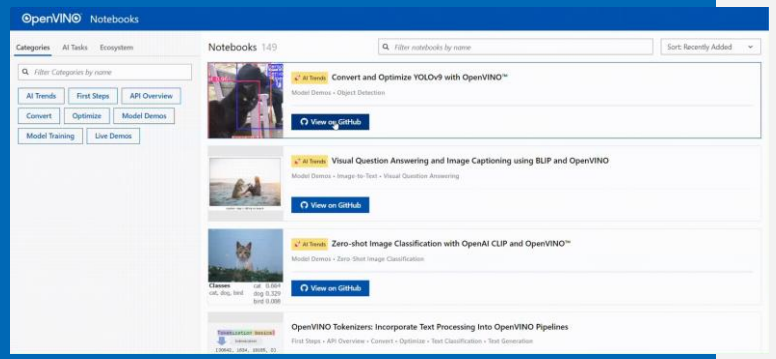
NPU

CPU

GPU

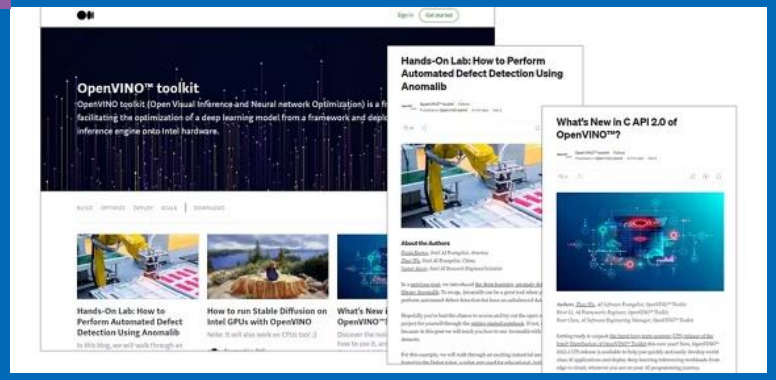
Developer Resources

Model Hub for AI Inference Benchmarks



OpenVINO™ Notebooks

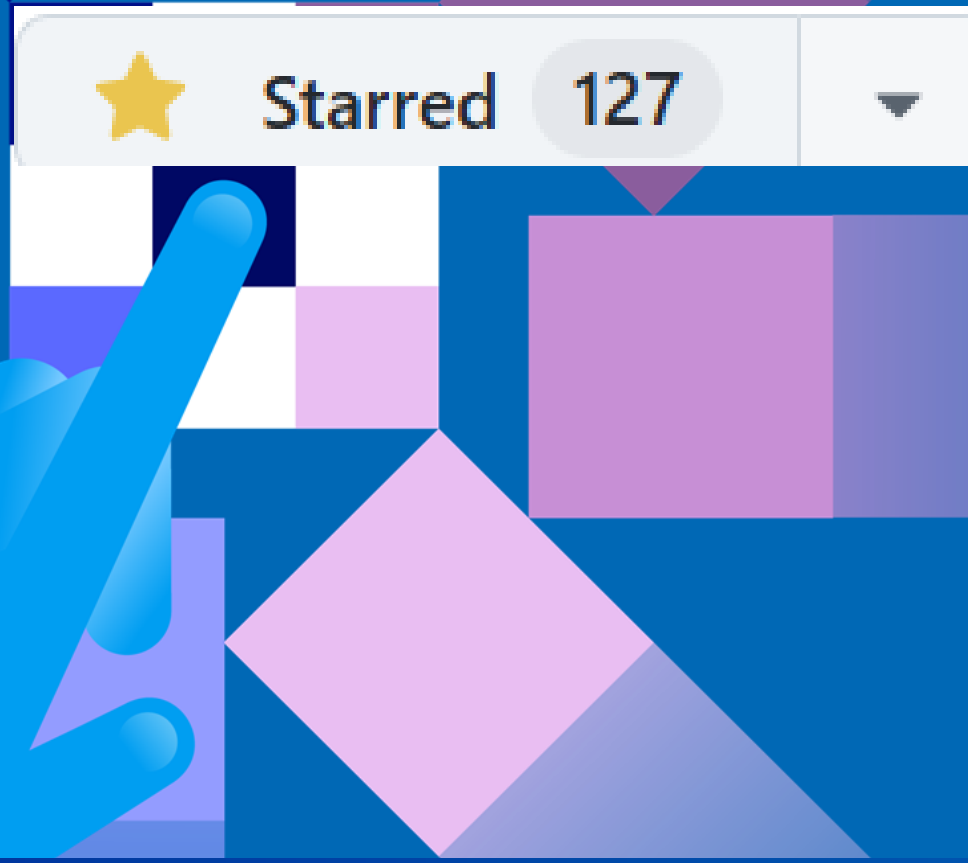
DevCon Workshop Series



Latest Blogs

Developer Clouds for Accelerated Computing

OpenVINO Kits



Thank You

Anisha Udayakumar

Notices & Disclaimers

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

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

intel®