OpenVINO Release Notes

2024.3 - 31 July 2024

System Requirements | Release policy | Installation Guides

What's new

- More Gen AI coverage and framework integrations to minimize code changes.
  - OpenVINO pre-optimized models are now available in Hugging Face making it easier for developers to get started with these models.
- Broader Large Language Model (LLM) support and more model compression techniques.
  - Significant improvement in LLM performance on Intel discrete GPUs with the addition of Multi-Head Attention (MHA) and OneDNN enhancements.
- More portability and performance to run AI at the edge, in the cloud, or locally.
  - Improved CPU performance when serving LLMs with the inclusion of vLLM and continuous batching in the OpenVINO Model Server (OVMS). vLLM is an easy-to-use open-source library that supports efficient LLM inferencing and model serving.
  - Ubuntu 24.04 is now officially supported.

OpenVINO™ Runtime

Common

- OpenVINO may now be used as a backend for vLLM, offering better CPU performance due to fully-connected layer optimization, fusing multiple fully-connected layers (MLP), U8 KV cache, and dynamic split fuse.
- Ubuntu 24.04 is now officially supported, which means OpenVINO is now validated on this system (preview support).
- The following have been improved:
  - Increasing support for models like YoloV10 or PixArt-XL-2, thanks to enabling Squeeze and Concat layers.