Accelerating Text-to-Image Diffusion Models using OpenVINO with Intel Xeon Processors

Planning / Implementation

Text-to-image diffusion models, such as Stable Diffusion XL, represent the frontier of generative AI, offering the ability to create realistic images directly from textual descriptions. These models, however, are computationally intensive and often require high-performance hardware typically found in GPU-based systems. This dependency on GPUs can present cost, availability, and power efficiency challenges—particularly at scale.

To address these limitations, this paper explores an alternative approach: accelerating text-to-image model inference on CPU-based infrastructure using Intel's OpenVINO toolkit and the AI acceleration capabilities of Intel Advanced Matrix Extensions (AMX), available in 5th and 6th Gen Intel Xeon Scalable processors.

Figure 1 illustrates sample images generated using OpenVINO with Stable Diffusion XL.

Figure 1: Example images generated using stabilityai's stable-diffusion-xl model with OpenVINO

This paper describes the following:

- Introduction to the Intel Xeon platform, explaining how AMX enhances deep learning performance
- A step-by-step walk-through of the implementation of Stable Diffusion XL using OpenVINO
- Performance benchmarking data comparing native Hugging Face inference with OpenVINO-accelerated pipelines

Intel Xeon processors and Intel AMX

At the heart of this solution are Lenovo ThinkSystem servers with Intel Xeon processors. These processors include Intel Advanced Matrix Extensions (Intel AMX), which enable direct acceleration of deep learning inference and training workloads on the CPU.

By eliminating the dependency on discrete accelerators for many AI tasks, organizations benefit from simplified infrastructure, streamlined deployment, and a reduced total cost of ownership. With backward compatibility to prior generation Intel Xeon platforms, businesses can also preserve existing investments and accelerate time to value through seamless system upgrades.

In this paper we are using the ThinkSystem SR650 V3 as the basis for our testing. The SR650 V3 is based on 5th Gen Intel Xeon processors. These processors deliver up to:

- 14× PyTorch inference performance improvement over 3rd Gen Intel Xeon
- 5× better performance per watt compared to 4th Gen AMD EPYC
- Sub-100ms latency on large language models with up to 20B parameters

It is expected that you would get even greater results using the ThinkSystem SR650 V4 with the latest Intel Xeon 6700P processors.

Diffusion workflow

This section outlines how to run an accelerated text-to-image diffusion model using OpenVINO on Intel CPUs.

The steps are as follows:

1. Create a virtual environment (Optional)

   We recommend you install the uv package to create a virtual python environment to avoid dependency issues and quickly download needed python libraries. Otherwise, if you have a clean install or a containerized machine, you can install the needed libraries directly without a virtual environment.

   Run the following in your Linux terminal:

   ```
   curl -LsSf https://astral.sh/uv/install.sh | sh
   uv venv diffusion-env
   source diffusion-env/bin/activate
   ```

2. Download dependencies

   Install the necessary OpenVINO python libraries as well as the optimum command line tool to easily obtain pre-trained models from huggingface.

   ```
   uv install optimum[openvino]
   uv pip install -U --pre --extra-index-url https://storage.openvinotoolk
   it.org/simple/wheels/nightly openvino openvino-tokenizers openvino-gena
   i
   ```

3. Download Models & Convert to OpenVINO-IR

   Using the optimum command line interface download a huggingface model of your choosing and convert it to the OpenVINO Intermediate Representation.

   Below is an example using the stable-diffusion-xl model due to its visually impressive results.

```
optimum-cli export openvino --model stabilityai/stable-diffusion-xl-bas
e-1.0 /~
/.cache/huggingface/hub/stable-diffusion-xl-base-1.0-ov
```

4. Create Inference Pipeline and Generate Image

   Begin using the model to generate a sample image using the following python script. The results should be a 1024x1024 image similar to Figure 2.

```
model_path = "/home/lenovoai/.cache/huggingface/hub/stable-diffusion-xl
-base-1.0-ov/" prompt = "A fox wearing a detective trench coat solving
a mystery"
pipe = ov_genai.Text2ImagePipeline(model_path, "CPU")
image_tensor = pipe.generate(prompt, num_inference_steps=40, rng_seed=4
2)

# Post-Process and Save Image
image = Image.fromarray(image_tensor.data[0]) image.save("output-img.pn
g")
```

Figure 2: Example output from prompt: “A fox wearing a detective trench coat solving a mystery”

OpenVINO performance

To assess the performance benefits of OpenVision and Intel Xeon processor-optimized inference, a benchmark was conducted comparing Stable Diffusion XL-1.0 executed with native Hugging Face libraries with the OpenVINO toolkit on 5th Gen Intel Xeon processors. The results are shown in Figure 3.

Stable Diffusion XL-1.0 was selected due to its computational intensity and model complexity, representing a demanding use case for generative AI workloads. The test involved generating 30 images, each using 40 inference steps.

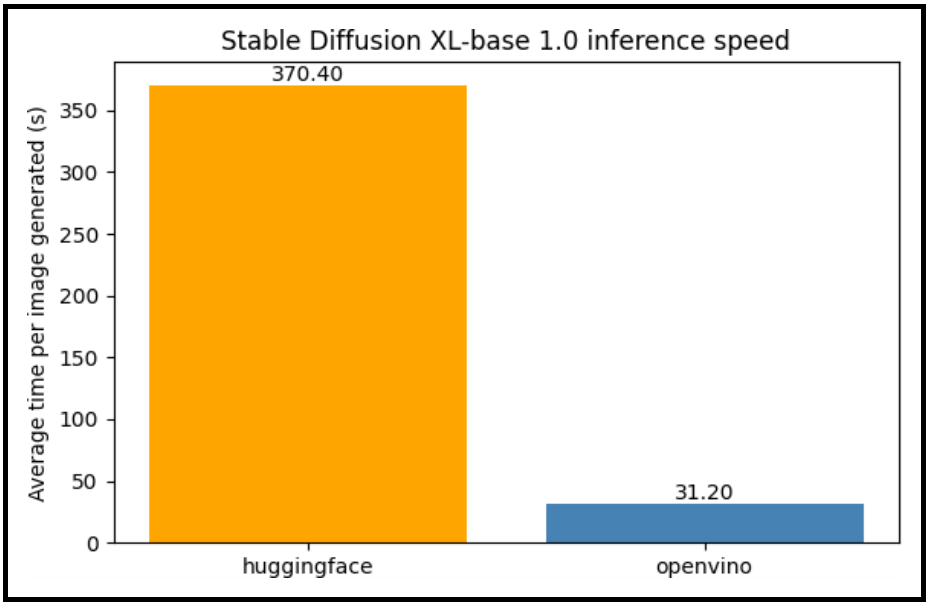

Figure 3. Inference speed comparison on Stable Diffusion XL-1.0 averaged over 30 images with 40 inference steps each; see Hardware Details for exact hardware specifications

Results demonstrate that native Hugging Face pipelines are not fully optimized to take advantage of the AI acceleration capabilities of Intel Advanced Matrix Extensions (Intel AMX) as implemented in 5th Gen Intel Xeon processors. Using the native implementation, the average time to generate a single image was approximately six minutes.

In contrast, OpenVINO significantly improves inference speed by applying graph-level optimizations and utilizing low-level hardware acceleration resulting in the average to perform the same task of thirty seconds, a 12x performance improvement. These enhancements enable more efficient execution of compute-intensive diffusion models on CPU-based infrastructure, delivering faster time-to-result and improving system throughput for AI-driven image generation workloads. This performance uplift makes CPU-based inference a viable, cost-effective alternative for organizations deploying large-scale generative AI applications.

Conclusion

This paper highlights a powerful, cost-effective solution for Generative AI: high-performance text-to-image inference on CPU-based infrastructure using Lenovo ThinkSystem servers with Intel Xeon Scalable processors and the OpenVINO toolkit. By harnessing Intel AMX acceleration in 5th Gen Intel Xeon Processors and OpenVINO's optimized model execution, organizations can achieve up to 12× faster inference over native libraries—without relying on discrete GPUs.

The result is a highly scalable, energy-efficient solution that reduces total cost of ownership, simplifies deployment, and supports sustainable AI operations. For enterprises looking to scale generative AI across production environments, Lenovo and Intel deliver a robust, ready-to-deploy platform that meets today's performance, efficiency, and operational demands.

The latest Intel Xeon 6 processors expand on the capabilities of their predecessors by offering broader support for Intel AMX, higher core counts, and improved power efficiency. Users can expect even faster inference times, greater throughput, and better energy-per-inference metrics. These enhancements further close the performance gap between CPU and GPU solutions while maintaining the flexibility and scalability of general-purpose server infrastructure. This makes Intel Xeon 6 an ideal foundation for next-generation generative AI deployments across enterprise, edge, and cloud environments.

Ready to accelerate your AI strategy? Explore how Lenovo and Intel can help your organization deploy high-performance, efficient, and reliable generative AI solutions at scale.

Hardware Details

The following table lists the key components of the server we used in our performance tests.

Table 1. Hardware Details

Component	Description
Server	Lenovo ThinkSystem SR650 V3
Processor	2x Intel Xeon Platinum 8592+ CPU @ 2.50GHz
Installed Memory	16x 32GB TruDDR5 4800MHz (2Rx8) RDIMM
Disk	1x 7450 960GB M.2 NVMe SSD, 1x 2.0TB M.2 NVMe SSD
OS	Ubuntu 22.04.5 LTS (GNU/Linux 6.8.0-60-generic x86_64)
OpenVINO	2025.2.0

Author

Eric Page is an AI Engineer at Lenovo. He has 6 years of practical experience developing Machine Learning solutions for various applications ranging from weather-forecasting to pose-estimation. He enjoys solving practical problems using data and AI/ML.

Related product families

Product families related to this document are the following:

- Artificial Intelligence
- Processors

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service. Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
8001 Development Drive
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION ”AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary. Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk. Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document, LP2261, was created or updated on July 30, 2025.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2261
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2261.

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at https://www.lenovo.com/us/en/legal/copytrade/.

The following terms are trademarks of Lenovo in the United States, other countries, or both:
Lenovo®
ThinkSystem®

The following terms are trademarks of other companies:

AMD and AMD EPYC™ are trademarks of Advanced Micro Devices, Inc.

Intel®, OpenVINO®, and Xeon® are trademarks of Intel Corporation or its subsidiaries.

Linux® is the trademark of Linus Torvalds in the U.S. and other countries.

Other company, product, or service names may be trademarks or service marks of others.