The bridge to possible

Unleashing Creativity with Generative AI

Utilizing 5th Gen and 4th Gen Intel® Xeon® Scalable Processors in Cisco® UCS® for Inferencing

The bridge to possible

Benefits

Generative AI is revolutionizing industries, enabling text-to-image generation, realistic voice synthesis, and even the creation of novel scientific materials. However, unleashing the full potential of these powerful models requires a robust and optimized infrastructure.

Generative AI models typically require massive amounts of data and complex algorithms, leading to significant computational demands during inference. Challenges include:

- High computational workloads: Inference often involves processing large amounts of data through complex neural networks, requiring high-performance computing resources.
- Memory bandwidth demands: Large models often require substantial memory bandwidth to handle data transfer efficiently.
- Latency requirements: Many applications require low latency inference to ensure real-time responsiveness.

Why Cisco UCS with 5th Gen and 4th Gen Intel Xeon Scalable Processors Excel for Generative AI

Cisco UCS®, powered by Intel® Xeon® Scalable Processors, delivers a compelling solution for overcoming these challenges and maximizing Generative AI performance. Some of the benefits are:

- Faster inference: UCS with Intel Xeon processors delivers faster inference speeds, enabling real-time applications and lower latency response times.
- Increased model throughput: The platform efficiently handles large-scale Generative AI workloads, allowing you to process more data and achieve higher throughput.
- Reduced operational costs: Lower TCO and improved energy efficiency lead to significant cost savings, making Generative AI more accessible and affordable.
- Simplified infrastructure management: Cisco Intersight® streamlines infrastructure management, freeing up valuable resources to focus on innovation and development.

Cisco UCS X-Series Modular System, and C240 and C220 rack servers support 5th Gen and 4th Gen Intel Xeon Scalable processors so that you have the option to run inferencing in the data center or at edge using either a modular or a rack form-factor.

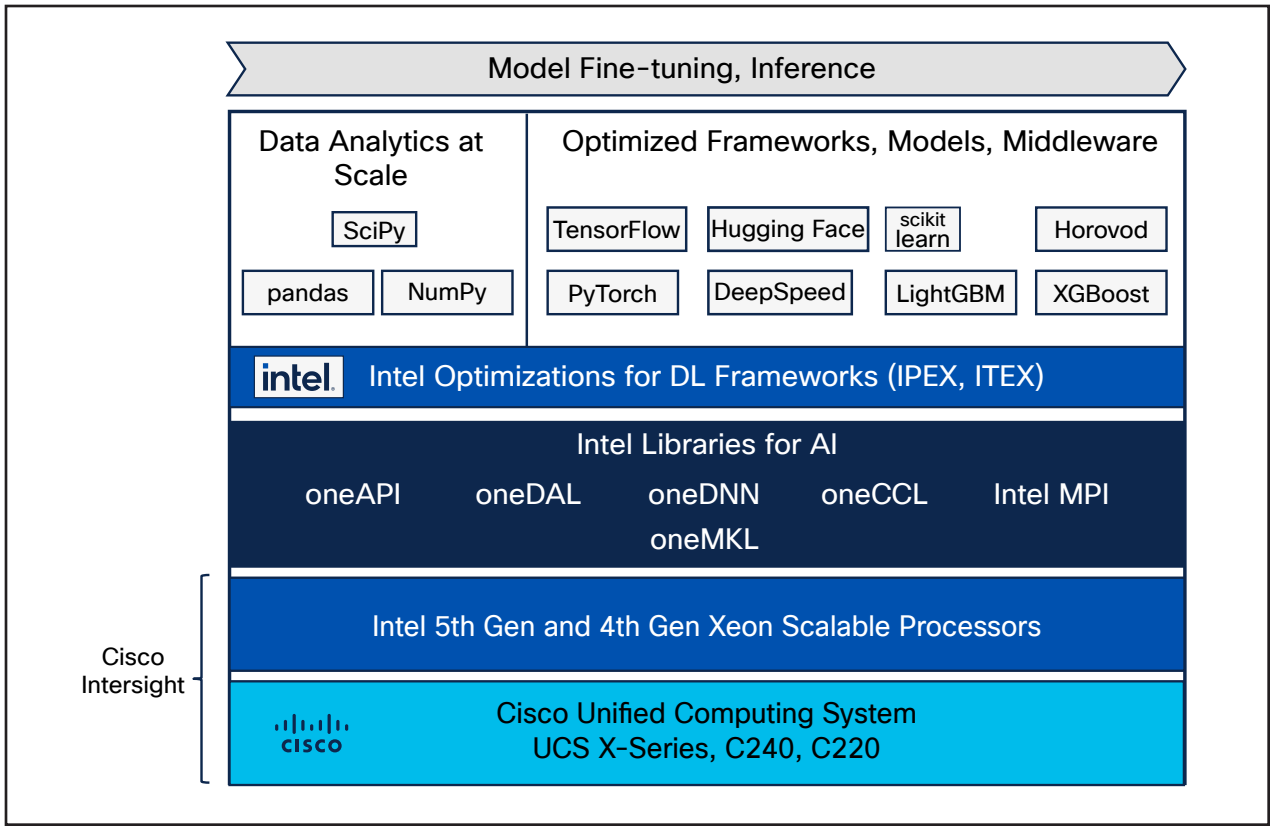


Figure 1. Reference architecture for deploying Generative AI on Cisco UCS with 5th Gen and 4th Gen Intel Xeon Scalable Processors

Inference everywhere with the leading CPU for AI

5th Gen and 4th Gen Intel Xeon Scalable processors offer several advantages for running Generative AI inferencing, including:

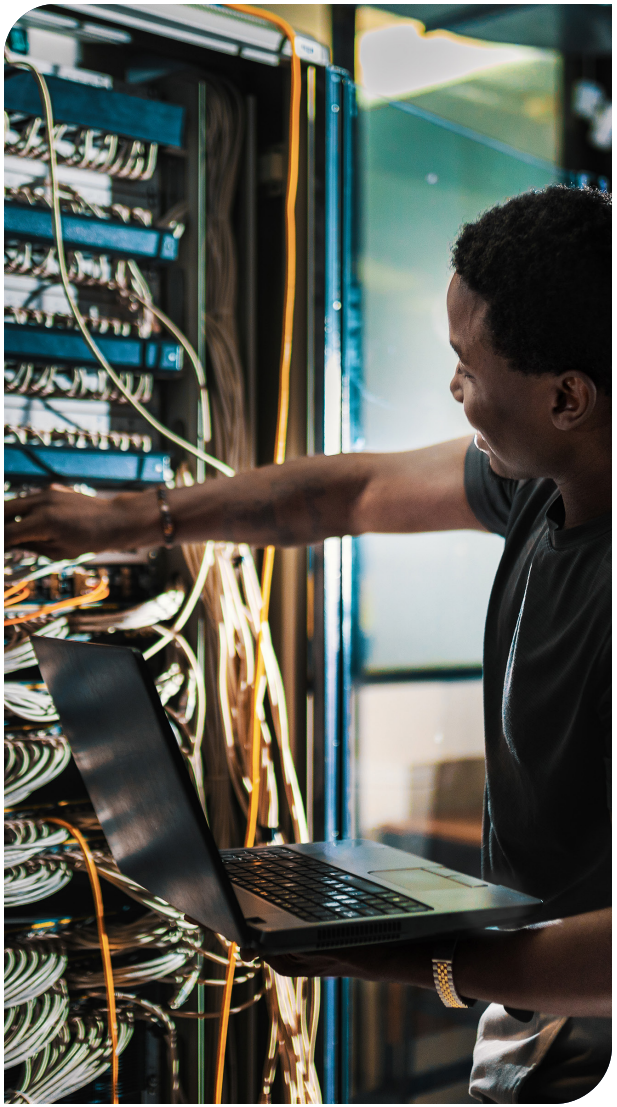
- High performance:
  - Built-in AI accelerator: Intel® Advanced Matrix Extensions (Intel AMX) is built into each core, so that computations that rely on matrix math are accelerated within the processor itself. This accelerator supports BF16 for inferencing and fine-tuning, offering a performance boost through accelerated computation, reduced memory bandwidth pressure, and half the memory consumption of FP32-based computations. It also supports quantization with support for INT8 for inferencing, which leads to significant reduction in memory requirements.
  - Larger last-level cache to help with data locality.
  - Higher core frequency and faster memory with DDR5, which is 1.5x faster than DDR4, for fast compute processing and memory access.
  - Intel® Advanced Vector Extensions 512 (Intel AVX-512) for help with non-deep learning vector computations.
- Software suite of optimized open-source frameworks and tools: Optimizations for Intel Xeon processors are already integrated into the mainstream distributions of popular deep learning frameworks, including TensorFlow and PyTorch.
- TCO benefits: Offers a balance of performance and cost of hardware that you already use for your other workloads by using built-in accelerators rather than dedicated accelerators, which results in enhanced energy efficiency, lower operational costs, and a smaller environmental footprint.

The bridge to possible

Mainstream AI infrastructure with Cisco UCS

Cisco UCS (Unified Computing System) is a comprehensive platform developed by Cisco that aims to simplify and automate the management of servers, networking, and storage resources within a data center. Cisco UCS offers several advantages as the foundation of your AI infrastructure:

- Unified management platform with 100-percent programmability: From the very beginning, Cisco UCS was designed with the entire state of each server—identity, configuration, and connectivity—abstracted into software. Cisco Intersight provides cloud-based operations to you with a single pane of glass for management of UCS infrastructure, simplifying deployment, monitoring, and troubleshooting. Cisco UCS open APIs can set server state and thus automate the integration of servers into systems.
- Automate AI deployment: Cisco UCS simplifies and automates many of the tasks involved in server provisioning, such as installing operating systems and configuring software. Cisco Validated Designs (CVDs), with detailed blueprints for deployments, can reduce deployment time of new applications by up to 60 percent and ensure that installation is done correctly and reliably, reducing the risks associated with running complex AI infrastructure. Additionally, CVDs are accompanied with automation playbooks (available in the Cisco UCS Solutions GitHub repository) that greatly simplify the deployment of the AI stack.
- Secure AI platform: Cisco UCS servers provide security features such as role-based access control, secure boot so that only digitally signed Cisco firmware images can be installed and run on the servers, and data encryption. These features can help you protect your data and applications from unauthorized access.
- Reduced TCO: Cisco UCS servers are often more cost-effective than traditional servers, especially when you consider the costs of additional hardware, software, and personnel. Cisco UCS simplifies management and reduces the need for manual intervention, which can save you time and money. Cisco UCS is a highly reliable platform, which can help you reduce downtime and improve the availability of your applications.



The bridge to possible

Learn more

For more information about Cisco UCS servers with 5th Generation Intel Xeon Scalable processors, refer to the At-A-Glances for Cisco UCS-X Modular System, Cisco UCS X-210c M7 Compute Node, Cisco UCS C240 M7 Rack Server, and Cisco UCS C220 Rack Servers.

- Cisco AI Readiness
- Cisco® AI

Efficient inferencing for Large Language Models (LLMs) on Cisco UCS with Intel Xeon Scalable Processors

Large Language Models (LLMs) are a powerful type of Generative AI technology that excel in comprehending and generating human-like text. They are essentially complex algorithms trained on massive amounts of text data, allowing them to learn the patterns and nuances of language. This enables them to perform various tasks such as text generation, answering questions, language translation, writing different kinds of creative content, and more.

Inferencing refers to the process of using a trained LLM model to generate outputs based on new, unseen input data. This involves feeding the input data to the model and obtaining the model's predicted output, such as a text continuation, a translation, or an answer to a question.

Cisco engineers tested Llama-2-7B, an LLM with seven billion parameters, with 4th Gen Intel Xeon processors using Intel's inference benchmarking scripts, which test the text generation of LLMs. IPEX (Intel's open-source Extension for PyTorch) is leveraged to optimize deep learning performance on Intel Xeon processors. Computation on the Intel Xeon processors was done using Float32, BFloat16, and INT8 data formats.

The system under test was a Cisco UCS M7 X210c compute node with 4th Gen Intel Xeon Gold 6430 processors installed in a Cisco UCS X-Series chassis. Testing was done on a single instance (that is, a single node on a single socket) using the single-node benchmark test to test and validate the results. Cisco Intersight SaaS was used to set up the service profiles on UCS in the lab. The measured latency response was well within the recommended 100ms.

Such low latency for a Generative AI workload on a X-series CPU means you can cost-effectively use Cisco UCS servers with 5th Gen and 4th Gen Intel Xeon processors for demanding AI workloads without adding discrete accelerators to the servers.

C45-4111868-00 12/23