Mainstreaming Generative AI Inference Operations with 5th Gen Intel Xeon Scalable Processors in Cisco UCS

Generative AI has arrived, promising to revolutionize industries from healthcare to entertainment with its ability to create novel text, images, and even materials. However, harnessing the full potential of these powerful models requires a robust, scalable, and cost-effective infrastructure.

Generative AI inferencing: Challenges of at scale deployments

Generative AI workloads are characterized by their massive data-processing requirements, complex algorithms, and distributed computing architectures. Deploying Generative AI at scale presents unique challenges:

- **Intensive compute needs:** Large-scale models are computationally expensive. Suboptimal compute resources can result in slower response in practical scenarios. Moreover, different stages of the inference pipeline, from pre-processing to post-processing, require diverse compute capabilities.
- **Model complexity and size:** Large Language Models (LLMs) can have billions of parameters, exceeding the memory capacity of single devices. Distributed inferencing across multiple machines can introduce complexities in model partitioning and require model optimization techniques.
- **Massive network demands:** Real-time responsiveness is often crucial for AI applications, making low latency critical. Large scale models, distributed across servers, can generate high volumes of traffic between servers. Any degradation in performance will affect the Job Completion Time (JCT).
- **Infrastructure complexity:** Managing and orchestrating large-scale AI deployment requires robust infrastructure and intelligent automation.

Comprehensive scalable solution: Cisco UCS with 5th Gen Intel Xeon Scalable Processors and Cisco Nexus

A solution based on Cisco UCS® with Intel® Xeon® Scalable Processors and Cisco Nexus® offers a compelling and scalable foundation for deploying Generative AI at scale. This architecture offers a combination of:

- **Optimal performance:** Cisco UCS with Intel Xeon Scalable processors with specialized AI accelerators and optimized software frameworks significantly improves inferencing performance and scalability. Cisco Nexus 9000 switches provide high bandwidth, low latency, congestion management mechanisms, and telemetry to meet the demanding networking requirements of AI/ML applications.
- **Balanced architecture:** Cisco UCS excels in both Deep Learning and non-Deep Learning compute, critical for the entire inference pipeline. This balanced approach leads to better overall performance and resource utilization.
- **Scalability on demand:** Cisco UCS seamlessly scales with your Generative AI inferencing needs. Add or remove servers, adjust memory capacities, and configure resources in an automated manner as your models evolve and workloads grow using Cisco Intersight®.

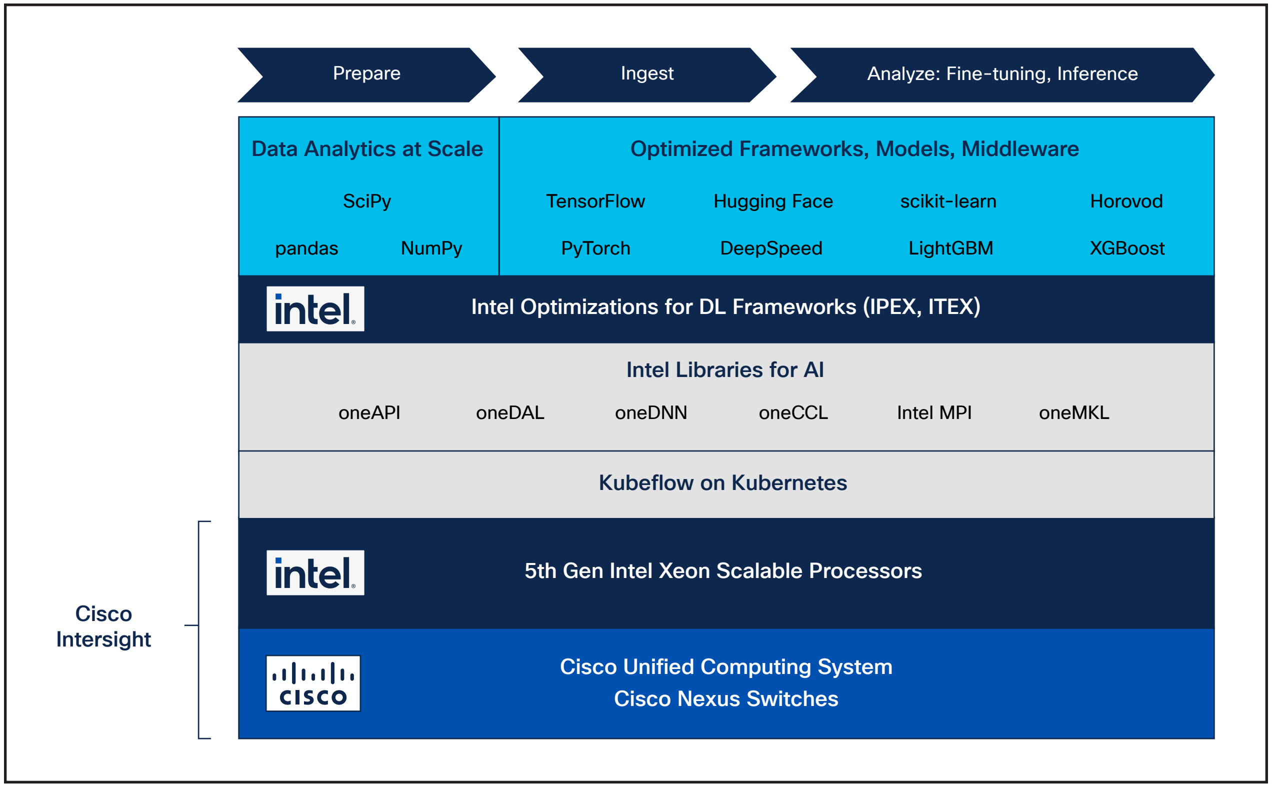
The Cisco UCS X-Series Modular System, and C240 and C220 rack servers, support 5th Gen Intel Xeon Scalable processors so that you have the option to run inferencing in the data center or at the edge, using either a modular or a rack form factor.

Deploying Kubernetes on top of a hardware infrastructure stack allows scale up or scale down of your inference workloads based on demand. This is crucial for handling fluctuating traffic, avoiding over-provisioning, and optimizing resource utilization. Kubeflow abstracts the components of Kubernetes and hides the complexity of containerization to expose a usable platform for developers and engineers.

Scalable Generative AI inferencing on 5th Gen Intel Xeon Scalable Processors

DeepSpeed integration with 5th Gen Intel Xeon Scalable processors offers the following benefits:

- **High performance:**
  - Built-in AI accelerator: Intel® Advanced Matrix Extensions (Intel AMX) accelerator is built into each core to significantly speed up deep-learning applications when 8-bit integer (INT8) or 16-bit float (bfloat16) data types are used.
  - Higher core frequency, larger last-level cache, and faster memory with DDR5 speed-up compute processing and memory access.
  - Intel® Advanced Vector Extensions 512 (Intel AVX-512) for help with non-deep learning vector computations.
- **Optimized DeepSpeed inferencing:** DeepSpeed provides high-performance inference support for large Transformer-based models with billions of parameters, through enablement of multi-CPU inferencing. It automatically partitions models across the specified number of CPUs and inserts necessary communications to run multi-CPU inferencing for the model. Intel has integrated optimizations into both DeepSpeed and Intel Extensions for PyTorch (IPEX), enabling users to seamlessly use the DeepSpeed trained models in the following manner:
  - Run on 5th Gen Intel Scalable processors without any modification.
  - Run across multiple cores.
  - Fully utilize CPU cores and reduce memory footprint as some transformer stages are fused to run together, and
  - Be bound to cores to reduce interference.



Prepare

Ingest

Analyze: Fine-tuning, Inference

Data Analytics at Scale

SciPy

pandas NumPy

Optimized Frameworks, Models, Middleware

TensorFlow Hugging Face scikit-learn Horovod

PyTorch DeepSpeed LightGBM XGBoost

intel Intel Optimizations for DL Frameworks (IPEX, ITEX)

Intel Libraries for AI

oneAPI oneDAL oneDNN oneCCL Intel MPI oneMKL

Kubeflow on Kubernetes

intel 5th Gen Intel Xeon Scalable Processors

Cisco Unified Computing System

Cisco Nexus Switches

Cisco Intersight

Figure 1. Reference architecture for deploying Generative AI on Cisco UCS with 5th Gen Intel Xeon processors

- **Cost-effectiveness:** DeepSpeed on 5th Gen Intel Xeon Scalable processors offers lower TCO by enabling use of built-in accelerators to scale-out inferencing performance rather than relying on discrete accelerators, making Generative AI more accessible and affordable.

High-performance Ethernet fabrics at scale for AI/ML

Cisco Nexus 9000 Switches maximize JCT performance by ensuring high-bandwidth and lossless communication, so that compute resources are not sitting idle waiting for networking resources. Cisco Data Center Networking Blueprint for AI/ML applications is built on the following capabilities:

- **State-of-the-art scalable platforms:** Cisco Nexus 9000 fixed and modular switches offer a choice of highly dense 100G/400G/800G interface speeds for one IP/Ethernet network vs. dedicated front-end/ back-end networks. Intelligent buffering and rich telemetry/ visibility are available.
- **Lossless networking:** Explicit Congestion Notification (ECN), Priority Flow Control (PFC), and Approximate Fair Drop (AFD) algorithms are used for dynamic congestion management to ensure lossless networking.
- **High-performance transport layer:** Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE v2) is used to provide high throughput, low-latency transport over IP/Ethernet networks.

Benefits of deploying Generative AI inferencing on Cisco UCS

The Cisco Unified Computing System™ (Cisco UCS) is a next-generation data-center platform that unites computing, networking, storage access, and virtualization resources into a cohesive system designed to reduce Total Cost of Ownership (TCO) and increase business agility. Some of the benefits are:

- **Faster time-to-value:** Fully-software-defined compute platform to rapidly deploy and scale Generative AI models, accelerating your journey to innovation. Cisco Intersight, a cloud-based infrastructure management platform, can manage your entire Cisco UCS infrastructure, including servers, storage, networking, and virtual machines, from a single pane of glass. It can automate provisioning, configuration, and patching to streamline your operations with consistent policies.
- **High-performance connectivity:** Connect up to 200 Gbps of line-rate bandwidth per server in both modular and fixed form factors. Traffic between any two blade servers or racks requires only one network hop, thereby enabling low-latency and consistency.
- **Reduced operational costs:** streamlined management, infrastructure automation, efficient resource utilization, greater energy efficiency, and advanced power management translate to significant cost savings. Cisco UCS X-Series earned the 2023 SEAL Sustainable Product Award, which honors products that are “purpose-built” for a sustainable future.

High-performance, elastic inferencing for Large Language Models (LLMs) using Kubernetes and DeepSpeed across multiple 5th Gen Intel Xeon Scalable Processors

Large Language Models (LLMs) are sophisticated AI models designed to understand, generate, and process human language at an advanced level. These models are characterized by their enormous size, with millions or even billions of parameters, enabling them to capture complex linguistic patterns and relationships. Inferencing with LLMs involves utilizing pre-trained models to process and generate language-based outputs based on new inputs or prompts.

Inferencing with LLMs typically requires significant infrastructure resources due to their vast number of parameters. As shown in Figure 2, three different deployment models can be considered, depending on the size and location of the inferencing deployment. You can use DeepSpeed integration with 5th Gen Intel Xeon processors to speed up inferencing and use Kubeflow to run inferencing on Kubernetes regardless of the deployment model.

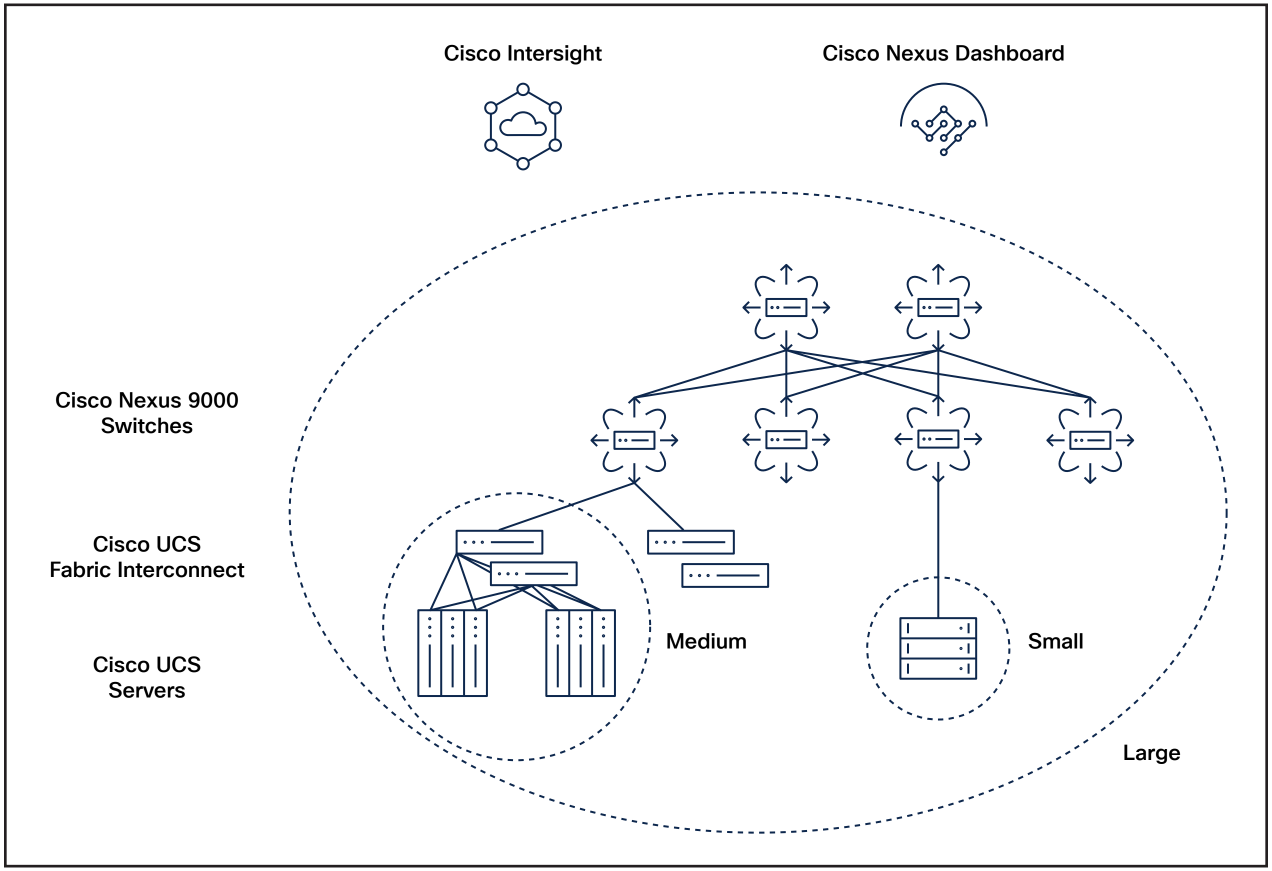


Figure 2. One operational model for small, medium, and large deployment models

Small deployment model: Inferencing is deployed on Cisco UCS rack servers at the edge that are directly attached to Intersight through 10G/25G/100G connections.

Medium deployment model: Inferencing is deployed in the data center on a small cluster of servers that are in the same Cisco UCS X-Series Modular Systems or on UCS X-Series Modular Systems that are connected to the same UCS fabric interconnect. All the inferencing traffic is confined to a UCS fabric interconnect. Each server can be statically pinned to the ingress and egress downlink ports of the same fabric interconnect to make sure that there is no more than a single hop between the servers to maintain the lowest possible latency. Lossless class of service is utilized.

Large deployment model: A very large cluster of UCS servers can be deployed using the Cisco Data Center Networking Blueprint for AI/ML Applications. The Nexus 9000 switches are configured in a spine-leaf architecture to keep the number of hops to a minimum. Congestion management is enabled through Quality of Service (QoS) to maintain lossless operation. RoCEv2 transport layer is used to transfer data between servers at the memory-memory level, without burdening the CPU, to achieve low-latency throughput.

Cisco UCS with Intel Xeon Scalable Processors and Cisco Nexus together provide a combination of unmatched performance, intelligent management, robust networking, and optimized DeepSpeed integration to empower organizations to unlock the transformative potential of this revolutionary technology.

Learn more

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

- Cisco Data Center Networking Blueprint for AI/ML Applications