Power a Spectrum of AI Applications

Cisco UCS X215c M8 Compute Node and AMD EPYC processors

Artificial Intelligence (AI) is appearing in many new contexts, placing a spectrum of new demands on enterprise computing—whether in the core data center, branch, remote, and industrial sites, or at the network edge:

- Companies are integrating AI features, such as recommendation engines, into customer-facing web sites and applications. AI inference using these small- to medium-sized models can often be accomplished with powerful server CPUs.
- Enterprise application vendors are integrating AI features into traditional applications such as enterprise resource planning and database management systems. These AI features can be supported with a modest boost from GPU accelerators.
- Organizations training and fine tuning models, or deploying generative AI solutions to interact with customers, need multiple high-performance GPUs to optimize performance.

In this rapidly changing environment, it's hard to make infrastructure choices when you don't know what you will need in a month or a year.

A flexible, modular system designed to take you into the future

When you need infrastructure that can adapt as fast as the new demands your software places, the Cisco UCS® X-Series Modular System offers and ideal perfect solution. With eight slots that can accommodate compute nodes with configurable combinations of disk storage and GPU acceleration, and a PCIe node that can accommodate a number of full-sized GPUs, you can easily configure servers to meet your demands today and well into the future.

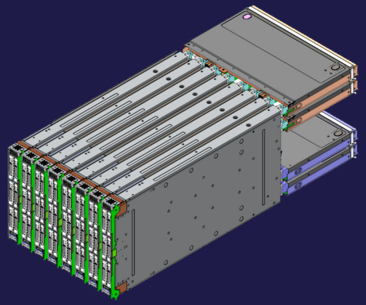
Highlights

- Address a spectrum of AI infrastructure needs in a single flexible, modular chassis
- Address today's needs and be ready to adapt to future requirements
- Deploy either in the core data center or as self-contained pods in remote and branch offices and industrial locations
- Propel your AI infrastructure with the only modular system powered by AMD EPYC™ processors: the Cisco UCS X215c M8 Compute Node

It doesn’t get simpler than this

The Cisco UCS X-Series chassis houses technology you need today with an approach that embraces the future.

Without an I/O midplane, the chassis can adapt to new networking and I/O technology as it emerges. Vertically-oriented compute nodes intersect with horizontally oriented I/O components with blind-mating connectors.



Today, each Cisco UCS X-Series compute node powered by AMD EPYC processors connects to up to 200 Gbps of network bandwidth and 32 lanes of PCIe Gen 4 connectivity to modular devices such as GPU accelerators.

The system is engineered around a chassis without an I/O midplane, enabling server, I/O, and networking technologies to evolve without having to overhaul your infrastructure.

Cloud-operating model

All server state is abstracted from the hardware, so you can change and adapt configurations through the Cisco Intersight® IT operations platform. Not only can you update the server configuration as your needs change, you also can deploy servers consistently wherever they reside by capturing your IT best practices in templates that are used for consistent configuration directions.

Deploy where you need AI

The Cisco UCS X-Series is at home everywhere you need to support AI applications. In the data center, you can deploy multiple chassis with a set of Cisco UCS fabric interconnects that act as a single point of connectivity and management for the system. If you prefer to deploy individual AI pods in the data center or in branch and remote offices, the Cisco UCS X-Series Direct option enables you to configure self-contained chassis with connectivity and management built in (Figure 1).

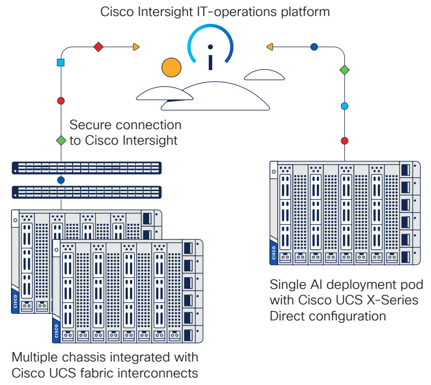

Cisco Intersight IT-operations platform
Secure connection to Cisco Intersight
Single AI deployment pod with Cisco UCS X-Series Direct configuration
Multiple chassis integrated with Cisco UCS fabric interconnects

Figure 1. Deploy your AI infrastructure with multiple chassis integrated with Cisco UCS fabric interconnects or with a Cisco UCS X-Series Direct configuration

Incorporate sustainability into AI infrastructure

Sustainability is important, especially in the face of AI initiatives demanding more from the power grid. With shared power supplies, zoned cooling, and high-voltage power distribution within the chassis, the Cisco UCS X-Series can help reduce resource consumption, one of the reasons why it has been recognized as SEAL Sustainable Product of the Year for 2023. In addition to its power efficiency, the product is designed for longevity and reuse, resulting in roughly 50 percent less raw material consumed over the course of three years compared to using traditional rack servers. This flexibility and reusability earned the Cisco UCS X-Series the Silverlining Best Cloud Sustainability Initiative award for 2023.

Cisco UCS X215c M8 Compute Node

The only modular compute node built with 4th and 5th Gen AMD EPYC processors, the Cisco UCS X215c M8 is key to high-performance, flexible AI infrastructure that can be deployed anywhere.

- Up to two AMD EPYC 9004 Series processors with up to 128 cores per socket; or up to two AMD EPYC 9005 Series processors with up to 160 cores per socket
- 24 DIMM slots for up to 6 TB of memory
- Up to 200 Gbps Cisco® unified fabric network connectivity and 32 PCIe Gen4 lanes connecting to the Cisco UCS X-Fabric
- Configure with up to 6 front-facing small-form-factor (SFF) SAS, SATA, or NVMe drives with optional RAID controller or up to 2 front-facing drives and two half-height GPU accelerators
- Support for Cisco UCS Virtual Interface Card 15000 Series
- Internal dual M.2 drive options

Use the only modular system powered by AMD EPYC processors

The Cisco UCS X-Series is the only modular system offering a node powered by 4th and 5th Gen AMD EPYC processors: the Cisco UCS X215c M8 Compute Node. They are equipped to handle AMD EPYC 9004 Series processors with up to 128 cores per socket, and AMD EPYC 9005 Series processors with up to 160 cores per socket. These CPUs are ideal for use in artificial intelligence. AI performance in the AMD EPYC 9005 Series has been enhanced by doubling most data paths in the processor's 'Zen 5' core to 512 bits and adding more integer arithmetic-logic units (ALUs) to process data through a wider pipeline than in prior-generation CPUs. In addition, the 64-core AMD EPYC 9575F CPU is designed to enhance AI performance with a frequency-optimized CPU that accelerates CPU-based inference operations in addition to speeding data to GPUs when needed.

Cisco UCS X215c M8 Compute Node

This compute node unleashes the power of AMD EPYC AI-optimized CPUs with up to two AMD EPYC 9004 or 9005 Series processors up to 400W. The node supports up to 24 DIMMs of DDR5-6000 memory for up to 6 TB of capacity. The node comes equipped with a front mezzanine that hosts up to six SAS/SATA/NVMe drives, or one that supports up to two drives and two half-height GPUs. These configuration options enable you to power your AI initiatives with exactly the acceleration you need, regardless of where they fall on the spectrum.

Power the spectrum of AI infrastructure

Propelling your AI initiatives with the Cisco UCS X215c M8 Compute Node can help you deliver superior performance:

- **CPU-powered inference:** a broad class of AI workloads can execute small- and medium-size models with CPU-based inference (Figure 2). These include some large-language models, classical image detection, decision trees, and recommendation engines.

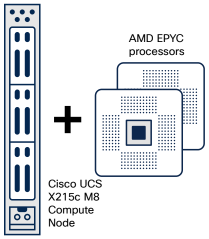
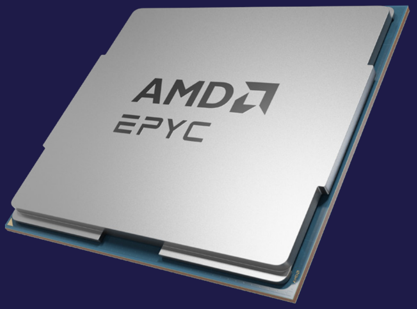


Figure 2. Support a large set of workloads with CPU-based inference

AMD EPYC processors



When you choose AMD EPYC processors to power your virtual desktop infrastructure, you gain such benefits as:

- **Compute density**, with up to 192 cores per processor (160 cores available in the Cisco UCS X215c M8 Compute Node), delivering leading performance while contributing to space, power, and cooling reductions
- **AMD Infinity Guard** features that promote security in virtualized environments with virtual machines encrypted with keys only the CPU knows
- **High-frequency options** that are optimized for AI operations with up to 64 frequency-optimized cores
- **Large cache sizes** (up to 768 MB L3 cache in the AMD EPYC 9004 Series) to help propel AI applications with large memory footprints

AMD's white paper AI Inferencing with AMD EPYC Processors demonstrates superior performance from the 96-core AMD EPYC 9654 CPU in inference operations including classification on random decision forests,[SP5-184A] Multi-Gate Mixture-of-Experts (MMoE)[SP5-183A] that helps predict customer behavior, and clustering dense vectors.[SP5-185A] Not only can you perform AI inference on CPU-only infrastructure, you can do so efficiently. AMD EPYC processors power the most energy efficient servers, delivering exceptional performance and helping reduce energy costs.[EPYC-028D]

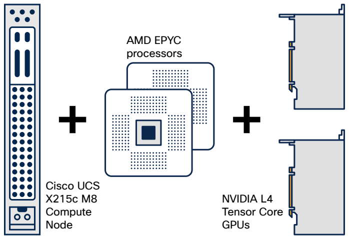


Figure 3. Boost performance of AI-enabled applications with node-resident GPU accelerators

- **AI-enabled enterprise applications:** these often need the kind of boost that can be accommodated within the Cisco UCS X215c M8 Compute Node, which supports up to two half-height GPUs, such as the NVIDIA L4 Tensor Core GPU, which has been certified for use in the Cisco UCS X-Series Modular System (Figure 3).

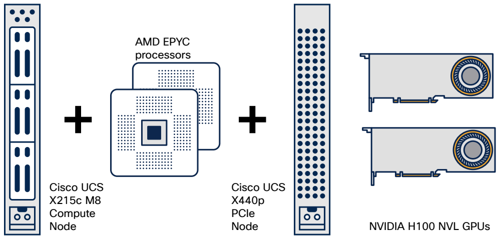


Figure 4. Undertake challenging workloads including generative AI and machine-learning training with PCIe nodes equipped with fast GPU accelerators

- **AI training and heavyweight inferencing:** the capabilities of the Cisco UCS X215c M8 can be extended through the Cisco UCS X-Fabric, named 'X' because it is a variable that can change over time as new connection capabilities evolve (Figure 4). Today, the Cisco UCS X-Fabric can connect up to four nodes in the system, each to a Cisco UCS X440p PICe Node. Each PCIe node accommodates up to four half-height cards such as the NVIDIA L4, Tensor Core GPU, or two full-height, full-width cards such as the NVIDIA A16, L40S, or H100 NVL, all of which are certified for the platform. The importance of using the best CPUs does not diminish for workloads mostly executed on the GPU accelerators. The CPU plays a crucial role in preparing data, moving it to GPU memory, and performing any post-processing tasks. With the performance of AMD EPYC CPUs, it should be no surprise that the processors demonstrate superior performance in this context. Servers tested

Financing to help you achieve your objectives

Cisco Capital® can help you acquire the technology you need to achieve your objectives and stay competitive. We can help you reduce CapEx, accelerate your growth, and optimize your investment dollars and ROI. Cisco Capital financing gives you flexibility in acquiring hardware, software, services, and complementary third-party equipment. And there's just one predictable payment. Cisco Capital is available in more than 100 countries. Learn more.

with eight NVIDIA GPUs and the AI-optimized AMD EPYC 9575F turn around inference and training tasks more quickly. [9xx5-014, 9xx5-015]

Why Cisco?

The advantage of using the Cisco UCS X-Series Modular System, the Cisco UCS X215c M8 Compute Node, and AMD EPYC processors is that you can support the AI infrastructure needs you have today, and easily adapt to supporting different requirements in the future. Only Cisco offers a modular approach using AMD EPYC processors, You even can support multiple points along the infrastructure spectrum in the same chassis. For example, Figure 5 illustrates two nodes each for CPU-based inference, internal GPU boost, and full GPU acceleration.

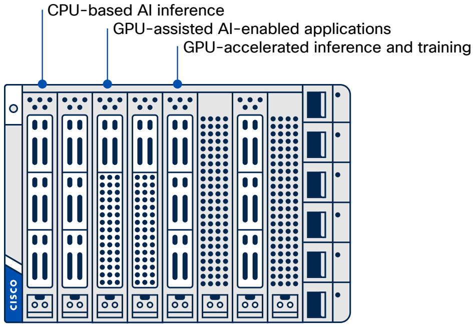

CPU-based AI inference
GPU-assisted AI-enabled applications
GPU-accelerated inference and training
cisco

Figure 5. Configure infrastructure for multiple AI applications in a single chassis: two nodes for CPU-based AI inference; two nodes for GPU-assisted AI-enabled applications; and two nodes for GPU-accelerated inference and training

With Cisco, you have the benefit of networking you can trust, and an internal I/O system based on the Cisco UCS X-Fabric that is ready to grow into the future with you. Whether you assemble multiple chassis into a single system with Cisco UCS fabric interconnects, or use a single chassis with built-in interconnects as an AI deployment pod, each Cisco UCS X215c M8 Compute Node provides up to 200 Gbps of unified fabric bandwidth to speed AI and machine learning (ML) data between GPUs and networked or Fibre Channel storage.

With a flexible, modular AI deployment solution for both your core data center, remote, branch, and industrial locations, and at the network edge, there's no better choice for a consistent deployment model that is ready for you now and into the future.