Accelerate Deep Learning

With the Cisco UCS C480 ML M5 Rack Server designed for AI and ML workloads

Artificial intelligence in the data center at scale

Cisco® machine learning computing solutions ease the challenges faced by IT organizations and data scientists: supporting the needs of machine learning (ML) workloads while making them part of the enterprise data center. With Cisco solutions, you can power artificial intelligence (AI) workloads at scale and help extract more intelligence out of data to make better decisions.

A no-compromise approach for deep learning workloads

With the addition of the Cisco UCS® C480 ML M5 for machine learning, we now offer a complete array of computing options sized to each element of the AI lifecycle: data collection and analysis near the edge, data preparation and training in the data center core, and real-time inference at the heart of AI. Our cloud-based management makes it easy to extend accelerated computing to the right locations across an increasingly distributed IT landscape.

- **Gain performance and capacity**: Cisco UCS C480 ML M5 offers flexible options for CPU, memory, networking, and storage while providing outstanding GPU acceleration
- **Demystify your machine-learning** software ecosystem with validated solutions
- **Simplify operations** with the Cisco Intersight™ platform to extend accelerated computing to the locations where it is needed

Benefits

- **Accelerate insights and decisions:** high-performance systems support constantly changing data-intensive workloads.
- **Demystify ML stacks:** proven solutions deliver a faster and more reliable deployment.
- **Reduce cost and complexity:** cloud-based management enables consistent and unified operations across your entire computing landscape.

AI at work

- **Retail:** predict shopping patterns and optimize supply chains, help prevent loss, and implement “frictionless” commerce by eliminating checkout stands.
- **Healthcare:** quickly screen or triage radiology, pathology, and dermatology images; predict patient outcomes; and help guide treatment.
- **Smart cities:** process video faster to recognize faces, license plates, and suspicious objects; observe traffic patterns of cars, bicycles, and pedestrians; and watch for intrusions into secure areas.

We support your entire AI and ML lifecycle

When IDC sees a 30 percent growth rate in an industry, it’s a good indication that things will be changing soon in your IT organization. Artificial intelligence, machine learning, and deep learning applications are helping businesses and governmental agencies to make faster, more informed decisions.

However, traditional data center technology is not designed to handle the data volume, velocity, and variability of AI at production scale. This requires a very different model than traditional business applications. The massive amount of data required and its ingestion speed are fundamentally changing how applications behave. Applications are now being shaped by data and require high-performance systems that can adapt to these new workloads. This leads your IT teams struggling to keep up. They are trying to keep up with data scientists who are constantly changing data sources, and software stacks that have changing infrastructure requirements. At the same time, the data scientists are struggling to turn machine learning into a competitive business tool.

Artificial intelligence projects have a lifecycle that begins with acquiring and preparing data, then developing and testing machine learning software. Once ready, the software is trained with massive amounts of data and is then used to make inferences that help guide decision making (Figure 1).

Power deep learning workloads

The Cisco UCS C480 ML M5 Rack Server is designed for the most compute-intensive phase of the AI and ML lifecycle: deep learning. This server integrates GPUs and high-speed interconnect technology combined with large storage capacity and up to 100-Gbps network connectivity.

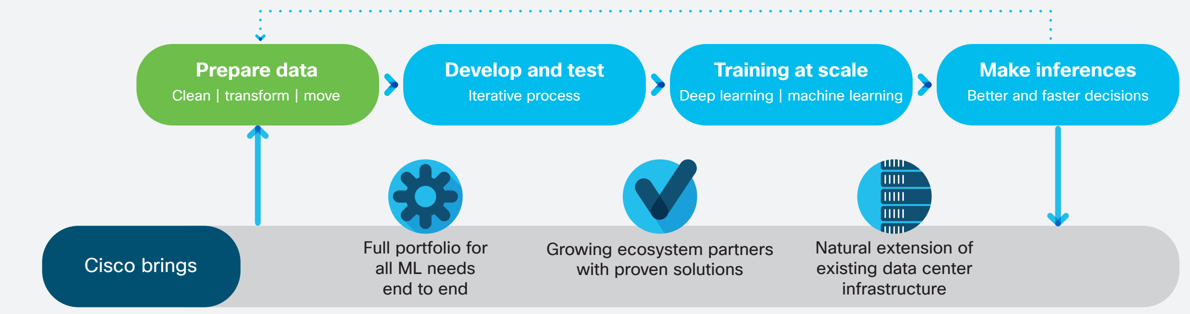
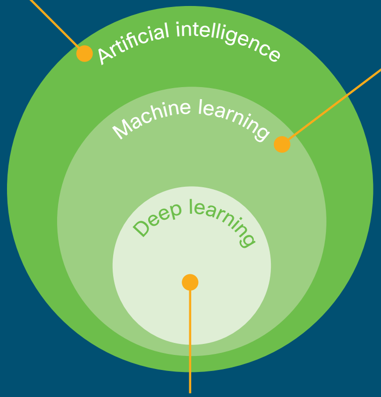

Prepare data
Clean | transform | move
Develop and test
Iterative process
Training at scale
Deep learning | machine learning
Make inferences
Better and faster decisions
Cisco brings
Full portfolio for all ML needs end to end
Growing ecosystem partners with proven solutions
Natural extension of existing data center infrastructure

Figure 1 Our complete portfolio of AI and ML specialized servers helps you efficiently support data scientists

Evolve to deep learning

Perform basic chores faster than a human: for example, classify images or recognize speech

Use AI techniques to parse data, learn from it, and make decisions: for example, detect spam



Engage neural networks to sort through vast amounts of data and make distinctions: for example: identify cancer in a medical image

The Cisco UCS C480 ML M5 features (Figure 2):

- **GPU acceleration:** eight NVIDIA V100 SXM2 32-GB modules are interconnected with NVIDIA NVLink for fast communication across GPUs to accelerate computing. NVIDIA specifies TensorFlow performance of up to 125 teraFLOPs per module for a total of up to one petaFLOP of processing capability per server.
- **The latest Intel® Xeon® Scalable CPUs:** two CPUs with up to 28 cores each manage the machine learning process and dispatch calculations to the GPUs.
- **Storage capacity and performance:** data locality can be important for deep learning applications, and up to 24 hard disk drives or SSDs store data close to where it is used and are accessed through a midplane-resident RAID controller. Up to six-disk-drive slots can be used for NVMe drives, offering best-in-class performance.
- **Up to 3 TB of main memory:** with fast 2666-MHz DDR4 DIMMs.
- **High-speed networking:** two built-in 10 Gigabit Ethernet interfaces speed the flow of data to and from the server.
- **PCIe expandability:** 4 PCIe switches feed four x16 PCIe slots for high-performance networking. Options include Cisco UCS virtual interface cards (VICs) and third-party NICs for up to 100 Gbps connectivity.
- **Unified management:** by expanding our Cisco UCS portfolio with this new Cisco UCS C480 ML M5 server, we continue to support any workload without adding management complexity.

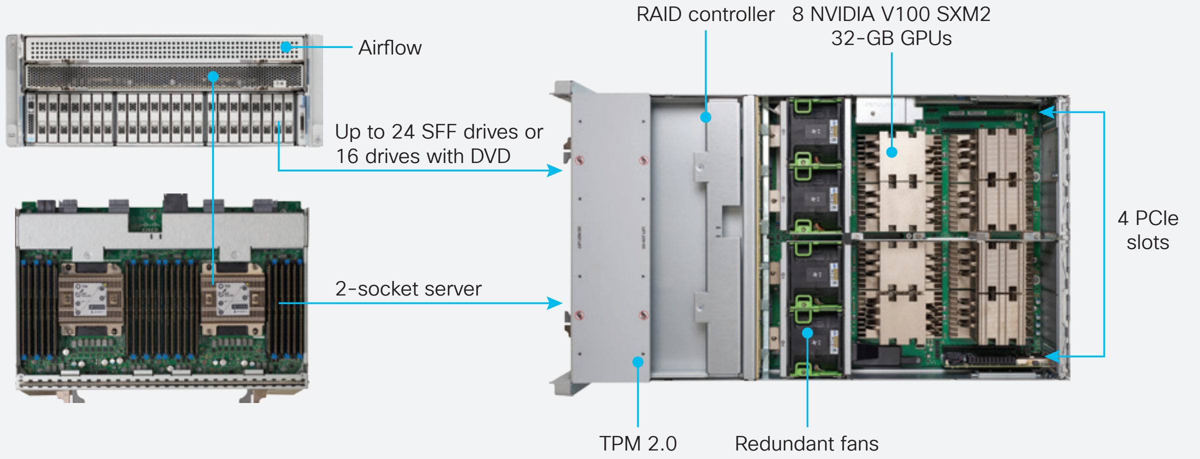

Airflow
Up to 24 SFF drives or 16 drives with DVD
2-socket server
RAID controller
8 NVIDIA V100 SXM2 32-GB GPUs
4 PCIe slots
TPM 2.0
Redundant fans

Figure 2 Cisco UCS C480 ML M5 Rack Server physical design

For more information

- cisco.com/go/ai-compute
- Cisco UCS C480 ML M5 data sheet

Why Cisco

Power the full AI and ML data lifecycle

Cisco has experience helping customers integrate changing data sources as part of a dynamic data pipeline. We can help you extend your big data environment to AI and ML by integrating GPUs into Cisco Unified Computing System™ (Cisco UCS) and Cisco HyperFlex™ systems, so you can capitalize on the adaptability and programmability of the Cisco UCS platform to power AI workloads at scale.

Eliminate silos

With Cisco Intersight, we can make it easy to deploy new technologies anywhere, eliminating islands of standalone servers regardless of where they are located, in the data center, multisite remote and branch offices, or at the edge.

Demystify ML stacks

We invest in testing and validating AI and ML solutions so that you can deploy our servers with confidence. Cisco Validated Designs provide practices for all aspects of solution deployment. With Cisco engineering doing the deployment validation, you can deploy solutions faster and with less risk.

AI helps you learn from your data and make better, faster decisions. We offer a portfolio of computing solutions that addresses all the stages of the AI lifecycle. Cisco Services, with certified partners, provide the right mix of analytics, deep-learning, and automation capabilities to deliver faster data center transformation.