Learn More with FlexPod Datacenter for AI

You want to use AI and ML to increase revenue and efficiency. The FlexPod® Datacenter for AI solution with Cisco UCS® C480 ML M5 and Cisco UCS C-Series M5 servers stands ready to train your models for faster insight.

The promise of artificial intelligence

Deep learning (DL) and machine learning (ML) techniques offer the potential for unparalleled access to accelerated insights. With the capability to learn from data and make more informed, faster decisions, your organization is better positioned to deliver innovative products and services in an increasingly competitive marketplace. Whether you need to make discoveries, analyze patterns, detect fraud, improve customer relationships, optimize supply chains, or automate processes, DL and ML techniques can help you use digital information for business advantage.

Traditional IT infrastructure falls short

The evolution of artificial intelligence (AI) continues to push the bounds of traditional IT infrastructure (Figure 1). Increasingly complex tasks require unprecedented levels of computing power and large amounts of scalable storage. An explosion of data results in deep learning models that take days or weeks to train. Computing nodes, storage systems, and networks often are unable to handle the data volume, velocity, and variability of AI, ML, and DL applications at scale. These are not the only challenges you may face.

- **Do-it-yourself integrations are complex.** Assembling and integrating off-the-shelf hardware and software components increases complexity and lengthens deployment times. As a result, valuable data science resources are wasted on systems integration work that often results in islands of IT resources that are difficult to manage and require deep expertise to optimize and control.
- **Achieving predictable and scalable performance is hard.** Scaling with traditional solutions can lead to downtime. These disruptions not only reduce the productivity of data scientists, they can result in a chain reaction that reduces developer productivity and causes operational expenses to spin out of control.

FlexPod Datacenter for AI

- Accelerates AI/ML initiatives with a validated solution that demystifies deployment
- Scales to more than 20 PB in a single namespace to support very large learning data sets with ONTAP FlexGroups
- Reduces data storage capacity requirements up to 10 times with deduplication and compression techniques
- Supports development, testing, training, and inferencing environments

Cisco UCS C480 ML M5 Rack Server

Designed for AI/ML deployments, the new Cisco UCS C480 ML M5 Rack Server complements your FlexPod deployments and offers:

- High-performance computing with two of the latest Intel® Xeon® Scalable processors
- Unparalleled GPU acceleration with eight NVIDIA Tesla V100-32GB Tensor Core GPUs in a 4-rack-unit (4RU) form factor
- NVIDIA NVLink technology for high bandwidth and massive scalability in multi-GPU configurations
- Flexible options for network, storage, memory, and OS
  - Up to 3 TB of memory
  - Up to 24 HDDs or SSDs
  - Up to 6 NVMe drives
  - Up to 4 Cisco UCS virtual interface cards

Look at your data in new ways

Data stored on your FlexPod infrastructure holds tremendous value. If your existing AI and ML solutions are failing to keep pace—or if you have not implemented AI or ML solutions yet—deploying the FlexPod Datacenter for AI solution with Cisco UCS C-Series Rack Servers can open the door to better insight. This is especially true when data gravity, security, and regulatory requirements dictate that model training be performed on the premises, where your data lives.

FlexPod Datacenter for AI solution

The FlexPod Datacenter for AI solution provides converged infrastructure that is optimized for analytic workloads. Building on the popular FlexPod Datacenter platform, the solution includes Cisco UCS blade and rack servers, Cisco Nexus® 9000 Series switches, Cisco UCS 6000 Series Fabric Interconnects, and NetApp® AFF A-Series flash storage arrays with NetApp ONTAP® data management software.

- **Cisco UCS C480 ML M5 Rack Server.** This no-compromise, purpose-built server integrates graphics processing units (GPUs) and high-speed interconnect technology with fast networking to accelerate deep-learning tasks. The server features two CPUs with up to 28 cores each and up to eight NVIDIA Tesla V100-32GB Tensor Core GPUs that are interconnected with NVIDIA NVLink for fast communication across GPUs to accelerate computation. NVIDIA specifies TensorFlow performance of up to 125 teraFLOPs per module for up to 1 petaFLOP of processing capability per server.
- **Cisco UCS C-Series Rack Servers.** These servers can be equipped with GPU accelerators to meet the needs of other phases of the AI/ML/DL lifecycle, including data aggregation, cleanup, transformation, and inferencing. All of these workloads do not always require the full performance of a deep-learning-optimized server such as the Cisco UCS C480 ML M5. The Cisco UCS C240 M5 Rack Server can host up to four NVIDIA T4 Tensor Core GPUs for AI inferencing, or up to two NVIDIA Tesla V100 Tensor Core GPUs for training workloads. The compact, 1RU Cisco UCS C220 M5 Rack Server can host up to two NVIDIA T4 Tensor Core GPUs.
- **NetApp ONTAP.** The ONTAP software built into NetApp A-Series storage systems makes it easy to create a seamless data lake that spans your distributed data

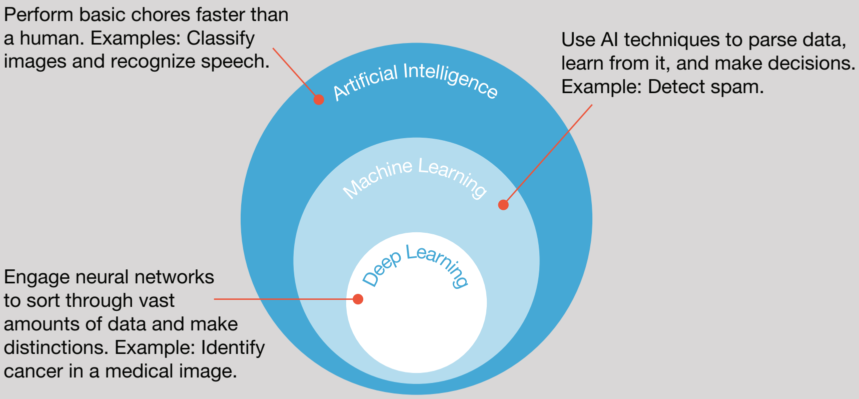

Perform basic chores faster than a human. Examples: Classify images and recognize speech.
Artificial Intelligence
Use AI techniques to parse data, learn from it, and make decisions. Example: Detect spam.
Machine Learning
Deep Learning
Engage neural networks to sort through vast amounts of data and make distinctions. Example: Identify cancer in a medical image.

Figure 1) Evolution to Deep Learning

sources. Your data lake can stream data from the all-flash arrays into your training environment at high speed and with low latency, supporting many I/O streams in parallel. After training completes, the resulting inference models can quickly be moved to a repository and subjected to inference testing and hypothesis validation by Cisco UCS C480 ML M5 servers with massive GPU acceleration for fast results.

Deployment architecture

In the solution, new Cisco UCS C480 ML M5 computing engines place massive GPU acceleration close to the data stored on your FlexPod infrastructure (Figure 2). Cisco UCS C480 ML M5 servers are connected through the system's fabric interconnects, just like the other Cisco UCS blade and rack servers in your FlexPod deployment. Your AI and ML models and applications run on the server, with the NetApp Data Fabric in your FlexPod infrastructure moving data from its collection point or storage location to the computing engines at high speed. This is accomplished with NetApp ONTAP, which helps simplify, accelerate, and integrate your data pipeline. With this integrated approach, you can reap the benefits of a consistent architecture and accelerate your learning models and AI workloads.

Achieve IT and business advantage

The FlexPod Datacenter for AI solution is fully equipped to power your AI, ML, and DL workloads and databases. By deploying this highly scalable architecture, your organization can take advantage of built-in technology advancements and a unified approach to management to achieve many IT and business benefits. The solution integrates with Kubeflow Pipelines to foster collaboration across multiple private and public cloud platforms and provide broad access to artificial intelligence capabilities.

Simplify management

The Cisco UCS C480 ML, C240 M5, and C220 M5 servers used for AI operations can be managed alongside your existing FlexPod infrastructure and data sources. Your IT staff can manage the server with the familiar tools they use to administer

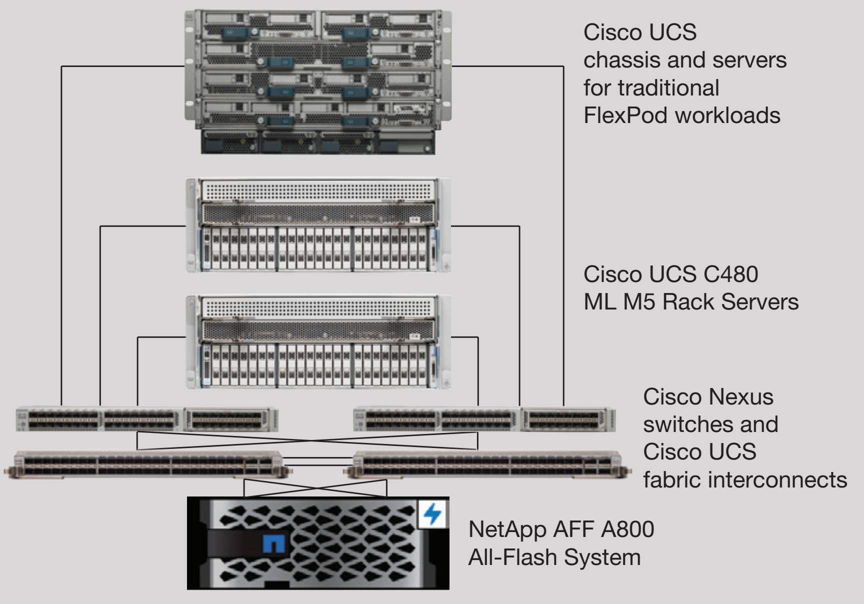


Figure 2) Solution Architecture

Supporting the AI/ML Needs of any Industry

Financial

- Fraud detection
- Cryptocurrencies
- Algorithmic trading

Healthcare

- Medical image screening
- Cancer cell detection
- Drug discovery
- Medical research

Manufacturing

- Inspection
- Quality assurance
- Automation

Media and entertainment

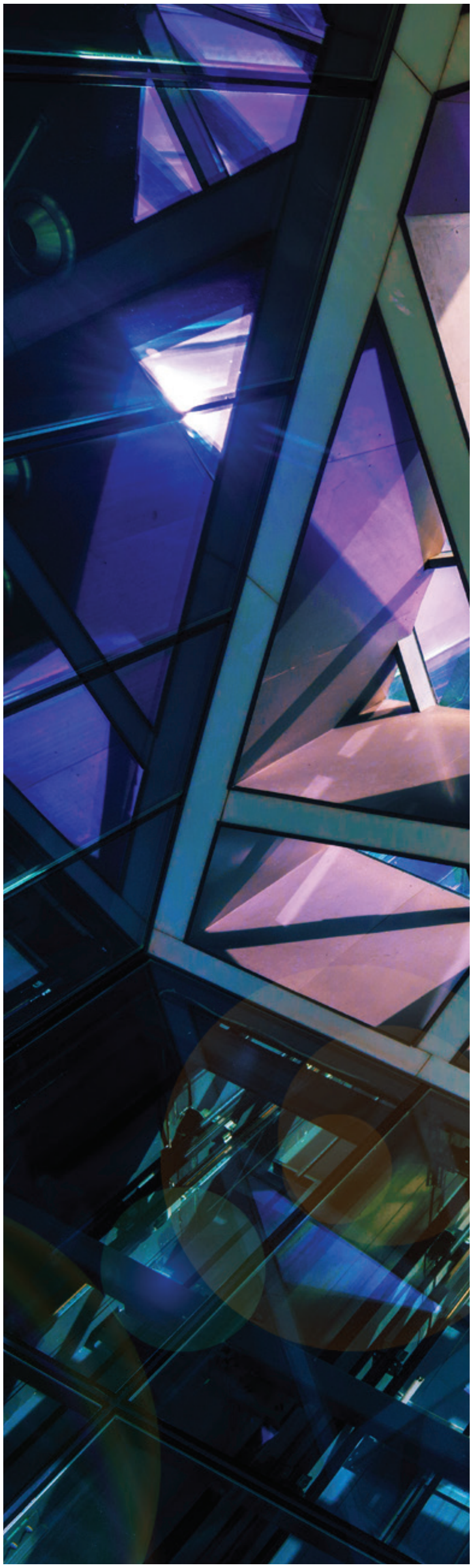
- Video captioning
- Content-based search
- Virtual reality (VR) and augmented reality (AR)

Retail

- Shopping pattern prediction
- Supply chain optimization
- Automated checkout
- Theft detection
- Targeted marketing

Smart cities

- Facial, license plate, and suspicious object recognition
- Traffic pattern analysis
- Intrusion detection
- Cybersecurity measures



your FlexPod systems, reducing costs and administrative overhead. Using Cisco UCS Manager, your IT staff can manage fabrics and logical servers and use models that deliver consistent, error-free, policy-based alignment of server personalities with workloads.

Support continuous integration and development

The FlexClone® capability built into your FlexPod infrastructure makes it easy for your staff to create development and test environments for sandboxes and configuration testing or copies of data sets. Traditional copies can take many minutes or hours to make. With FlexClone technology, even the largest volumes can be cloned in a matter of seconds, enabling you to rapidly improve model accuracy and development velocity as your developers and test engineers spend less time waiting for access to data sets and more time doing productive work.

Scale computing, storage, and GPU capacity

You can expand your AI and ML deployments as demand and data sets change. You can purchase exactly the infrastructure you need for your applications today and scale up (by adding more resources to the FlexPod system or Cisco UCS C-Series Rack Servers or scale out (by adding more FlexPod instances or compute engines). With the massive scalability created with Cisco Nexus 9000 Series Switches and ONTAP software, you can deploy environments that scale to 20 PB and beyond in a single namespace to support very large data sets, resulting in better data models.

The FlexGroup capability in ONTAP creates scale-out NAS volumes consisting of multiple storage components that automatically and transparently share the traffic. Combined with automatic load distribution, FlexGroups make it easy to use infrastructure resources to serve workloads that require massive scalability, high throughput, and low latency, without complicating storage management.

Access existing data sets with ease

In traditional infrastructure deployments, accessing and copying large volumes of data is slow and results in poorly used storage systems. The ONTAP software included in the FlexPod Datacenter for AI solution provides fast access to existing data sets, without the need to copy large volumes of data onto the HDDs of newly deployed FlexPod platforms.

Deliver extreme performance

Delivering better results in less time requires speed at every layer of infrastructure. The extreme low latency performance of NetApp All Flash storage systems and fast networking interconnects supports the accelerated movement of data from where it resides to where it is processed. After its arrival, the multi-GPU configuration of the Cisco UCS C480 ML M5 helps ensure that your data is processed quickly, accessing up to 1 petaFLOP of processing capability per server.

Deliver performance economy

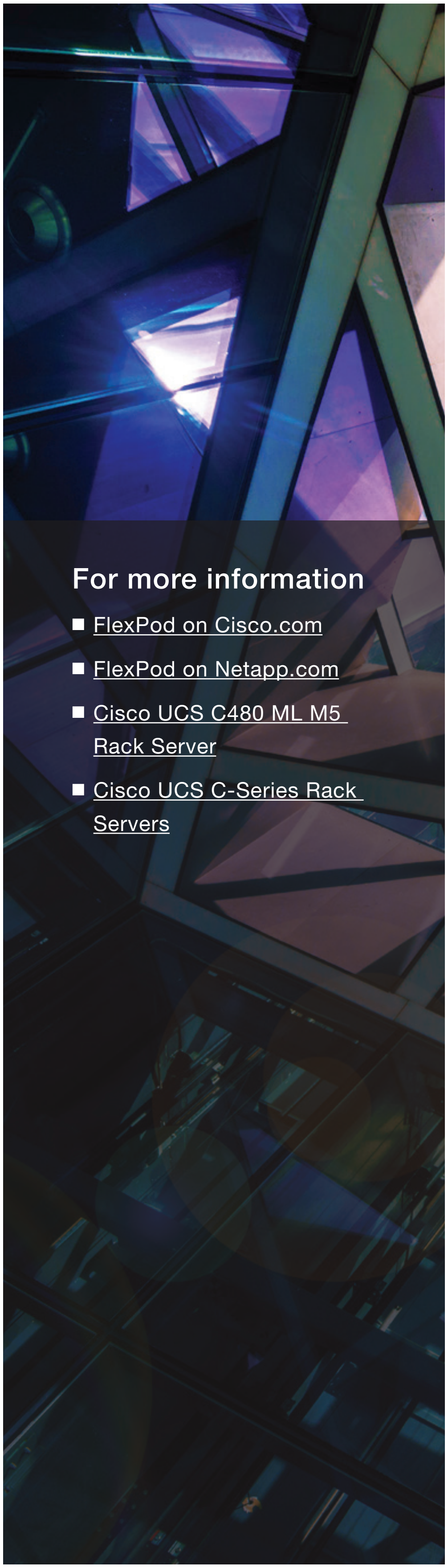
Not all workloads require the power of the Cisco UCS C480 ML M5, and as you deploy your AI models for actual use in inferencing and prediction, you can use lower-cost Cisco UCS C-Series Rack Servers with a smaller number of NVIDIA GPUs to more cost-effectively put your AI models to work.

Reduce risk

Our validation of proven combinations of technologies allows you to extract more intelligence out of all stages of the data lifecycle. In addition, your AI and ML data stays protected through the use of continuous data protection methods that provide near zero recovery time (RT) and recovery point objectives (RPOs). As your AI and ML data sets evolve, you can nondisruptively integrate newer and faster flash technologies such as NVMe, NVMe over fabrics (NVMeoF), and storage-class memory into your FlexPod systems to continuously improve performance without a forklift upgrade.

Learn more

If your organization needs to accelerate access to business insight, consider a FlexPod Datacenter solution. This popular architecture has helped thousands of organizations worldwide deploy infrastructure that is optimized for enterprise applications and databases, and naturally extends to AI/ML environments. With technology advancements and infrastructure optimizations for analytic workloads, the solution makes it easy to power your AI and modern applications to achieve better and faster insight. And if you have an existing FlexPod deployment, you can add the Cisco UCS C480 ML M5 computing engine or Cisco UCS C240 or C220 M5 Rack Servers and extract more intelligence from the valuable data already stored in your data center.



For more information

- FlexPod on Cisco.com
- FlexPod on Netapp.com
- Cisco UCS C480 ML M5 Rack Server
- Cisco UCS C-Series Rack Servers

LE-68702-01 05/19