FlexPod®

Make Your Business Smarter with FlexPod Datacenter for AI

Powered by Cisco UCS X-Series with Intersight

The promise of artificial intelligence

Deep learning (DL) and machine learning (ML) 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 to advantage your business.

You want to use AI and ML to increase revenue and efficiency. The FlexPod® Datacenter for AI solution with the Cisco UCS® X-Series Modular System stands ready to train your models for faster insights.

The need for flexible and fast infrastructure

Artificial intelligence (AI) pushes the bounds of traditional IT infrastructure. Increasingly complex tasks require unprecedented levels of computing power and large amounts of fast storage. The need to test, train, and deploy models with vast amounts of data presses computing, network, and storage resources to the limit with jobs that can take days to complete. But dedicated, specialized resources are a challenge to IT organizations because they require staff to be trained in configuration, management, and maintenance of specialized devices.

You want to use AI and ML to make better, faster decisions and improve business processes. The FlexPod® Datacenter for AI solution with the Cisco UCS X-Series Modular System and NetApp® AFF A-Series and C-Series storage stands ready to deliver the computing and storage resources to test, train, and deploy AI algorithms.

The Cisco UCS X-Series Modular System simplifies your data-center environment, adapting to the specialized needs of DL and ML workloads while also supporting your traditional scale-out and

enterprise workloads. It reduces the number of server types to maintain, helping to improve operational efficiency and agility as it helps reduce complexity. Powered by Cisco Intersight™, a cloud-delivered infrastructure operations platform for compute, storage, and networking infrastructure resources, it shifts your focus from administrative details to business outcomes.

Look at your data in new ways

The flexibility of the Cisco UCS X-Series Modular System means that you can eliminate silos and colocate your enterprise applications that generate data with your ML and DL training workloads that consume it. This helps you unlock the value of your data with a single consistent, on-premises infrastructure solution. This is especially true when data gravity, security, and regulatory requirements dictate that model training be performed on premises, where your data lives.

FlexPod Datacenter for AI solution

FlexPod Datacenter for AI provides converged infrastructure that is optimized for analytic workloads. Building on the popular FlexPod Datacenter platform, the solution includes the Cisco UCS X-Series Modular System with Cisco UCS X-Series compute nodes, Cisco Nexus® 9000 Series Switches, Cisco UCS 6500 Series Fabric Interconnects, and NetApp AFF A-Series and C-Series flash storage arrays with NetApp ONTAP® data management software.

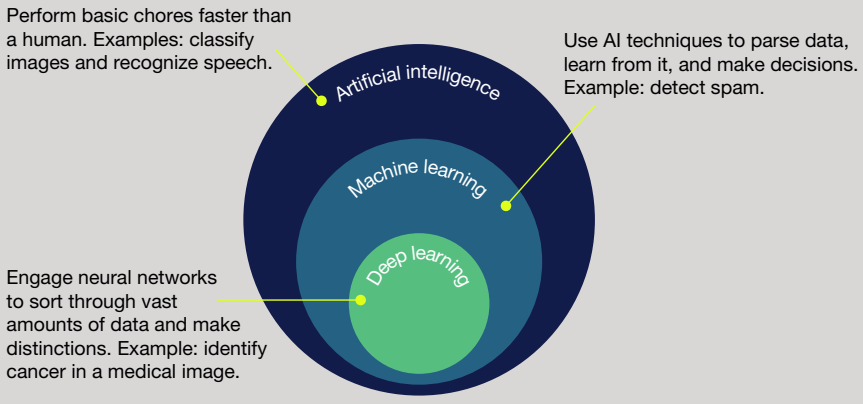

Figure 1) Evolution to deep learning

Supporting the AI/ML needs of the following industries and organizations:

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

Cisco UCS X-Series Modular System

Shaped through Cisco Intersight to match workload needs, including flexible GPU options

The Cisco UCS X210c M7 Compute Node includes:

- High-performance computing with two 4th Gen Intel® Xeon® Scalable Processors
- Up to 8 TB of DDR5 memory
- GPU acceleration with up to two half-height onboard and up to four more half-height or two full-height GPUs with the Cisco UCS X440p PCIe Node
- Up to six SAS/SATA/NVMe drives
- Up to two SATA/NVMe M.2 boot drives
- Up to two Cisco UCS virtual interface cards for up to 200 Gpbs of unified fabric connectivity

The Cisco UCS X410c M7 Compute Node delivers all of the above plus:

- Up to four 4th Gen Intel Xeon processors with up to 60 cores per processor
- Up to 16 TB of memory

- **Cisco UCS X-Series Modular System.** The Cisco UCS X-Series Modular System is ready to serve your data center well into the future with an architecture designed to support many future generations of server I/O and networking technologies. With cloud-based lifecycle management through the Cisco Intersight platform, you can simplify your data center in multiple ways: Simplify with cloud-operated infrastructure that can respond at the speed and scale of your business by shaping resources to workload requirements. Simplify with an adaptable system designed for modern applications—the system can meet the needs of AI and ML software as easily as it supports your enterprise applications. And simplify with a system engineered for the future. Cisco is known for blade server chassis that stand the test of time, and the X-Series designed to serve your data center well into the future by accommodating new technologies as they arise.
- **Cisco UCS X210c M6 and X210c M7 compute nodes.** With your choice of two 3rd or 4th Gen Intel Xeon Processors (respectively), you have the computing power to handle the massive amount of data you need for ML training, and the low latency needed for real-time inferencing. Cisco UCS X440p PCIe Nodes add the capability to add Intel or NVIDIA GPU accelerators to your compute nodes, with up to two double-height GPUs typically used for training, or up to four single-height GPUs typically used for inferencing.With the X210c compute node's capability to support up to two half-height GPUs on board, you can bring the total number of NVIDIA or Intel GPU accelerators to six per node.
- **Cisco UCS X410c M6 Compute Nodes.** When your ML and DL workloads require higher levels of CPU power, the Cisco UCS X410c M6 supports four 4th Gen Intel Xeon processors and 16 TB of memory for staging massive amounts of training data for processing. Its power can be augmented as with the Cisco UCS X210c nodes, with up to six half-height GPUs
- **NetApp AFF A-Series and C-Series with ONTAP.** The NetApp AFF A-Series flash storage arrays deliver the industry's highest performance, superior flexibility, and best-in-class data services. With AI/ML pipeline integration, it helps accelerate, manage, and protect business-critical,

AI/ML data across your enterprise. The systems are smart, powerful, and trusted arrays that take advantage of modern cloud and the latest AI/ML technologies to deliver the speed, efficiency, and security your AI/ML applications need. NetApp AFF C-Series storage leverages budget-friendly, high-capacity, all-flash storage. This storage platform increases efficiency and sustainability with seamless data management and outstanding price/performance. You can purchase what you need now and painlessly scale with your data.

Defend against ransomware attacks and security breaches with new ONTAP autonomous protection and a zero-trust architecture. The ONTAP software built into NetApp storage systems makes it easy to create a seamless data lake that spans your distributed data sources and helps your data scientists share data. Your data lake can stream 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

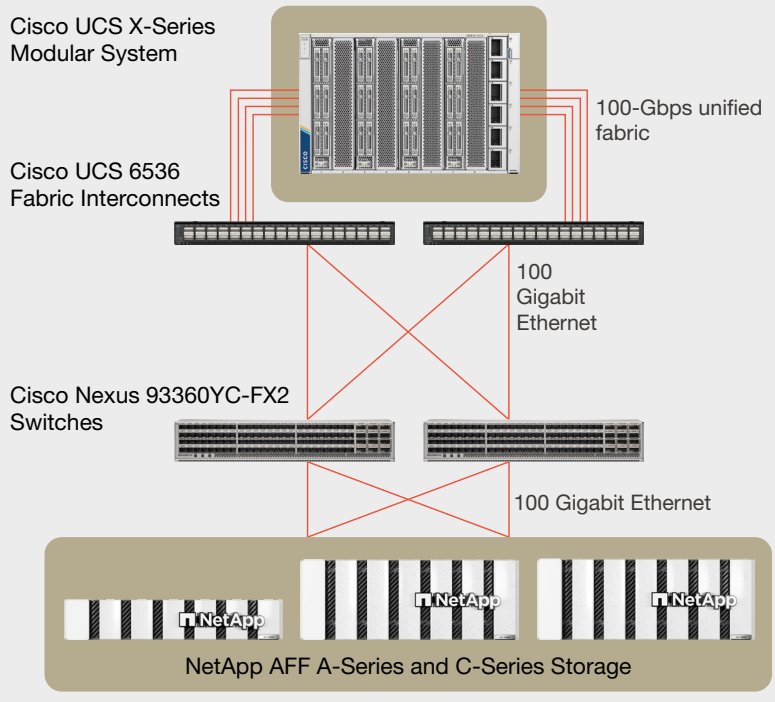

Figure 2) Example FlexPod Datacenter AI configuration with 100-Gbps connectivity

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
- Ensures data security with tamper-proof snapshots, automatic ransomware detection, hardened security, defense against malicious files, and increased flexibility by concurrently accessing storage as file or objects from the same data source

FlexPod: a modern foundation for AI

- **Secure:** FlexPod integrates security at every layer of infrastructure and operations. Built on a zero-trust security model incorporating authentication, encryption, visibility, and control, FlexPod delivers real-time vulnerability notifications, remediation, and data recovery capabilities.
- **Smart:** FlexPod supports AI workloads and helps to reduce risk and accelerate time to value. FlexPod can increase productivity, improve performance, and reduce TCO.
- **Sustainable:** FlexPod introduces our most energy-efficient systems ever! They are engineered to address the entire sustainability lifecycle and enable the continuous activation of new operational capabilities without incremental infrastructure.

repository used in the data center, close to customers at the edge, or in the cloud.

Deployment architecture

The FlexPod Datacenter for AI solution features GPU-accelerated computing engines in the Cisco UCS X-Series chassis combined with NetApp storage. AI applications typically need to store and move massive amounts of data for ML training, and this is facilitated with 100-Gbps throughout thanks to the Cisco® 5th generation fabric—the 100-Gbps Cisco UCS fabric interconnects combined with Cisco Nexus switching. The NetApp data fabric is supported by 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. 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.

Simplify management

The Cisco UCS X210c M6 or M7 and Cisco UCS X420c M7 compute nodes used for AI operations can be managed alongside your existing FlexPod infrastructure and data sources through the Cisco Intersight platform, helping to reduce costs and administrative overhead. With Cisco Intersight, your IT staff can deliver consistent, error-free, policy-based alignment of server personalities and storage characteristics with workloads. The work of data scientists is accelerated by AI-specific management tools.

Accelerate data science

Two NetApp tools help data scientists collaborate and move more quickly. With the NetApp AI Control Plane, Kubeflow integration fosters collaboration across multiple private and public cloud platforms and provides broad access to artificial intelligence capabilities. The NetApp Data Science Toolkit

helps data scientists carry out data management tasks from their existing workspaces without having to learn about storage infrastructure.

Support continuous integration and development

The FlexClone® capability built into your FlexPod infrastructure makes it easy for data scientists 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.

Flexibly scale computing, storage, and GPU capacity

The future-ready, modular architecture of the Cisco UCS X-Series disaggregates computing elements and enables you to nondisruptively scale and adopt new technologies without forklift upgrades. The key to data-center success is flexibility. Because many organizations frequently move workloads between the on-premises data center and the cloud, they require infrastructure that can be shaped to the workloads it needs to serve. With the Cisco UCS X-Series, you can easily support GPU-intensive ML and DL training quickly and easily.

The ONTAP FlexGroup capability creates scale-out NAS volumes consisting of multiple storage components that automatically and transparently share 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

Agile and efficient: how FlexPod drives data-center modernization

65% more time spent on innovation and new projects

43% fewer staff needed to manage

32% faster software installation and management

23% savings on cloud computing

34% reduction in data-center floor space

28% savings on services, outsourcing, and consulting

29% less time spent on monitoring, troubleshooting, and remediation

24% CapEx reduction for both hardware and software

23% savings on annual maintenance fees

29% savings on power and cooling

Source: IDC document #US45212519

the HDDs of newly deployed FlexPod platforms.

Performance density and bandwidth

Delivering better results in less time requires speed at every infrastructure layer. The extreme low-latency performance of NetApp AFF storage systems and 100-Gbps networking supports the accelerated movement of data from where it resides to where it is processed. Support for up to 24 GPUs in the chassis helps ensure that your data is processed quickly.

Accelerated compute flexibility

Not all workloads require the power of multiple double-height GPU accelerators. Once you put your trained models to work, many inferencing applications can use a single Cisco UCS X210c Compute Node with one or two onboard single-height GPUs for cost-effective ongoing AI operations. This is one of the many ways in which the Cisco UCS X-Series Modular System can be shaped to meet your workload needs.

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 short 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 fabric (NVMeoF), and storage-class memory into your FlexPod systems to continuously improve performance without forklift upgrades.

Learn more

- netapp.com/flexpod
- cisco.com/go/flexpod
- Read the blog: FlexPod for your full-stack AI and complete AI lifecycle

Accelerate business insights

FlexPod Datacenter for AI solutions optimized for AI/ML environments. It combines GPU, storage, and networking technologies into a single solution that can power AI applications while helping to improve operational efficiency and agility as it reduces complexity.