FlexPod®

FlexPod Datacenter with generative AI inferencing Solution brief

FlexPod Datacenter: Built for generative AI

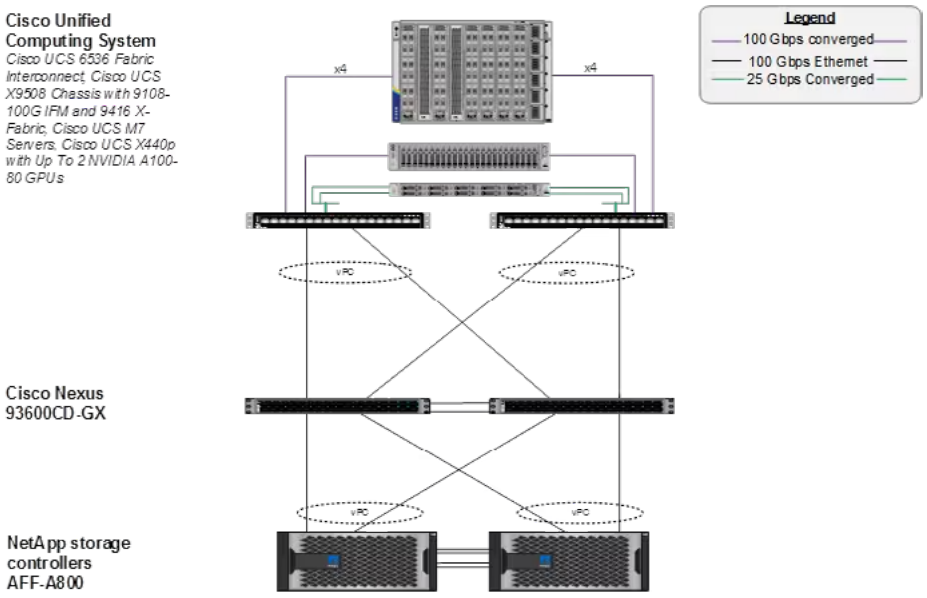
Artificial intelligence workloads generate massive amounts of structured and unstructured data. Cisco, NetApp, and NVIDIA have collaborated to create a design and deployment guide that seamlessly integrates Cisco UCS Servers, Cisco Nexus switches, and NetApp® storage with NVIDIA GPUs. The FlexPod® Cisco Validated Design (CVD) provides a comprehensive and streamlined approach for organizations to configure and customize their environments to provide robust performance for virtually any AI workload.

The intelligent platform for GPU-intensive workloads

Powered by fourth-generation Xeon Scalable processors, Cisco x210c M7 compute nodes provide up to 120 cores and up to 8TB of RAM per server, which work in tandem with Cisco UCS X440p PCI Express nodes. Each PCIe node houses NVIDIA A100-80 GPUs, connecting to compute nodes over Cisco UCS X-Fabric. The CVD validates the NetApp AFF A800 all-flash storage system, with NetApp Astra Trident™ layered on top of NVMe-TCP with 100Gbe providing high-performance persistent storage for containerized workloads. Joining the compute and storage are the latest Cisco Nexus 9000 series switches and Cisco UCS 6500 series fabric interconnects. This combination provides the high performance that generative AI inferencing software and models require.

Programmable infrastructure simplifies deployments

The Cisco Validated Design covers both the design and deployment in one document. By referencing the Infrastructure as Code (IaC) CVD, Ansible is used to simplify the deployment of Red Hat OCP on top of vSphere 8. This results in a 3-node management cluster with 5 nodes to handle containerized AI workloads. Each large language and image generation model runs as a container, creating a persistent volume claim through Astra Trident, which enables persistent, high-performing storage direct to the model of choice.



Cisco Intersight enables sustainable visibility

The FlexPod Generative AI CVD covers benchmarks for some of the industry's leading large language and image generation models, demonstrating low latency and high-performance inferencing. Monitoring this performance is easy with Cisco Intersight, which offers a Power & Energy Metrics Dashboard.

This dashboard offers a high-level overview of the top five power-consuming blade and rack servers, as well as individual host power usage refreshed by the minute to monitor power consumption the moment your AI workloads kick off.

To read more about how FlexPod Datacenter with Generative AI can benefit your business, check out the full Cisco Validated Design.