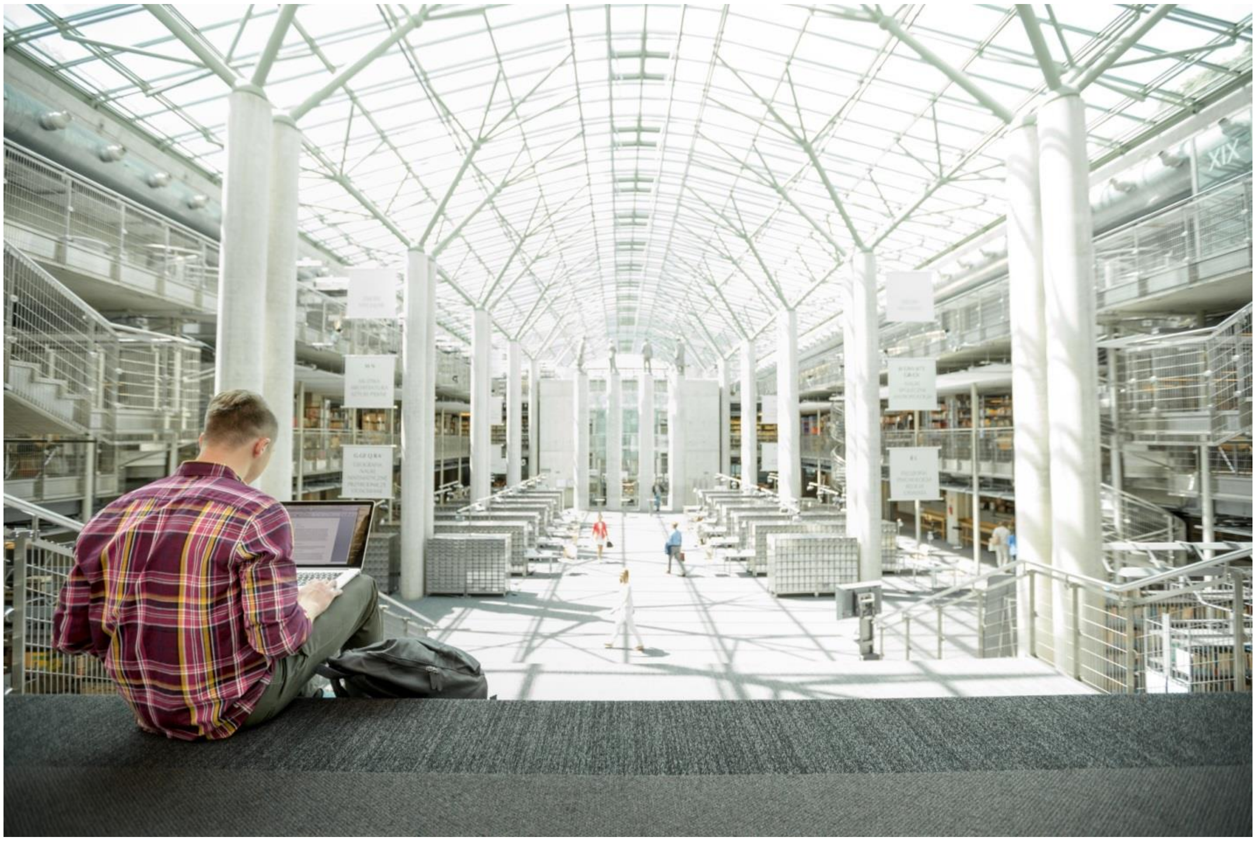


FlexPod Datacenter for AI/ML with Cisco UCS 480 ML for Deep Learning Design Guide

Published: January 16, 2020

About the Cisco Validated Design Program

The Cisco Validated Design (CVD) program consists of systems and solutions designed, tested, and documented to facilitate faster, more reliable, and more predictable customer deployments. For more information, go to:

http://www.cisco.com/go/designzone.

ALL DESIGNS, SPECIFICATIONS, STATEMENTS, INFORMATION, AND RECOMMENDATIONS (COLLECTIVELY, " DESIGNS") IN THIS MANUAL ARE PRESENTED " AS IS," WITH ALL FAULTS. CISCO AND ITS SUPPLIERS DISCLAIM ALL WARRANTIES, INCLUDING, WITHOUT LIMITATION, THE WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OR ARISING FROM A COURSE OF DEALING, USAGE, OR TRADE PRACTICE. IN NO EVENT SHALL CISCO OR ITS SUPPLIERS BE LIABLE FOR ANY INDIRECT, SPECIAL, CONSEQUENTIAL, OR INCIDENTAL DAMAGES, INCLUDING, WITHOUT LIMITATION, LOST PROFITS OR LOSS OR DAMAGE TO DATA ARISING OUT OF THE USE OR INABILITY TO USE THE DESIGNS, EVEN IF CISCO OR ITS SUPPLIERS HAVE BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

THE DESIGNS ARE SUBJECT TO CHANGE WITHOUT NOTICE. USERS ARE SOLELY RESPONSIBLE FOR THEIR APPLICATION OF THE DESIGNS. THE DESIGNS DO NOT CONSTITUTE THE TECHNICAL OR OTHER PROFESSIONAL ADVICE OF CISCO, ITS SUPPLIERS OR PARTNERS. USERS SHOULD CONSULT THEIR OWN TECHNICAL ADVISORS BEFORE IMPLEMENTING THE DESIGNS. RESULTS MAY VARY DEPENDING ON FACTORS NOT TESTED BY CISCO.

CCDE, CCENT, Cisco Eos, Cisco Lumin, Cisco Nexus, Cisco StadiumVision, Cisco TelePresence, Cisco WebEx, the Cisco logo, DCE, and Welcome to the Human Network are trademarks; Changing the Way We Work, Live, Play, and Learn and Cisco Store are service marks; and Access Registrar, Aironet, AsyncOS, Bringing the Meeting To You, Catalyst, CCDA, CCDP, CCIE, CCIP, CCNA, CCNP, CCSP, CCVP, Cisco, the Cisco Certified Internetwork Expert logo, Cisco IOS, Cisco Press, Cisco Systems, Cisco Systems Capital, the Cisco Systems logo, Cisco Unified Computing System (Cisco UCS), Cisco UCS B-Series Blade Servers, Cisco UCS C-Series Rack Servers, Cisco UCS S-Series Storage Servers, Cisco UCS Manager, Cisco UCS Management Software, Cisco Unified Fabric, Cisco Application Centric Infrastructure, Cisco Nexus 9000 Series, Cisco Nexus 7000 Series. Cisco Prime Data Center Network Manager, Cisco NX-OS Software, Cisco MDS Series, Cisco Unity, Collaboration Without Limitation, EtherFast, EtherSwitch, Event Center, Fast Step, Follow Me Browsing, FormShare, GigaDrive, HomeLink, Internet Quotient, IOS, iPhone, iQuick Study, LightStream, Linksys, MediaTone, MeetingPlace, MeetingPlace Chime Sound, MGX, Networkers, Networking Academy, Network Registrar, PCNow, PIX, PowerPanels, ProConnect, ScriptShare, SenderBase, SMARTnet, Spectrum Expert, StackWise, The Fastest Way to Increase Your Internet Quotient, TransPath, WebEx, and the WebEx logo are registered trademarks of Cisco Systems, Inc. and/or its affiliates in the United States and certain other countries.

All other trademarks mentioned in this document or website are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (0809R)

© 2020 Cisco Systems, Inc. All rights reserved.

Table of Contents

Executive Summary

Cisco Validated Designs (CVDs) deliver systems and solutions that are designed, tested, and documented to facilitate and improve customer deployments. These designs incorporate a wide range of technologies and products into a portfolio of solutions that have been developed to address the business needs of the customers and to guide them from design to deployment.

Customers looking to deploy applications using a shared datacenter infrastructure face several challenges. A recurring infrastructure challenge is to achieve the required levels of IT agility and efficiency that can effectively meet the company's business objectives. Addressing these challenges requires having an optimal solution with the following key characteristics:

- Availability: Help ensure applications and services availability at all times with no single point of failure
- Flexibility: Ability to support new services without requiring underlying infrastructure modifications
- Efficiency: Facilitate efficient operation of the infrastructure through re-usable policies
- Manageability: Ease of deployment and ongoing management to minimize operating costs
- Scalability: Ability to expand and grow with significant investment protection
- Compatibility: Minimize risk by ensuring compatibility of integrated components

Cisco and NetApp have partnered to deliver a series of FlexPod solutions that enable strategic datacenter platforms with the above characteristics. FlexPod solution delivers an integrated architecture that incorporates compute, storage, and network design best practices thereby minimizing IT risks by validating the integrated architecture to ensure compatibility between various components. The solution also addresses IT pain points by providing documented design guidance, deployment guidance and support that can be used in various stages (planning, designing and implementation) of a deployment.

Artificial Intelligence (AI) and Machine Learning (ML) initiatives have seen a tremendous growth due to the recent advances in GPU computing technology and Deep learning (DL) and ML techniques offer the potential for unparalleled access to accelerated insights. With the capability to learn from data and make informed and faster decisions, an organization is better positioned to deliver innovative products and services in an increasingly competitive marketplace. ML and AI help organizations make new discoveries, analyze patterns, detect fraud, improve customer relationships, automate processes, and optimize supply chains for unique business advantages. Designing, configuring and maintaining a reliable infrastructure to satisfy the compute, network and storage requirements for these initiatives is the top of mind for all major customers and their IT departments. However, due to intense and rather unique network, storage and processing requirements of the AI/ML workloads, the successful integration of these new platforms into customer environments requires a lot of time and expertise.

This document is intended to provide design details around the integration of the Cisco UCS C480 ML M5 platform into the FlexPod datacenter solution to deliver a unified approach for providing AI and ML capabilities within the converged infrastructure. By providing customers the ability to manage the AI/ML servers with the familiar tools they use to administer traditional FlexPod systems, the administrative overhead as well as the cost of deploying deep learning platform is greatly reduced. The design presented in this CVD also includes other Cisco UCS platforms such as C220 M5 server with 2 NVIDIA T4 GPUs and C240 M5 server equipped with 2 NVIDIA V100 32GB PCIe cards as additional options for AI/ML workloads.

Solution Overview

Introduction

It is well understood that assembling and integrating off-the-shelf hardware and software components increases solution complexity and lengthens deployment times. As a result, valuable IT resources are wasted on systems integration work that can result in fragmented resources which are difficult to manage and require in-depth expertise to optimize and control various deployments.

The FlexPod Datacenter for AI/ML with Cisco UCS C480 ML M5 solution aims to deliver a seamless integration of the Cisco UCS C480 ML platform into the current FlexPod portfolio to enable the customers to easily utilize the platform's extensive GPU capabilities for their workloads without requiring extra time and resources for a successful deployment. FlexPod Datacenter solution is a pre-designed, integrated and validated architecture for datacenter that combines Cisco UCS servers, Cisco Nexus family of switches, Cisco MDS fabric switches and NetApp Storage Arrays into a single, flexible architecture. FlexPod solutions are designed for high availability, with no single points of failure, while maintaining cost-effectiveness and flexibility in the design to support a wide variety of workloads. FlexPod design can support different hypervisor options, bare metal servers and can also be sized and optimized based on customer workload requirements. FlexPod design discussed in this document has been validated for resiliency (under fair load) and fault tolerance during system upgrades, component failures, and partial as well as total power loss scenarios.

Audience

The intended audience of this document includes but is not limited to data scientists, IT architects, sales engineers, field consultants, professional services, IT managers, partner engineering, and customers who want to take advantage of an infrastructure built to deliver IT efficiency and enable IT innovation.

What's New in this Release?

The following design elements distinguish this version of FlexPod from previous models:

- Optimized integration of Cisco UCS C480 ML M5 platform into the FlexPod design
- Integration of NetApp A800 NVMe based all flash storage system to support AI/ML dataset
- Showcase AI/ML workload acceleration using NVIDIA V100 32G GPUs on both Cisco UCS C480 ML M5 and Cisco UCS C240 M5 platforms
- Showcase AI/ML workload acceleration using NVIDIA Testa T4 GPUs on Cisco UCS C220 M5 platform
- Showcase NVIDIA Virtual Compute Servers (vComputeServer) and Virtual GPU (vGPU) capabilities on various Cisco UCS platforms
- Support for Intel 2nd Gen Intel Xeon Scalable Processors (Cascade Lake) processors*
- NetApp FlexGroup volumes and NetApp ONTAP 9.6 release.

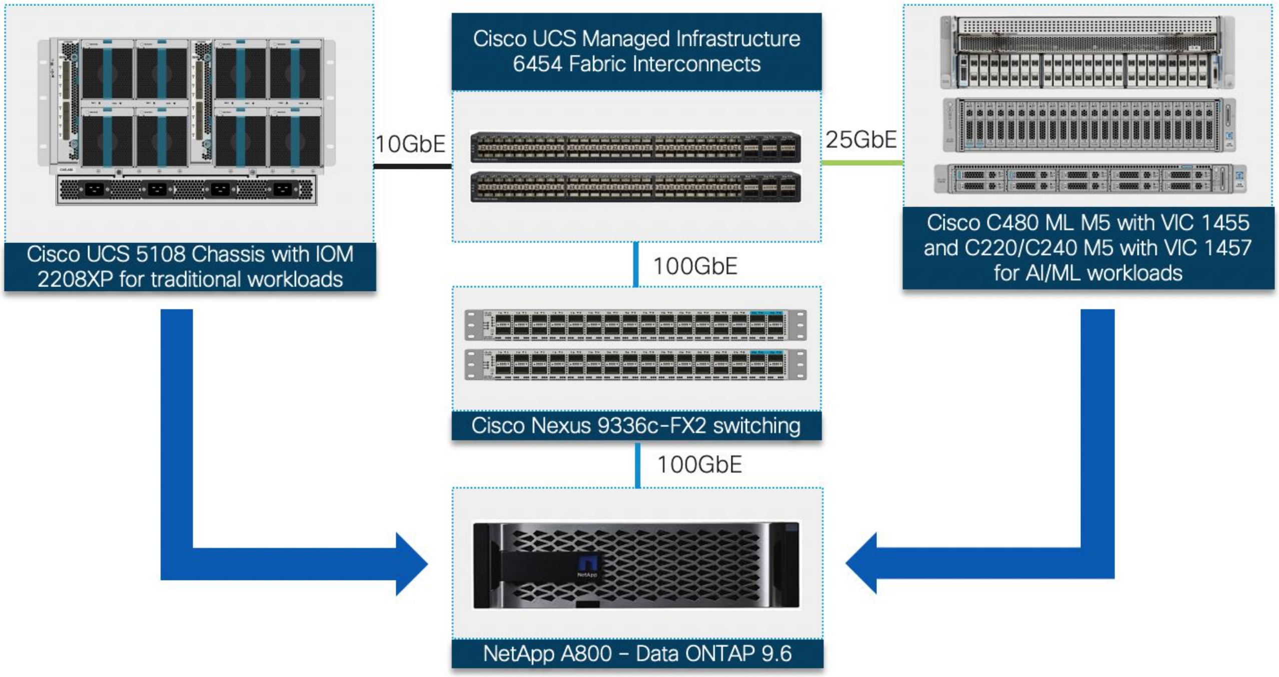
*** The Cisco UCS software version 4.0(4e) (described in this validation) and RHEL 7.6 support Cascade Lake CPUs on Cisco UCS C220 M5 and C240 M5 servers. Support for Cisco UCS C480ML M5 will be available in the upcoming Cisco UCS release.**

For more information about previous FlexPod designs, see: http://www.cisco.com/c/en/us/solutions/design-zone/data-center-design-guides/flexpod-design-guides.html.

Solution Summary

In the FlexPod Datacenter for AI/ML, Cisco UCS C480 ML M5 computing engines place massive GPU acceleration close to the data stored within the FlexPod infrastructure. Just like the other Cisco UCS blade and rack servers in the FlexPod deployment, Cisco UCS C480 ML M5 servers are connected and managed through the Cisco UCS fabric interconnects. The AI and ML workloads and applications run on the Cisco UCS C480 ML server with the NetApp All-Flash system providing storage access using high speed redundant paths. With this integrated approach, customers reap the benefits of a consistent, easily managed architecture where NetApp ONTAP helps simplify, accelerate, and integrate the data pipeline. Using Cisco UCS Manager, customer IT staff can manage fabrics and logical servers by using models that deliver consistent, error-free, policy-based alignment of server personalities with workloads. This architecture seamlessly supports deploying other Cisco UCS C-Series server models with GPUs (for example, Cisco UCS C240M5 and C220 M5) into the design. Figure 1 illustrates the high-level solution overview and connectivity.

Figure 1 FlexPod Datacenter for Deep Learning Powered by Cisco UCS C-Series M5 Servers


Cisco UCS Managed Infrastructure
6454 Fabric Interconnects
10GbE
25GbE
Cisco UCS 5108 Chassis with IOM 2208XP for traditional workloads
Cisco C480 ML M5 with VIC 1455 and C220/C240 M5 with VIC 1457 for AI/ML workloads
100GbE
Cisco Nexus 9336c-FX2 switching
100GbE
NetApp A800 – Data ONTAP 9.6

Like all other FlexPod solution designs, FlexPod Datacenter for Deep Learning is configurable according to the demand and usage. Customers can purchase exactly the infrastructure they need for their current applications requirements and can then scale-up (by adding more resources to the FlexPod system or the Cisco UCS servers) or scale-out (by adding more FlexPod instances). This FlexPod design also incorporates the FlexGroup capability in ONTAP to create scale-out NAS volumes. 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. These FlexGroup volumes can be seamlessly used across virtualized and non-virtualized environments such as environments setup with vGPU capabilities as well as the bare metal installation where workloads consume the entire GPU.

The validated design presented in this document combines proven combination of technologies that allow customers to extract more intelligence out of all stages of the data lifecycle. In addition, customer AI and ML data stays protected using continuous data protection methods (for example, volume-based snapshots) providing near zero recovery time (RT) and recovery point objectives (RPOs).

Technology Overview

FlexPod Datacenter

FlexPod is a datacenter architecture built using the following infrastructure components for compute, network, and storage:

- Cisco Unified Computing System (Cisco UCS)
- Cisco Nexus and Cisco MDS Switches
- NetApp Storage Systems (FAS, AFF, and so on)

These components are connected and configured according to the best practices of both Cisco and NetApp and provide an ideal platform for running a variety of workloads with confidence. One of the key benefits of FlexPod is the ability to maintain consistency at both scale-up and scale-out models. As illustrated in Figure 1 , the current solution comprises of following core components:

- Cisco UCS Manager on Cisco 4th generation 6454 Fabric Interconnects to support 10GbE, 25GbE and 100 GbE connectivity from various components
- Cisco UCS 5108 Chassis with Cisco UCS B200 M5 blade servers to support typical datacenter applications
- Cisco UCS C480 ML M5 server with 8 NVIDIA Tesla V100-32GB GPUs for AI/ML applications
- High-Speed Cisco NxOS based Nexus 9336C-FX2 switching design supporting up to 100GbE connectivity
- NetApp AFF A800 NVMe storage for both traditional and AI/ML workloads with 100GbE connectivity

Cisco UCS C220 M5 server(s) with NVIDIA 2 T4 GPUs or Cisco UCS C240 M5 server(s) with 2 NVIDIA Tesla V100-32GB PCIe GPUs can also be utilized for AI/ML workload processing depending on customer requirements. These platforms are also covered in this design guide.

Cisco Unified Computing System

Cisco Unified Computing System™ (Cisco UCS) is a next-generation datacenter platform that integrates computing, networking, storage access, and virtualization resources into a cohesive system designed to reduce total cost of ownership and increase business agility. The system integrates a low-latency, lossless unified network fabric with enterprise-class, x86-architecture servers. The system is an integrated, scalable, multi-chassis platform with a unified management domain for managing all resources.

The Cisco Unified Computing System consists of the following subsystems:

- Compute - The compute piece of the system incorporates servers based on latest Intel’s x86 processors. Servers are available in blade and rack form factor, managed by Cisco UCS Manager.
- Network - The integrated network fabric in the system provides a low-latency, lossless, 10/25/40/100 Gbps Ethernet fabric. Networks for LAN, SAN and management access are consolidated within the fabric. The unified fabric uses the innovative Single Connect technology to lowers costs by reducing the number of network adapters, switches, and cables. This lowers the power and cooling needs of the system.

- Virtualization - The system unleashes the full potential of virtualization by enhancing the scalability, performance, and operational control of virtual environments. Cisco security, policy enforcement, and diagnostic features are now extended into virtual environments to support evolving business needs.
- Storage access – Cisco UCS system provides consolidated access to both SAN storage and Network Attached Storage over the unified fabric. This provides customers with storage choices and investment protection. Also, the server administrators can pre-assign storage-access policies to storage resources, for simplified storage connectivity and management leading to increased productivity.
- Management: The system uniquely integrates compute, network and storage access subsystems, enabling it to be managed as a single entity through Cisco UCS Manager software. Cisco UCS Manager increases IT staff productivity by enabling storage, network, and server administrators to collaborate on Service Profiles that define the desired physical configurations and infrastructure policies for applications. Service Profiles increase business agility by enabling IT to automate and provision resources in minutes instead of days.

Cisco UCS Manager

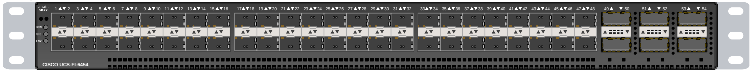
Cisco UCS Manager (UCSM) provides unified, integrated management for all software and hardware components in Cisco UCS. Cisco UCSM manages, controls, and administers multiple blades and chassis enabling administrators to manage the entire Cisco Unified Computing System as a single logical entity through an intuitive GUI, a CLI, as well as a robust API. Cisco UCS Manager is embedded into the Cisco UCS Fabric Interconnects and offers comprehensive set of XML API for third party application integration. Cisco UCSM exposes thousands of integration points to facilitates custom development for automation, orchestration, and to achieve new levels of system visibility and control.

Cisco UCS Fabric Interconnects

The Cisco UCS Fabric Interconnects (FIs) provide a single point for connectivity and management for the entire Cisco UCS system. Typically deployed as an active-active pair, the system's FIs integrate all components into a single, highly-available management domain controlled by the Cisco UCS Manager. Cisco UCS FIs provide a single unified fabric for the system, with low-latency, lossless, cut-through switching that supports LAN, SAN and management traffic using a single set of cables.

The Cisco UCS 6454 (Figure 2) provides the management and communication backbone for the Cisco UCS B-Series Blade Servers, Cisco UCS 5108 B-Series Server Chassis and Cisco UCS Managed C-Series Rack Servers. All servers attached to the Cisco UCS 6454 Fabric Interconnect become part of a single, highly available management domain. In addition, by supporting a unified fabric, the Cisco UCS 6454 provides both the LAN and SAN connectivity for all servers within its domain. The Cisco UCS 6454 supports deterministic, low-latency, line-rate 10/25/40/100 Gigabit Ethernet ports, a switching capacity of 3.82 Tbps, and 320 Gbps bandwidth between FI 6454 and IOM 2208 per 5108 blade chassis, independent of packet size and enabled services.

Figure 2 Cisco UCS 6454 Fabric Interconnect



Cisco UCS VIC 1400

The Cisco UCS Virtual Interface Card (VIC) 1400 Series provides complete programmability of the Cisco UCS I/O infrastructure by presenting virtual NICs (vNICs) as well as virtual HBAs (vHBAs) from the same adapter according to the provisioning specifications within UCSM.

The Cisco UCS VIC 1455 is a quad-port Small Form-Factor Pluggable (SFP28) half-height PCIe card designed for the M5 generation of Cisco UCS C-Series Rack Servers. The card supports 10/25-Gbps Ethernet or FCoE.

The card can present PCIe standards-compliant interfaces to the host, and these can be dynamically configured as either NICs or HBAs. In this CVD, the Cisco VIC 1455 was installed in Cisco UCS C480 ML M5 server.

The Cisco UCS VIC 1457 is a quad-port Small Form-Factor Pluggable (SFP28) mLOM card designed for the M5 generation of Cisco UCS C-Series Rack Servers. The card supports 10/25-Gbps Ethernet or FCoE. The card can present PCIe standards-compliant interfaces to the host, and these can be dynamically configured as either NICs or HBAs. In this CVD, Cisco VIC 1457 was installed in Cisco UCS C220 and Cisco UCS C240 M5 servers.

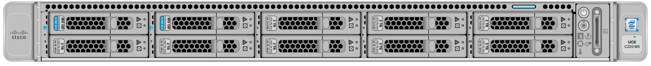
Cisco UCS C220 M5

The Cisco UCS C240 M5 Rack Server is a 2-socket, 1-Rack-Unit (1RU) rack server which supports a wide range of storage and I/O-intensive infrastructure workloads. This modular platform offers following capabilities:

- Up to 2 NVIDIA Tesla T4 enterprise 16GB PCIe GPU adapters
- Latest Intel Xeon Scalable CPUs with up to 28 cores per socket
- Up to 3TB of RAM (24 DDR4 DIMMs) for improved performance
- Support for the Intel Optane DC Persistent Memory (128G, 256G, 512G)
- Up to 10 Small-Form-Factor (SFF) 2.5-inch drives or 4 Large-Form-Factor (LFF) 3.5-inch drives
- Support for 12-Gbps SAS modular RAID controller in a dedicated slot, leaving the remaining PCIe Generation 3.0 slots available for other expansion cards
- Modular LAN-On-Motherboard (mLOM) slot that can be used to install a Cisco UCS Virtual Interface Card (VIC) without consuming a PCIe slot
- Dual embedded Intel x550 10GBASE-T LAN-On-Motherboard (LOM) ports

For more information about Cisco UCS C220 M5 servers, go to: https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/datasheet-c78-739281.html

Figure 3 Cisco UCS C220 M5 rack server



Cisco UCS C240 M5

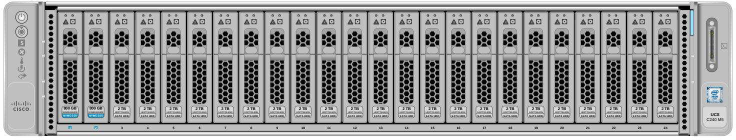
The Cisco UCS C240 M5 Rack Server is a 2-socket, 2-Rack-Unit (2RU) rack server which supports a wide range of storage and I/O-intensive infrastructure workloads. This modular platform offers following capabilities:

- Up to 2 NVIDIA Tesla V100 32GB PCIe GPU adapters
- Up to 6 NVIDIA Tesla T4 enterprise 16GB PCIe GPU adapters
- Latest Intel Xeon Scalable CPUs with up to 28 cores per socket
- Up to 3TB of RAM (24 DDR4 DIMMs) for improved performance
- Support for the Intel Optane DC Persistent Memory (128G, 256G, 512G)

- Up to 26 hot-swappable Small-Form-Factor (SFF) 2.5-inch drives, including 2 rear hot-swappable SFF drives (up to 10 support NVMe PCIe SSDs on the NVMe-optimized chassis version), or 12 Large-Form-Factor (LFF) 3.5-inch drives plus 2 rear hot-swappable SFF drives
- Support for 12-Gbps SAS modular RAID controller in a dedicated slot, leaving the remaining PCIe Generation 3.0 slots available for other expansion cards
- Modular LAN-On-Motherboard (mLOM) slot that can be used to install a Cisco UCS Virtual Interface Card (VIC) without consuming a PCIe slot
- Dual embedded Intel x550 10GBASE-T LAN-On-Motherboard (LOM) ports

For more information about Cisco UCS C240 M5 servers, go to: https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/datasheet-c78-739279.html.

Figure 4 Cisco UCS C240 M5 rack server



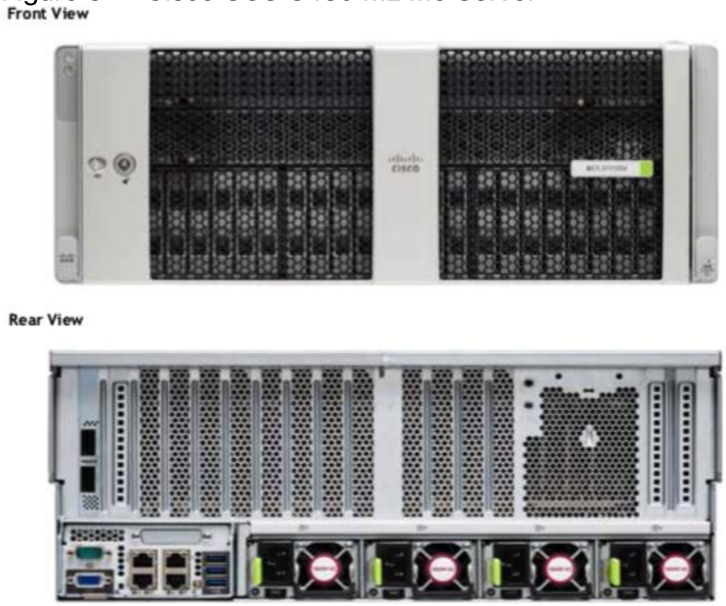
Cisco UCS C480 ML M5

The Cisco UCS C480 ML M5 Rack Server is a purpose-built server for deep learning and is storage and I/O optimized to deliver an industry-leading performance for various training models. The Cisco UCS C480 ML M5 delivers outstanding levels of storage expandability and performance options in standalone or Cisco UCS managed environments using a 4RU form factor (Figure 5). Because of a modular design, the platform offers following capabilities:

- 8 NVIDIA SXM2 V100 32GB modules with NVLink Interconnect
- Latest Intel Xeon Scalable processors with up to 28 cores per socket and support for two processor configurations
- 24 DIMM slots for up to 7.5 terabytes (TB) of total memory
- Support for the Intel Optane DC Persistent Memory (128G, 256G, 512G)
- 4 PCI Express (PCIe) 3.0 slots for multiple 10/25G, 40G or100G NICs
- Flexible storage options with support for up to 24 Small-Form-Factor (SFF) 2.5-inch, SAS/SATA Solid-State Disks (SSDs) and Hard-Disk Drives (HDDs)
- Up to 6 PCIe NVMe disk drives
- Cisco 12-Gbps SAS Modular RAID Controller in a dedicated slot
- Dual embedded 10 Gigabit Ethernet LAN-On-Motherboard (LOM) ports

For more information about Cisco UCS C480 ML M5 Server, go to: https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/c480m5-specsheet-ml-m5-server.pdf

Figure 5 Cisco UCS C480 ML M5 Server



Front View

Rear View

NVIDIA GPU

Graphics Processing Units or GPUs are specialized processors designed to render images, animation and video for computer displays. They perform these tasks by running many operations simultaneously. While the number and kinds of operations they can do are limited, GPUs can run many thousand operations in parallel making this massive parallelism extremely useful for deep learning. Deep learning relies on GPU acceleration for both training and inference and GPU accelerated datacenters deliver breakthrough performance with fewer servers at a lower cost. This CVD details the following NVIDIA GPUs:

- NVIDIA Testa V100 32GB

  NVIDIA Tesla V100 32GB, is an advanced datacenter GPU built to accelerate AI and ML workloads. Cisco UCS C480 ML platform supports 8 NVIDIA V100 SMX2 GPU connected using NVIDIA NVLINK fabric where the GPUs within the same server can communicate directly with each other at extremely high speeds (several times higher than the PCI bus speeds). Each NVLINK has a signaling rate of 25GB/sec in either direction and a single Tesla V100 SMX2 GPU supports up to 6 NVLINK connections which translates to a total bandwidth of 300 GB/sec per GPU. Cisco UCS C240 M5 supports up to 2 NVIDIA V100 PCIe GPUs.

Figure 6 NVIDIA V100 SMX2 GPU

Figure 7 NVIDIA V100 PCIe GPU

- NVIDIA T4 Tensor Core 16GB

  The NVIDIA T4 GPU accelerates diverse cloud workloads, including high-performance computing, deep learning training and inference, machine learning, data analytics, and graphics. Based on the new NVIDIA Turing™ architecture and packaged in an energy-efficient 70-watt, small PCIe form factor, T4 is optimized for mainstream computing environments and features multi-precision Turing Tensor Cores and new RT Cores. Cisco UCS C240 M5 supports up to 6 NVIDIA T4 GPUs and the Cisco UCS C220 M5 supports up to 2 NVIDIA T4 GPUs providing customers flexible deployment options for AI inference workloads.

Figure 8 NVIDIA T4 GPU

NVIDIA CUDA

GPUs are very good at running the same operation on multiple datasets simultaneously referred to as single instruction, multiple data, or SIMD. In addition to rendering graphics efficiently, many other computing problems also benefit from this approach. To support these new workloads, NVIDIA created CUDA. CUDA is a parallel computing platform and programming model that makes it possible to use a GPU for many general-purpose computing tasks through commonly used programming languages such as C and C++. In addition to the general-purpose computing capabilities that CUDA enables, a special CUDA library for deep learning, called the CUDA Deep Neural Network library or cuDNN, makes it easier to implement deep learning and machine learning architectures that take full advantage of the GPU's capabilities. The NVIDIA Collective Communications Library (NCCL) is also part of the CUDA library that enables communication between GPUs both inside a single server as well as across multiple servers. NCCL includes a set of communication primitives for multi-GPU and multi-node configurations enabling topology-awareness for DL training.

NVIDIA Docker

NVIDIA uses containers to develop, test, benchmark, and deploy deep learning frameworks and high-performance computing (HPC) applications. Since Docker does not natively support NVIDIA GPUs within containers, NVIDIA designed NVIDIA-Docker to enable portability in Docker images that leverage NVIDIA GPUs. NVIDIA-Docker is a wrapper around the docker commands that transparently provisions a container with the necessary components to execute code on the GPU. NVIDIA-Docker provides the two critical components needed for portable GPU-based containers: a) driver agnostic CUDA images and b) Docker command line wrapper that mounts the user mode components of the driver and the GPUs into the container at launch

NVIDIA Virtual Compute Server

NVIDIA Virtual Compute Server (vComputeServer) enables datacenters to accelerate server virtualization with GPUs so that the most compute-intensive workloads, such as artificial intelligence, deep learning, and data science, can be run in a virtual machine (VM).

vComputeServer software virtualizes NVIDIA GPUs to accelerate large workloads. With GPU sharing, multiple VMs can be powered by a single GPU, maximizing utilization and affordability, or a single VM can be powered by multiple virtual GPUs (vGPU), making even the most intensive workloads possible. With support for all major hypervisor virtualization platforms, datacenter admins can use the same management tools for their GPU-accelerated servers as they do for the rest of their datacenter.

TensorFlow

TensorFlow™ is an open source software library for high performance numerical computation. Its flexible architecture allows easy deployment of computation across a variety of platforms (CPUs, GPUs, TPUs), and from desktops to clusters of servers to mobile and edge devices. TensorFlow comes with strong support for machine learning and deep learning and the flexible numerical computation core is used across many other scientific domains.

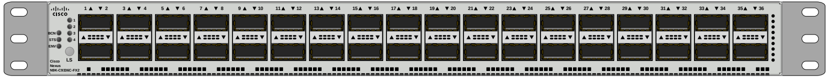
Cisco Nexus Switching Fabric

The Cisco Nexus 9000 Series Switches offer both modular and fixed 1/10/25/40/100 Gigabit Ethernet switch configurations with scalability up to 60 Tbps of non-blocking performance with less than five-microsecond latency, wire speed VXLAN gateway, bridging, and routing support.

The Nexus 9000 switch featured in this CVD is the Nexus 9336C-FX2 (Figure 9) configured in NX-OS standalone mode. NX-OS is a purpose-built datacenter operating system designed for performance, resiliency, scalability, manageability, and programmability at its foundation. It provides a robust and comprehensive feature set that meets the demanding requirements of virtualization and automation in present and future datacenters.

The Cisco Nexus 9336C-FX2 Switch is a 1RU switch that supports 36 ports, 7.2 Tbps of bandwidth and over 2.8 bpps. The switch can be configured to work as 1/10/25/40/100-Gbps offering flexible options in a compact form factor. Breakout is supported on all ports.

Figure 9 Nexus 9336C-FX2 Switch



NetApp AFF Systems

The NetApp all-flash A-Series systems have been designed to provide enterprise-class, scale-out, all-flash storage with the industry's most advanced data-management capabilities and cloud integration. These systems deliver industry-leading performance, capacity density, scalability, security, and network connectivity in highly dense form factors.

In the FlexPod Datacenter for deep learning solution, the NetApp AFF A800 system is utilized. The AFF A800 systems can deliver sub-200µs latency and a massive throughput of 300Gbps in a 24-node cluster scale-out architecture, enabling support for 60% more workloads or cutting application response time in half. The AFF A800 system backed by NVMe can process more data in less time, which is a critical benefit for data analytics, Artificial Intelligence (AI) and Deep Learning (DL) applications.

A few key highlights of the AFF A800 systems include:

- Accelerate artificial intelligence and machine-learning applications by providing an ultrafast end-to-end data path with sub-200µs latency and up to 300GBps throughput.
- Industry's first end-to-end NVMe over FC (NVMe/FC) host-to-flash array over 32Gb FC.
- Industry's first storage system to support 100GbE connectivity.
- Maximum effective capacity of 316.3PB.
- 15TB NVMe solid-state drives (NVMe SSDs) with multistream write (MSW).
- Reduced storage footprint by as much as 37x, 2PB SSD storage in a 2U drive shelf.

Because of all the features and functionality highlighted above, the AFF A800 is the top of the line storage system in terms of performance. There are several other models of all-flash storage systems that are designed to suit different end-user performance requirements and price points. There systems can also be utilized in an AI/ML environment. All these systems are managed by using the data management capabilities offered by NetApp ONTAP® 9.

ONTAP 9

NetApp storage systems harness the power of ONTAP to simplify the data infrastructure from edge, core, and cloud with a common set of data services and 99.9999% availability. NetApp ONTAP 9 data management software from NetApp enables customers to modernize their infrastructure and transition to a cloud-ready datacenter. The ONTAP 9 has a host of features to simplify data management, accelerate, and protect critical data and future-proof infrastructure across hybrid cloud architectures.

Simplifying Deployment and Data Management

The capability to deploy workloads in a matter of minutes, manage the data, and its mobility in a transparent and secure manner is critical for IT operations. The following are a few key features that enable these capabilities:

- NetApp OnCommand System Manager fast-provisioning workflows. Deploy key workloads with all best practices in a few minutes (from power-on to serving data).
- Best-in-class storage efficiency. Reduce wasted space with data compaction, increase effective capacity with deduplication and store more data in less space with compression.
- ONTAP FabricPool. Tier storage between all flash (performance) and object store (external capacity) as needed and reduce storage cost without compromising on performance, efficiency, or protection. From ONTAP 9.5, FabricPools can also use FlexGroup volumes for the performance tier.

- Quality of Service (QoS). Granular QoS controls that help critical workloads to maintain performance especially in highly shared environments.

Accelerate and Protect Critical Data

The fusion of all-flash storage arrays and ONTAP delivers superior levels of performance and data protection by using these key features:

- NVMe. Achieve extremely high throughput with microsecond latency, thus enabling vastly complex and highly resource-dependent applications to run at ease.
- ONTAP FlexGroup. Scale-out NAS containers with near-infinite capacity and predictable low-latency performance in metadata-heavy workloads with a single namespace.
- NetApp Data Availability Services. Replicate data securely from any ONTAP storage—on-premises or in the cloud. Back up ONTAP data directly to an object container. Easily search and recover lost data with a cloud-native management interface.
- NetApp Volume Encryption. Native volume-level encryption with support for onboard and external key management.
- NetApp MetroCluster™. Maintain business continuity and data availability for business-critical applications.

ONTAP provides excellent versatility by responding quickly to changing business requirements and allowing users to move data freely between on-premises environments and leading cloud providers.

Future-Proof Infrastructure

ONTAP can help customers meet the needs of constantly evolving datacenters and changing business with the following features:

- Cloud Integration. ONTAP is the most cloud-connected storage management software, with options for software-defined storage (NetApp ONTAP Select) and cloud-native instances in the form of NetApp Cloud Volumes Service (AWS and GCP), Cloud Volumes ONTAP and Azure NetApp Files.
- Seamless scaling and non-disruptive operations. ONTAP supports non-disruptive addition of capacity to existing controllers and scale-out clusters. Upgrade to the latest technologies such as NVMe and 32Gb FC without costly data migrations or outages.
- Integration with emerging applications. ONTAP provides enterprise-grade data services for next-generation platforms and applications such as OpenStack, Hadoop, MongoDB by using the same infrastructure that supports existing enterprise apps.

NetApp for Artificial Intelligence

NetApp is fully equipped to support applications/ workloads that are based on Artificial Intelligence and Deep Learning. This section discusses some of the key requirements to build a successful AI platform and how NetApp's portfolio meets these requirements.

Parallelization and Performance

The AI applications are massively parallel and the deep neural networks that are powered by GPUs are getting faster. The size of the datasets that these applications crunch is also getting bigger and richer over time. As part of this growth, traditional storage systems are unable to keep up with the speed of these parallel operations. ONTAP with the AFF A800 powered by NVMe makes sure that the storage system is also parallelized and thus supports throughputs as high as 300Gbps with extremely low latency. These high speeds of data transfer between the storage and GPUs enable the GPUs to perform at their peak throughput.

Hybrid Data Sources

The data that needs to be used for training and learning typically comes from diverse sources and applications that might use different storage protocols. These requirements are easily addressed by ONTAP's unified architecture that enables a multiprotocol storage system. The integration of ONTAP with several enterprise and emerging applications like SAP, Oracle, OpenStack, DB2, Splunk, MongoDB, Cassandra, Hadoop, and so on, significantly simplifies the data collection phase.

Data across Geographies – Cloud Integration

ONTAP provides superior data management from edge to core to cloud, which enables data to be consolidated from various locations using the same set of data management features, providing a unified experience to the end-user and a unified consumption model.

ONTAP Select at the edge can collect and process data in offline environments, which can then be moved to the on-premises datacenter or to the cloud.

NetApp Cloud Volumes make NFS capabilities available in AWS, Azure, and Google Cloud Platform and provides NetApp data management for both file and block storage in the cloud. With FabricPool, cold data can be tiered to the public cloud or to on-premises object storage and hot data can be tiered to run on all-flash backed storage.

By using these different capabilities of data mobility in the cloud and on-premises, the required data can be made available in the desired location in a transparent and secure manner.

Automation

Automation of storage features is essential to reduce the storage management impact on AI. ONTAP provides a powerful automation framework with OnCommand Workflow Automation, OnCommand Unified Manager, and OnCommand Performance Manager. The certified software packs are available for download in the Storage Automation Store providing automation capabilities for workflows, integration, data protection, insights, and reporting. NetApp Service Level Manager (SLM) helps to improve operations with predictable performance and cost by accelerating service activation with automated intelligent provisioning to optimize resource usage. NetApp qualified Ansible modules can be utilized for provisioning of storage objects and logical entities. These modules have been designed to meet the provisioning, replication, and general management needs of the ONTAP storage system.

AI-Aware Storage

In addition to providing the necessary storage for data in an AI environment, NetApp Active IQ provides predictive analytics and actionable intelligence in monitoring, securing, and optimizing storage systems and resolving performance issues.

Storage Scaling

AI systems typically process huge volumes of data in a short time frame and can therefore use large datasets to build an accurate algorithm. To handle huge volumes of data, the storage system should be able to scale to near infinite limits and still provide performance and efficiency at scale. With ONTAP FlexGroup volumes, administrators can easily provision a massive single namespace in a few seconds. A single FlexGroup volume can scale linearly up to 20PB and 400 billion files and provide automatic load balancing, superior performance at high capacities with predictable low latency.

Storage Efficiency

With the industry-leading data reduction capabilities from ONTAP that combine deduplication, compression, and compaction, the effective capacity can be increased by several folds. With a storage efficiency as low as 2:1 a 20PB FlexGroup volume can store 40 PB of application data. Although storage space savings vary depending on workloads and use cases, it is not uncommon to see storage efficiencies in the range of 5:1.

Data Durability, Security and Availability

NetApp Volume Encryption provides data-at-rest encryption at a volume level and maintains storage savings for the encrypted volumes. This feature makes sure that there are no additional capacity requirements for storing encrypted data and storage efficiencies stay intact. To maintain data redundancy, the volumes can be mirrored synchronously using NetApp SnapMirror® to maintain redundancy. NetApp Snapshot™ technology allows customers to roll back to a point-in-time dataset in the volume if required. The storage system can also be configured to tolerate multiple controller failures or even loss of an entire datacenter by using NetApp MetroCluster.

NVMe

NVMe (non-volatile memory express) is a host controller interface and storage protocol created to accelerate the transfer of data between enterprise and client systems and solid-state drives (SSDs) over a computer's high-speed Peripheral Component Interconnect Express (PCIe) bus. The protocol implementation is based on a subsystem consisting of specific NVMe controllers, namespaces, nonvolatile storage medium, hosts, ports, and an interface between the controller and storage medium. Although replacing SATA-based SSDs with NVMe SSDs might show some performance improvements, full benefit of the increased performance of NVMe is unlocked by implementing an NVMe over Fabric (NVMe-oF) design. The specifications describe an approach to extend NVMe across network fabrics at scale, allowing multiple storage arrays and hosts to exchange data at NVMe speeds. NVMe-oF supports four fabric options: Fiber Channel (FC), InfiniBand, RDMA over Converged Ethernet (RoCE), and Internet Wide Area RDMA Protocol (iWARP).

Transitioning to NVMe-oF using FC is a simpler option today because of the widespread usage of FC as a storage network. By using FC, the SAN is capable of simultaneously supporting both FCP traffic and NVMe/FC traffic. This enables a smooth transition with minimal changes and does not introduce any major technical design changes.

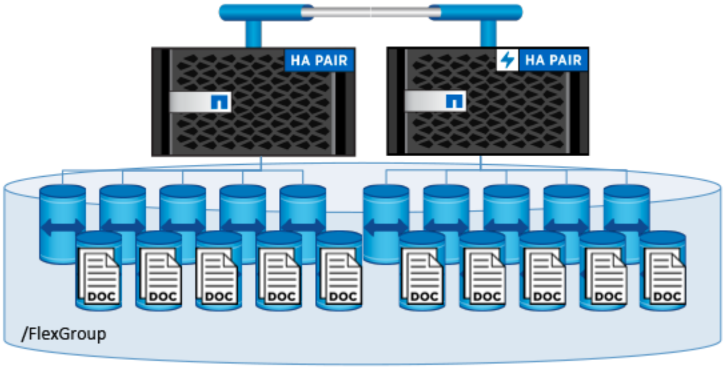
For more information, see the FlexPod End-to-End NVMe White Paper: https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/whitepaper-c11-741907.pdf

NetApp FlexGroup Volumes

A FlexGroup volume provides a massive single namespace that operates best for workloads that contain numerous small files or metadata operations. An AI training or learning dataset is typically a vast collection of files (sometimes billions) that could include structured data, unstructured data, or a combination of both. The GPUs across multiple servers, process this data in parallel, which requires data to be served from a storage system that can allow parallel processing. FlexGroup volumes provide parallelized operations in a scale-out NAS environment across CPUs, controller nodes, aggregates, and the constituent member NetApp FlexVol® volumes.

Additionally, FlexGroup volumes provide Automatic Load Balancing by using all the resources available in the storage cluster and can scale to multiple petabytes of capacity, offering optimal performance. Figure 10 illustrates the architecture of a FlexGroup volume.

Figure 10 NetApp FlexGroup Volume



Multiple FlexVol volumes are stitched together into a single namespace that behaves like a single FlexVol volume to clients and admins. The files are not striped across FlexVol volumes, instead, they are placed systematically into individual FlexVol member volumes that work together under a single namespace. For each new file created, ONTAP decides the best FlexVol member volume to store this file. This decision is based on several factors such as the available capacity across member FlexVol volumes, throughput, last accessed member, and other similar parameters. ONTAP is responsible for keeping the members balanced and for delivering predictable performance.

After the file creation, all read and write operations are performed directly on the member FlexVol volume with ONTAP providing the volume details to the client.

Solution Design

The FlexPod Datacenter for AI/ML with Cisco UCS 480 ML for deep learning solution focuses on the integration of the Cisco UCS GPU enabled platforms into the FlexPod datacenter solution to deliver a solution to support GPU intensive artificial intelligence and machine learning capabilities in the converged infrastructure. The key requirements and design details to deliver this new datacenter solution are outlined in this section.

Requirements

The FlexPod datacenter solution for machine learning closely aligns with latest NxOS based FlexPod CVD and meets the following general design requirements:

1. Resilient design across all layers of the infrastructure with no single point of failure.
2. Scalable design with the flexibility to add compute capacity, storage, or network bandwidth as needed.
3. Modular design that can be replicated to expand and grow as the needs of the business grow.
4. Flexible design that can support components beyond what is validated and documented in this guide.
5. Simplified design with ability to automate and integrate with external automation and orchestration tools.

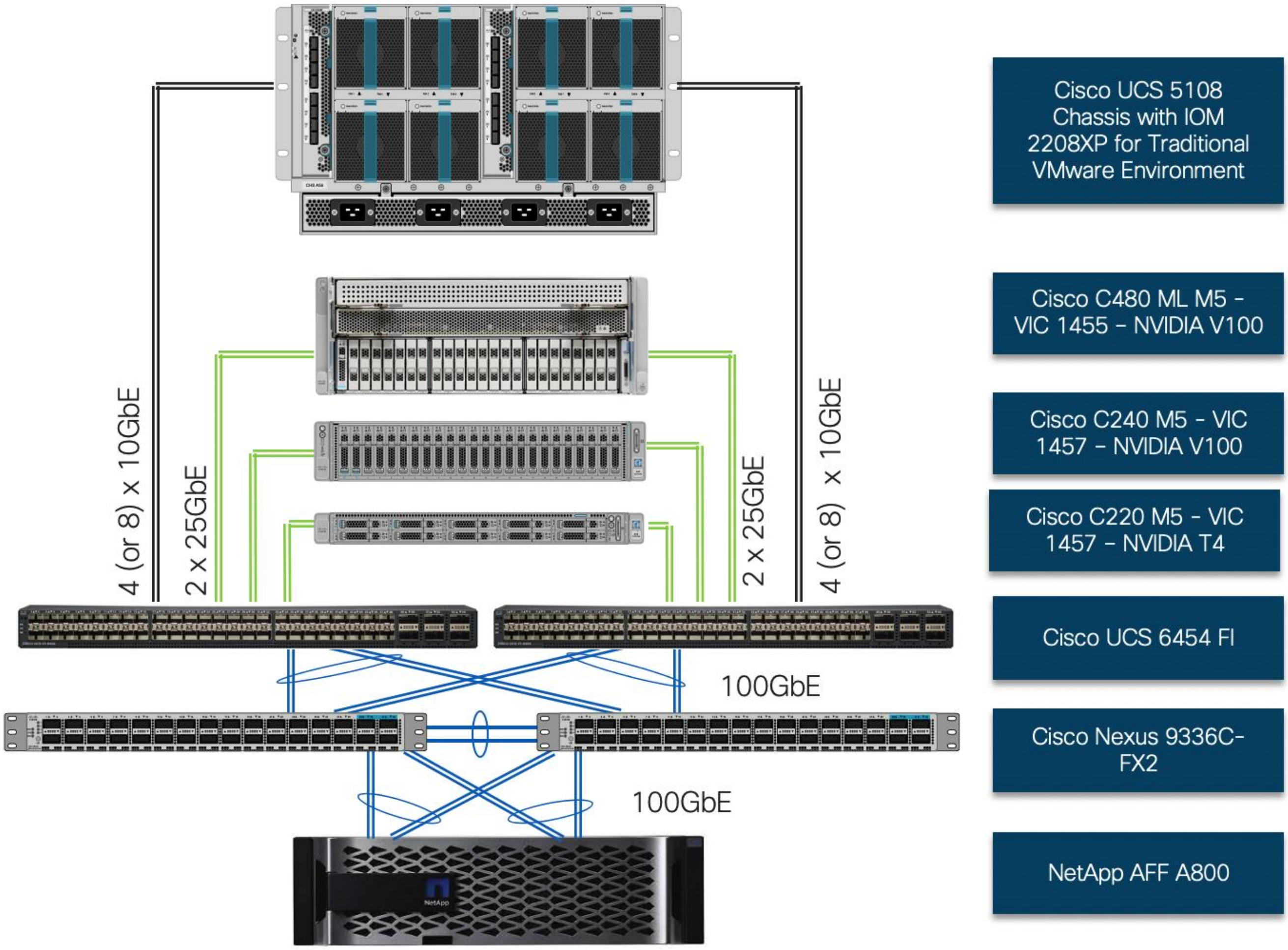
For Cisco UCS C-Series platform integration into a traditional FlexPod datacenter solution, following specific design considerations are also observed:

1. Ability of the Cisco UCS Manager to manage Cisco UCS C480 ML M5 like any other B-Series or C-Series compute node in the design.
2. Support for stateless compute design where the operating system disk is accessed using iSCSI. This operating system disk will coexist on the NetApp A800 controller being used for traditional FlexPod Datacenter environment.
3. High-availability of Cisco UCS C480 ML platform connectivity such that the system should be able to handle one or more link, FI or a storage controller failure.
4. Ability of the switching architecture to enable AI/ML platform to efficiently access AI/ML training and inference dataset from the NetApp A800 controller using NFS.
5. Ability to utilize the GPU capabilities in the VMware environment where multiple Virtual Machines (VMs) can share a GPU as well as a VM can utilize more than a single GPU.
6. Automatic load balancing and parallelized data access and scaling using NetApp FlexGroup volumes.

Physical Topology

The physical topology for the integration of Cisco UCS C-Series M5 platform(s) into a typical FlexPod datacenter design is shown in Figure 11 .

Figure 11 FlexPod for Deep Learning – Physical Topology


Cisco UCS 5108 Chassis with IOM 2208XP for Traditional VMware Environment
Cisco C480 ML M5 – VIC 1455 – NVIDIA V100
Cisco C240 M5 - VIC 1457 – NVIDIA V100
Cisco C220 M5 - VIC 1457 – NVIDIA T4
Cisco UCS 6454 FI
Cisco Nexus 9336C-FX2
NetApp AFF A800
4 (or 8) x 10GbE
2 x 25GbE
2 x 25GbE
4 (or 8) x 10GbE
100GbE
100GbE

To validate the Cisco UCS C480 ML M5 integration into FlexPod datacenter design, an environment with the following components were setup to support both virtual machines and bare metal AI/ML servers:

- Cisco UCS 6454 Fabric Interconnects (FI) to support Cisco UCS 5108 chassis and Cisco UCS C-Series M5 servers.
- Cisco UCS 5108 chassis connected to FIs using 2208XP IOMs. The 2208XP IOMs supports up to 8 10GbE connections to each FI. In this validation, 4 10GbE connections to each FI are utilized leaving the room for expansion in case customers need more bandwidth.
- Cisco UCS C480 ML M5 is connected to each FI using Cisco VIC 1455. Cisco VIC 1455 has 4 25GbE ports. The server was connected to each FI using 2 x 25GbE connections configured as port-channels.
- Cisco UCS C240 M5 and Cisco UCS C220 M5 are connected to each FI using Cisco VIC 1457. Cisco VIC 1457 has 4 25GbE ports. The servers were connected to each FI using 2 x 25GbE connections configured as port-channels.
- Cisco Nexus 9336C running in NxOS mode provides the switching fabric.
- Cisco UCS 6454 FI's 100GbE uplink ports were connected to Nexus 9336C as port-channels.
- NetApp AFF A800 controllers were also connected to Nexus 9336C switch using 100GbE port-channels.

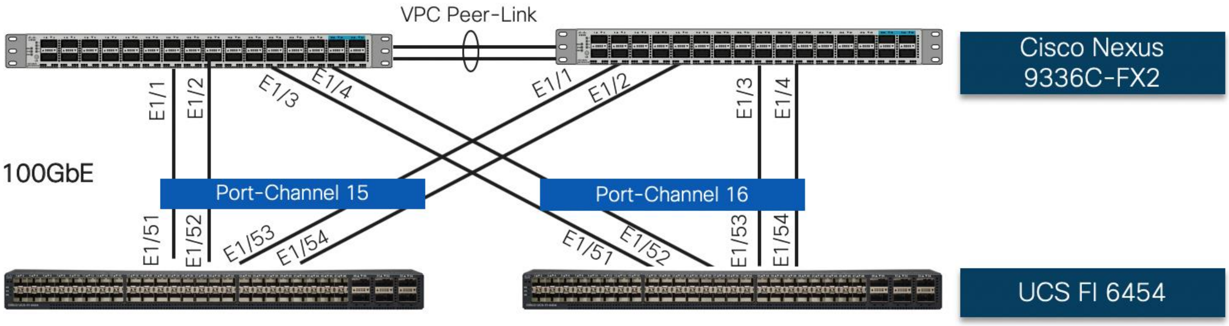
The following sections explain some of these connectivity options in greater detail.

Compute Design

Cisco UCS 6454 Fabric Interconnect Connectivity

Cisco UCS 6454 Fabric Interconnect (FI) is connected to the Nexus switch using 100GbE uplink ports as shown in Figure 12 . Each FI connects to each Nexus 9336C using 2 100GbE ports for a combined bandwidth of 400GbE from each FI to the switching fabric. The Nexus 9336C switches are configured for two separate vPCs, one for each FI.

Figure 12 Cisco UCS 6454 FI to Nexus 9336C Connectivity

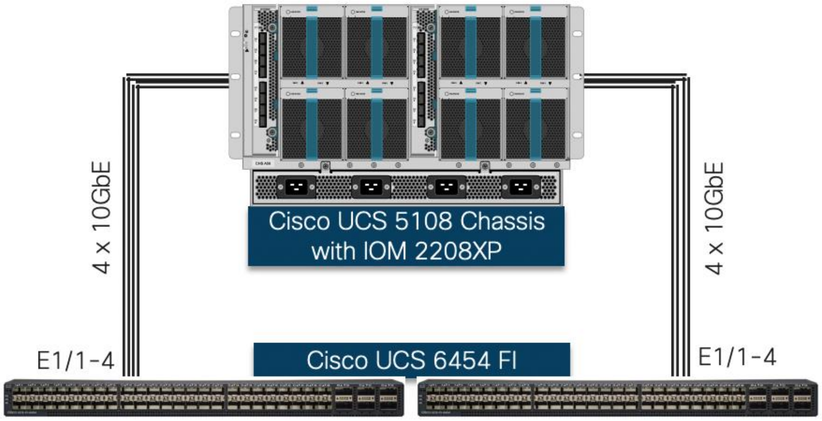


To support both Cisco UCS 5108 chassis and Cisco UCS C480 ML M5 servers, both the devices are physically connected to the same pair of Cisco 6454 FI as shown in Figure 13 and Figure 16 .

Cisco UCS 5108 Connectivity

The Cisco UCS B-Series servers have been utilized for setting up the virtualized environment within the converged infrastructure. Cisco UCS 5108 chassis is equipped with the Cisco UCS 2208XP IO Modules and populated with Cisco UCS B200 M5 blade servers containing Cisco VIC 1440. The servers are configured with appropriate vNICs and diskless iSCSI-based SAN boot to enable stateless compute environment for VMware. Since the B-Series blades are not equipped with GPUs, these ESXi servers host traditional workloads including infrastructure VMs such as AD, DNS, DHCP etc. Figure 13 shows the Cisco UCS 5108 chassis connected to each Cisco UCS 6454 FI using 4 10GbE ports. If the customers require more bandwidth, all 8 ports on Cisco UCS 2208XP IOM can be connected to FI to double the available bandwidth.

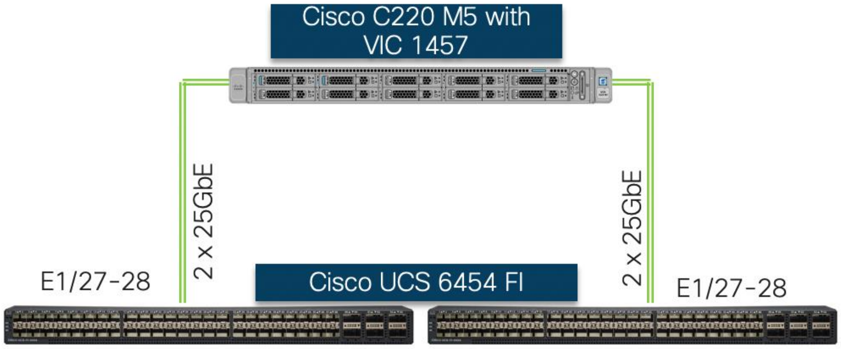
Figure 13 Cisco UCS 5108 to UCS 6454 FI Connectivity



Cisco UCS C220 M5 Connectivity

To manage the Cisco UCS C220 M5 platform with dual NVIDIA T4 GPUs using Cisco UCS Manager, the Cisco UCS C220 M5 is equipped with Cisco VIC 1457. Cisco VIC 1457 has four 25GbE ports which can be connected to the Cisco UCS 6454 FI in pairs such that ports 1 and 2 are connected to the Cisco UCS 6454 FI-A and the ports 3 and 4 are connected to the FI-B as shown in Figure 14 . The ports connected to a fabric interconnect form a port-channel providing an effective 50GbE bandwidth to each fabric interconnect.

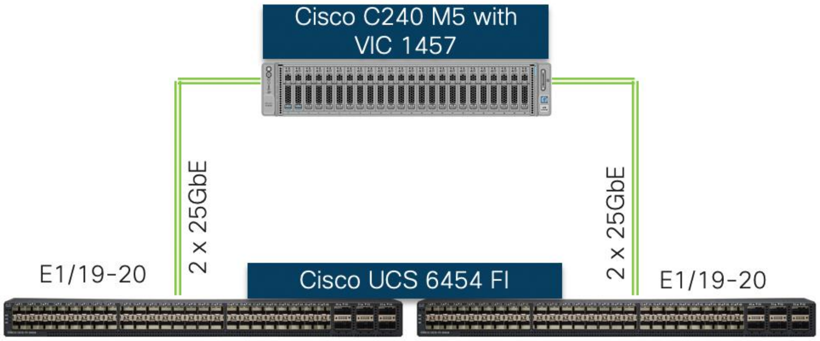
Figure 14 Cisco UCS C220 M5 to Cisco UCS 6454 FI Connectivity



Cisco UCS C240 M5 Connectivity

To manage the Cisco UCS C240 M5 platform with dual GPUs using Cisco UCS Manager, the C240 M5 is equipped with Cisco VIC 1457. Cisco VIC 1457 has four 25GbE ports which can be connected to the Cisco UCS 6454 FI in pairs such that ports 1 and 2 are connected to the Cisco UCS 6454 FI-A and the ports 3 and 4 are connected to the FI-B as shown in Figure 15 . The ports connected to a fabric interconnect form a port-channel providing an effective 50GbE bandwidth to each fabric interconnect.

Figure 15 Cisco UCS C240 M5 to Cisco UCS 6454 FI Connectivity



Managing Cisco UCS C480 ML M5 using Cisco UCS Manager

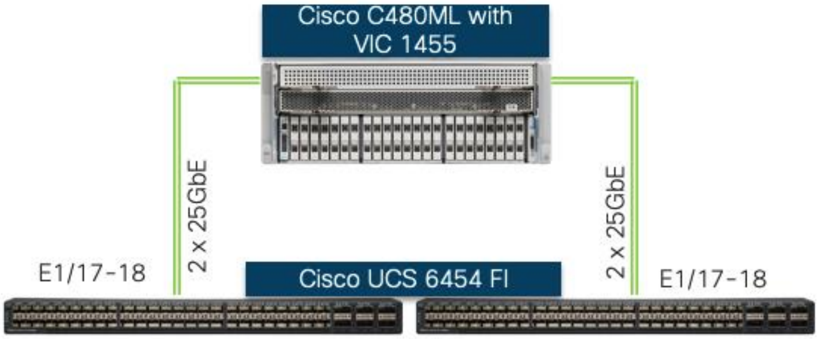
Cisco UCS C480 ML M5 Connectivity

To manage the Cisco UCS C480 ML M5 platform using Cisco UCS Manager, the Cisco UCS C480 platform is equipped with Cisco VIC 1455.

For Cisco UCS C480 ML M5 integration with Cisco UCS Manager, the Cisco VIC 1455 is installed in PCIe Slot 11.

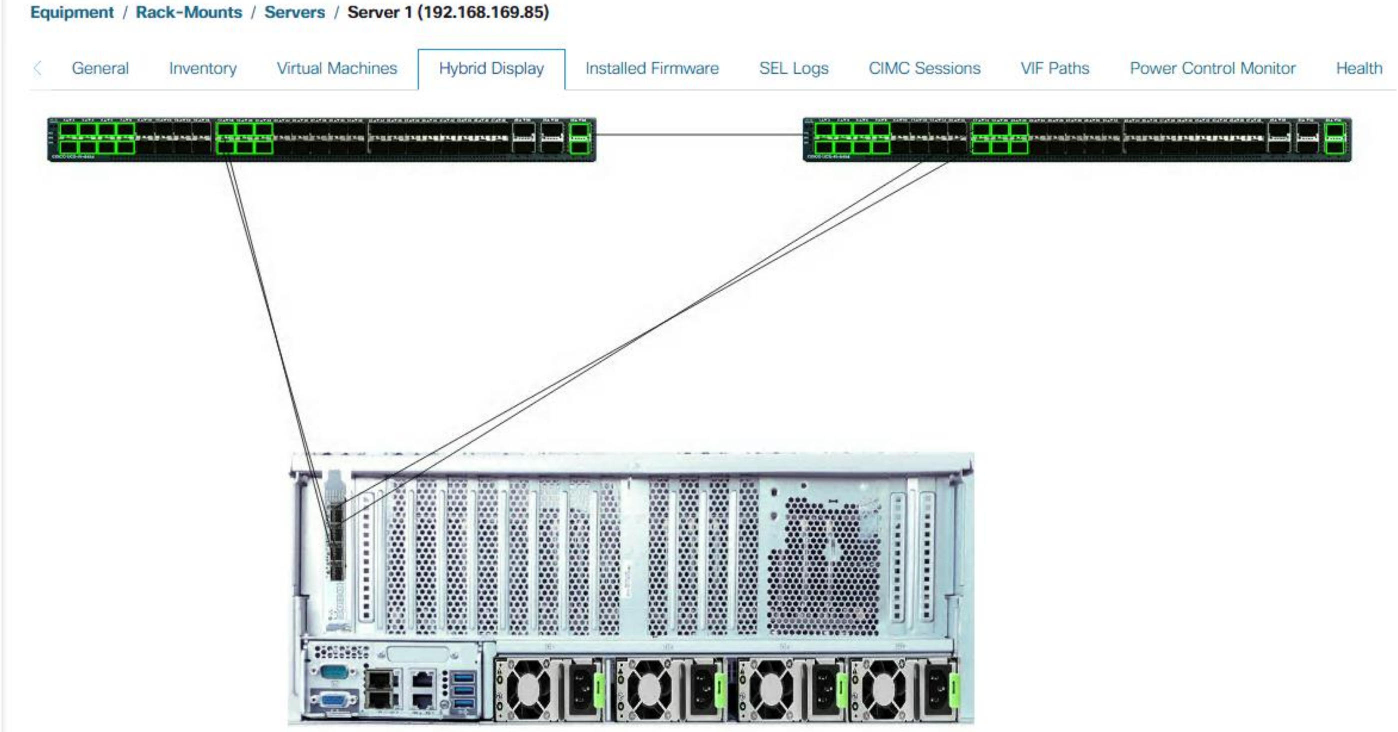
Cisco VIC 1455 has four 25GbE ports which can be connected to the Cisco UCS 6454 FI in pairs such that ports 1 and 2 are connected to the Cisco UCS 6454 FI-A and the ports 3 and 4 are connected to the FI-B as shown in Figure 16 . The ports connected to a fabric interconnect form a port-channel providing an effective 50GbE bandwidth to each fabric interconnect.

Figure 16 Cisco UCS C480 ML M5 to Cisco UCS 6454 FI Connectivity



On successful discovery of the Cisco UCS C480 ML platform within the UCSM managed environment, Cisco UCS C480 ML will appear in the Cisco UCS Manager as shown in some of the figures below. Figure 17 shows Cisco UCS C480 connectivity to the Cisco UCS 6454 FIs:

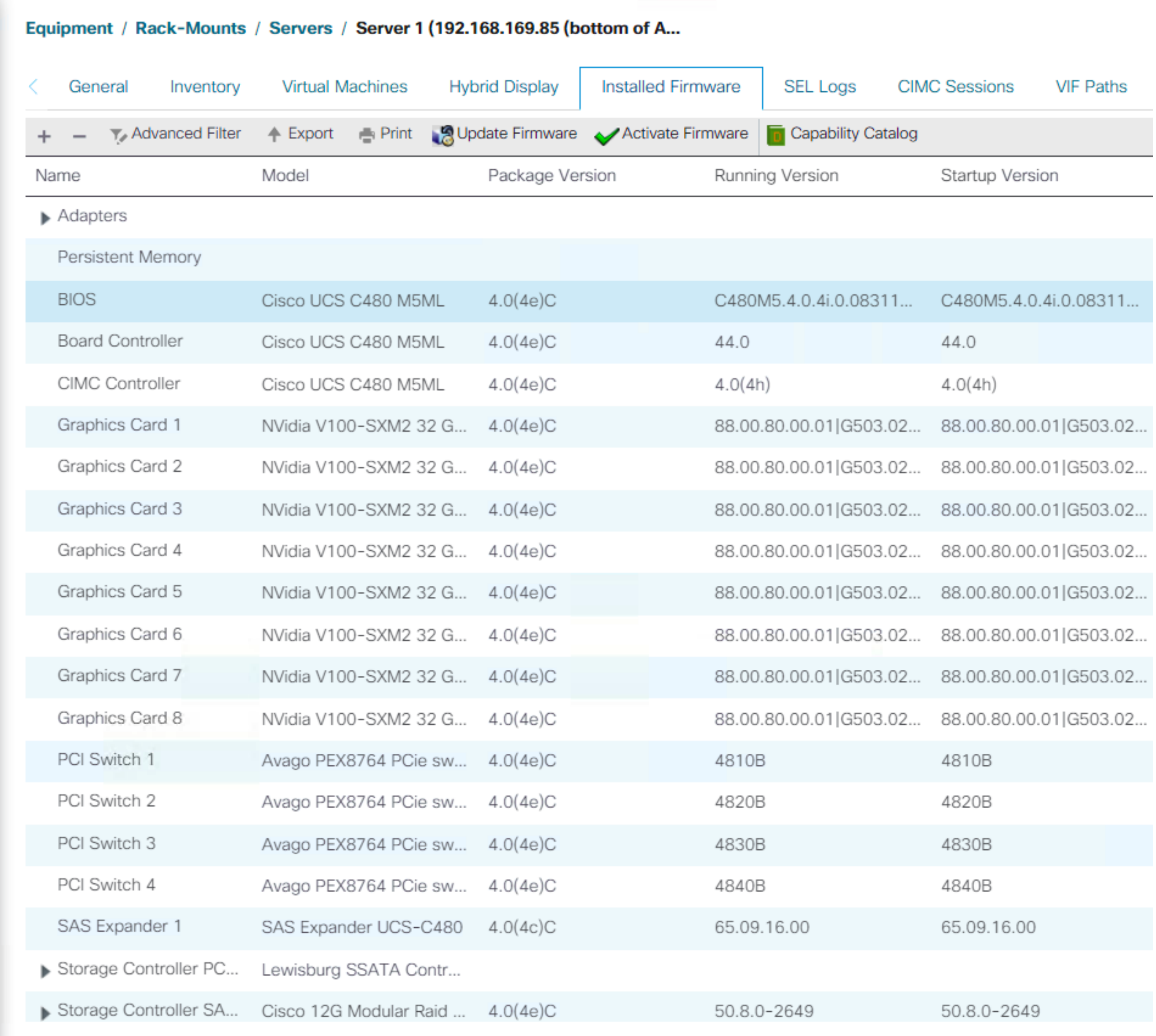
Figure 17 Cisco UCS C480 ML M5 Hybrid Display



Cisco UCS C480 ML M5 Inventory Information

Since the Cisco UCS C480 ML is now being managed by Cisco UCS Manager, the firmware management for the platform is also handled by Cisco UCS Manager. Figure 18 shows various firmware versions and upgrade options for the platform including the firmware installed on NVIDIA GPUs.

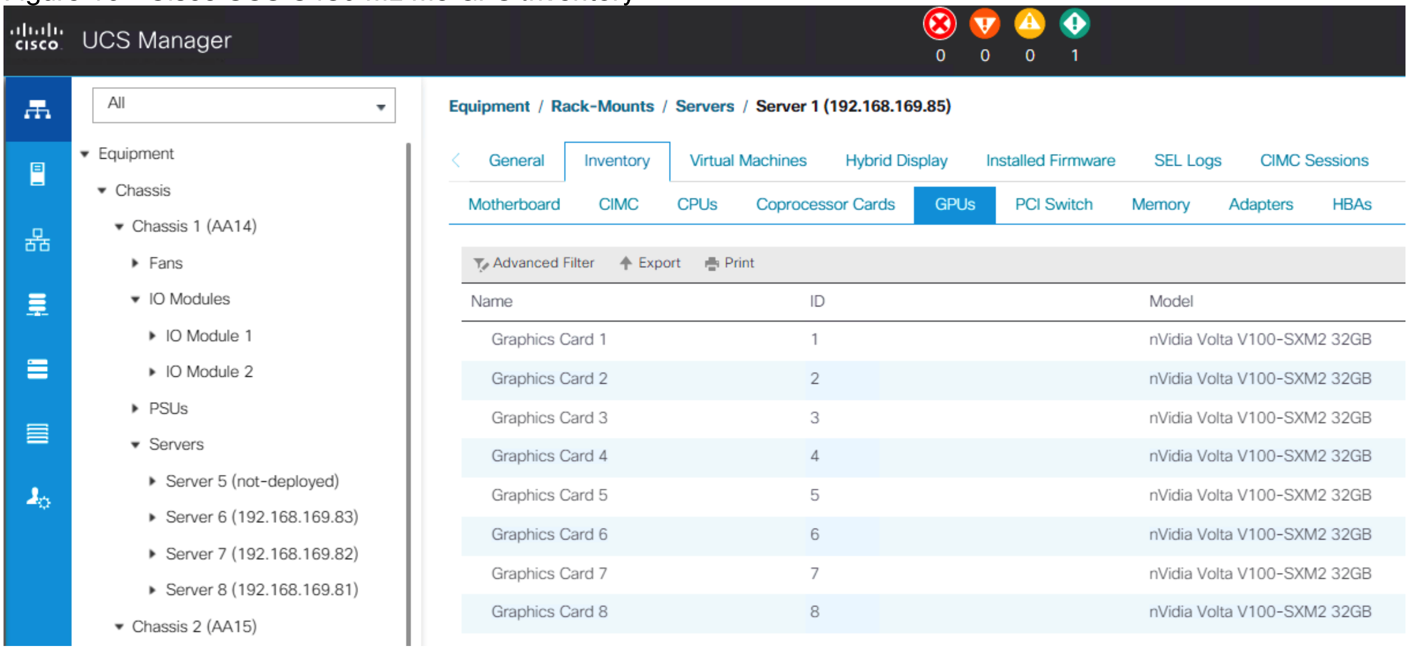
Figure 18 Cisco UCS C480 ML M5 Firmware Management



Name	Model	Package Version	Running Version	Startup Version
▶ Adapters				
Persistent Memory				
BIOS	Cisco UCS C480 M5ML	4.0(4e)C	C480M5.4.0.4i.0.08311...	C480M5.4.0.4i.0.08311...
Board Controller	Cisco UCS C480 M5ML	4.0(4e)C	44.0	44.0
CIMC Controller	Cisco UCS C480 M5ML	4.0(4e)C	4.0(4h)	4.0(4h)
Graphics Card 1	NVidia V100-SXM2 32 G...	4.0(4e)C	88.00.80.00.01|G503.02...	88.00.80.00.01|G503.02...
Graphics Card 2	NVidia V100-SXM2 32 G...	4.0(4e)C	88.00.80.00.01|G503.02...	88.00.80.00.01|G503.02...
Graphics Card 3	NVidia V100-SXM2 32 G...	4.0(4e)C	88.00.80.00.01|G503.02...	88.00.80.00.01|G503.02...
Graphics Card 4	NVidia V100-SXM2 32 G...	4.0(4e)C	88.00.80.00.01|G503.02...	88.00.80.00.01|G503.02...
Graphics Card 5	NVidia V100-SXM2 32 G...	4.0(4e)C	88.00.80.00.01|G503.02...	88.00.80.00.01|G503.02...
Graphics Card 6	NVidia V100-SXM2 32 G...	4.0(4e)C	88.00.80.00.01|G503.02...	88.00.80.00.01|G503.02...
Graphics Card 7	NVidia V100-SXM2 32 G...	4.0(4e)C	88.00.80.00.01|G503.02...	88.00.80.00.01|G503.02...
Graphics Card 8	NVidia V100-SXM2 32 G...	4.0(4e)C	88.00.80.00.01|G503.02...	88.00.80.00.01|G503.02...
PCI Switch 1	Avago PEX8764 PCIe sw...	4.0(4e)C	4810B	4810B
PCI Switch 2	Avago PEX8764 PCIe sw...	4.0(4e)C	4820B	4820B
PCI Switch 3	Avago PEX8764 PCIe sw...	4.0(4e)C	4830B	4830B
PCI Switch 4	Avago PEX8764 PCIe sw...	4.0(4e)C	4840B	4840B
SAS Expander 1	SAS Expander UCS-C480	4.0(4c)C	65.09.16.00	65.09.16.00
▶ Storage Controller PC...	Lewisburg SSATA Contr...			
▶ Storage Controller SA...	Cisco 12G Modular Raid ...	4.0(4e)C	50.8.0-2649	50.8.0-2649

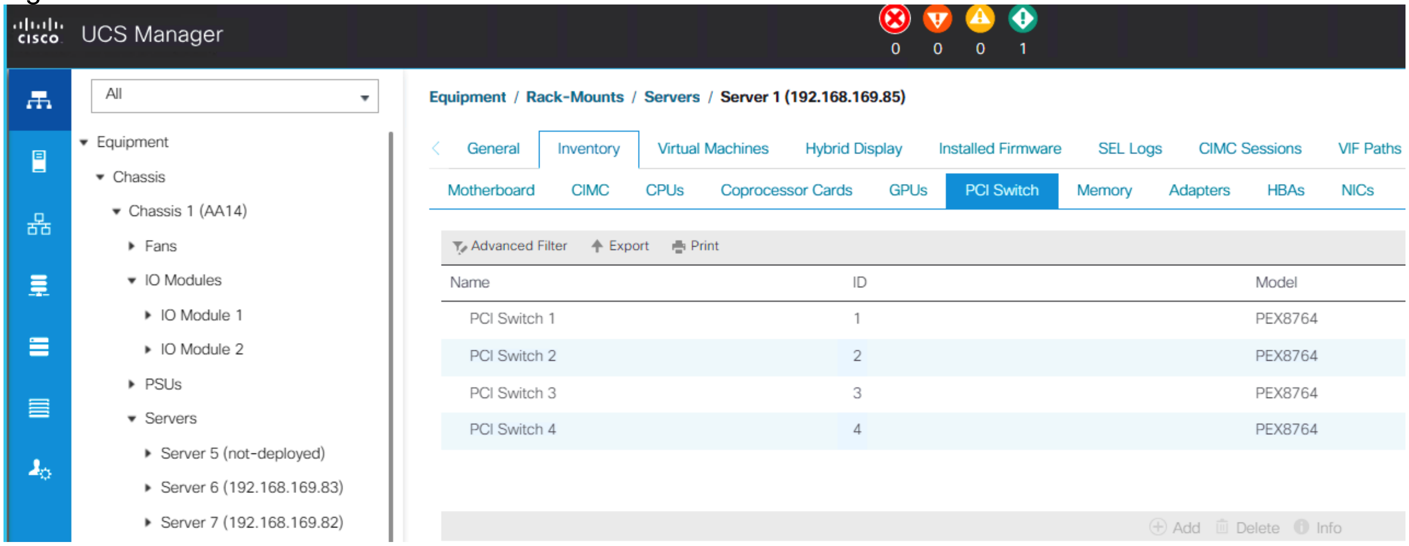
Additionally, Cisco UCS Manager also shows the number and models of the NVIDIA GPUs (Figure 19) in Cisco UCS C480 ML M5 platforms under the inventory page and shows the information about the PCI switches (Figure 20) as well.

Figure 19 Cisco UCS C480 ML M5 GPU Inventory



Name	ID	Model
Graphics Card 1	1	nVidia Volta V100-SXM2 32GB
Graphics Card 2	2	nVidia Volta V100-SXM2 32GB
Graphics Card 3	3	nVidia Volta V100-SXM2 32GB
Graphics Card 4	4	nVidia Volta V100-SXM2 32GB
Graphics Card 5	5	nVidia Volta V100-SXM2 32GB
Graphics Card 6	6	nVidia Volta V100-SXM2 32GB
Graphics Card 7	7	nVidia Volta V100-SXM2 32GB
Graphics Card 8	8	nVidia Volta V100-SXM2 32GB

Figure 20 Cisco UCS C480 ML M5 PCI Switch Information



Name	ID	Model
PCI Switch 1	1	PEX8764
PCI Switch 2	2	PEX8764
PCI Switch 3	3	PEX8764
PCI Switch 4	4	PEX8764

Service Profile Configuration

FlexPod design presented in this document outlines integration of traditional VMware environment and GPU equipped Cisco UCS C-Series servers for deep learning workloads. The service profile configurations for supporting GPU functionality in both virtualized and bare-metal environments is explained below.

VLANs Configuration

Table 1 list various VLANs configured for setting up the FlexPod environment including their usage.

Table 1 VLAN Usage

VLAN ID	Name	Usage
2	Native-VLAN	Use VLAN 2 as Native VLAN instead of default VLAN (1)

VLAN ID	Name	Usage
20	IB-MGMT-VLAN	Management VLAN to access and manage the servers
220	Data-Traffic	VLAN to carry data traffic for both VM and bare-metal Servers
3100	vMotion	VMware vMotion traffic
3111 (Fabric A only)	iSCSI-A	iSCSI-A path for booting both B-Series and C-Series servers
3121 (Fabric B only)	iSCSI-B	iSCSI-B path for booting both B-Series and C-Series servers
3151	ESXi-NFS-VLAN	NFS VLAN for mounting ESXi datastores in ESXi environment
3152	AI-ML-NFS	NFS VLAN to access AI/ML NFS volume hosting ImageNet data

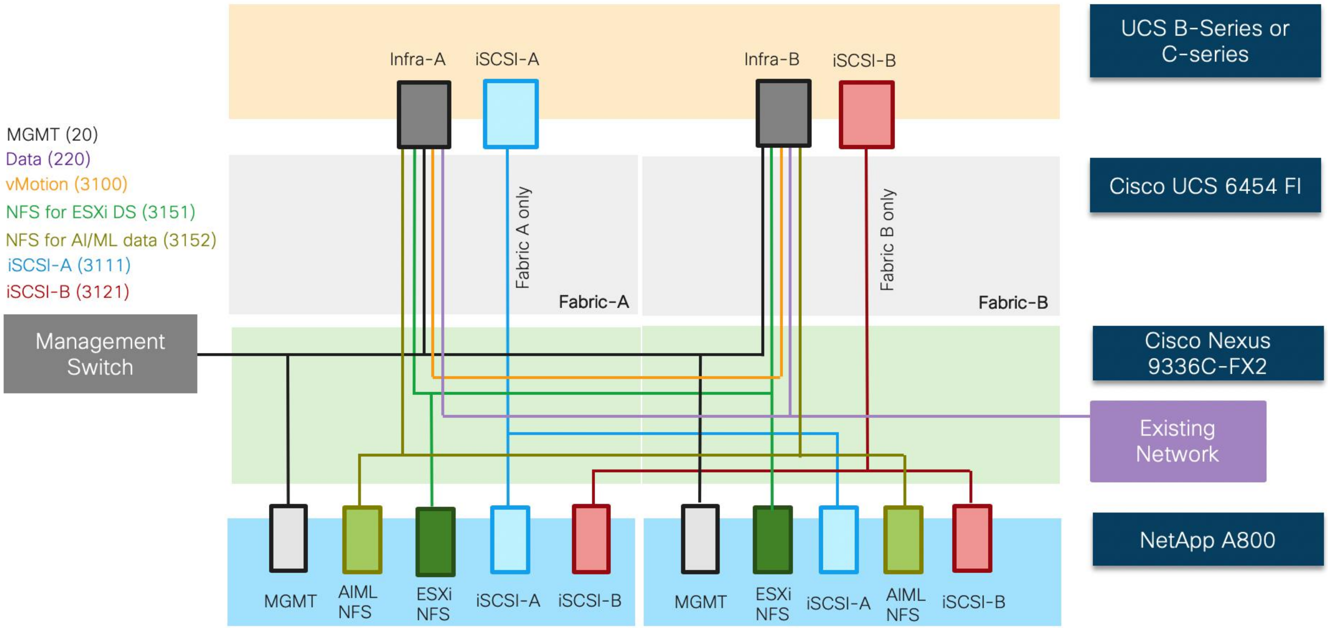
Some of the key highlights of VLAN usage are as follows:

- Both virtual machines and bare-metal servers are managed using same VLAN (20).
- An optional dedicated VLAN (220) is used for virtual machine and bare-metal data communication. Customers are encouraged to evaluate this VLAN usage according to their specific requirements.
- Utilizing separate NFS VLANs for datastores in traditional ESXi environment and for mounting data shares for AI/ML hosts provides path selection flexibility and the ability to configure specific QoS policies (if required).
- A common pair of iSCSI VLANs are utilized to access boot LUNs for ESXi servers as well as bare-metal (RHEL) servers. Customers can also use separate pairs of iSCSI VLANs when using dedicated SVMs for VMware and bare-metal environments.

Service Profile for VMware Hosts

In FlexPod Datacenter deployments, Cisco UCS Service Profiles are provisioned from Service Profile Templates to allow rapid deployment of servers with guaranteed configuration consistency. Each Cisco UCS server (B-Series or C-Series), equipped with a Cisco Virtual Interface Card (VIC), is configured for multiple virtual interfaces (vNICs) which appear as standards-compliant PCIe endpoints to the OS. The service profile configuration for an ESXi host is as shown in Figure 21

Figure 21 ESXi Service Profile



Each ESXi service profile supports:

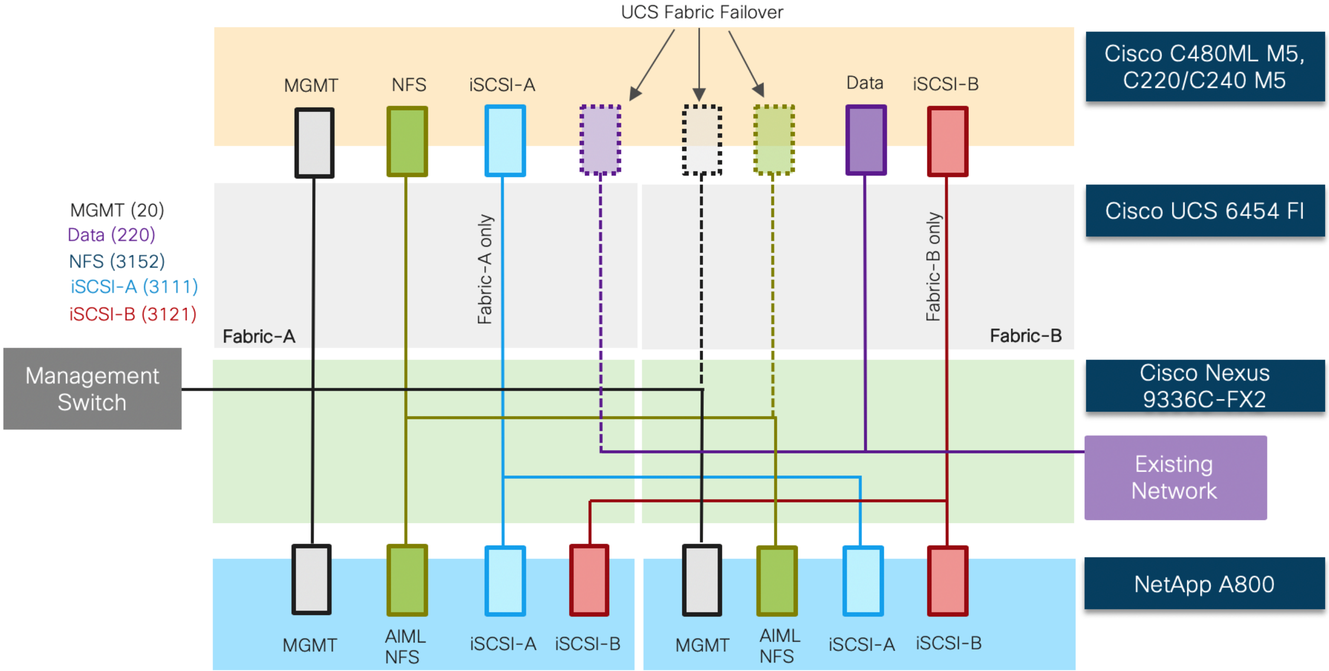
- Managing the ESXi hosts using a common management segment
- Diskless SAN boot using iSCSI with persistent operating system installation for true stateless computing
- Four vNICs where
  - 2 redundant vNICs (Infra-A and infra-B) carry management, vMotion, NFS and virtual machine data traffic VLANs. The MTU value for this interface is set as a Jumbo MTU (9000).
  - 1 iSCSI-A vNIC utilizes iSCSI-A VLAN (defined only on Fabric A) to provide access to iSCSI-A path. The MTU value for this interface is set as a Jumbo MTU (9000).
  - 1 iSCSI-B vNIC utilizes iSCSI-B VLAN (defined only on Fabric B) to provide access to iSCSI-B path. The MTU value for this interface is set as a Jumbo MTU (9000).
- Each ESXi host (blade) accesses NFS datastores hosted on NetApp A800 controllers to be used for deploying virtual machines.
- Each ESXi host allows VMs to access to ImageNet data using the AI-ML-NFS VLAN

The common ESXi service profile enables AI/ML NFS VLAN (3152) on B-Series ESXi hosts as well however the vGPU enabled VMs will only be deployed on C-Series based ESXi hosts.

Service Profile for Bare-Metal (RHEL) Hosts

Like ESXi servers, Cisco UCS Service Profiles are provisioned using Service Profile Templates to allow rapid server deployment with guaranteed configuration consistency. The service profile configuration for a Cisco UCS C-Series host is as shown in Figure 22

Figure 22 Bare-Metal Service Profile



The bare-metal service profile supports:

- Bare-Metal installation of RedHat Enterprise Linux (RHEL) 7.6 with appropriate NVIDIA and CUDA drivers as well as various workload packages (dockers, TensorFlow, and so on).
- Managing the RHEL hosts using a common management segment.
- Cisco UCS Fabric Failover for the vNIC where if one fabric interconnect fails, the surviving fabric interconnect takes over vNIC operations seamlessly. The fabric failover option allows high availability without the need to configure multiple NICs in the host operating system.
- Diskless SAN boot using iSCSI with persistent operating system installation for true stateless computing.
- Five vNICs using Cisco VIC where:
  - 1 management vNIC interface where management VLAN (20) is configured as native VLAN (to avoid VLAN tagging on the RHEL host). The management interface is configured on Fabric A with fabric failover enabled. This vNIC uses standard MTU value of 1500.
  - 1 iSCSI-A vNIC utilizes iSCSI-A VLAN (3111 - defined only on Fabric A) as the native VLAN to provide access to iSCSI-A path. The MTU value for this interface is set as a Jumbo MTU (9000).
  - 1 iSCSI-B vNIC utilizes iSCSI-B VLAN (3121 - defined only on Fabric B) as a native VLAN to provide access to iSCSI-B path. The MTU value for this interface is set as a Jumbo MTU (9000).
  - 1 NFS vNIC interface where NFS VLAN (3152) is configured as native VLAN. The NFS interface is configured on Fabric A with fabric failover is enabled. The MTU value for this interface is set as a Jumbo MTU (9000).
  - (Optional) 1 Data vNIC interface where data traffic VLAN (220) is configured as native VLAN. The Data interface is configured on Fabric B with fabric failover enabled. The MTU value for this interface is set as a Jumbo MTU (9000).

- For handling the high-speed data efficiently, the NFS traffic and the data traffic vNICs (if required) are configured to use separate FIs.

All Cisco UCS C-Series M5 servers including C220, C240 and C480ML equipped with Cisco VIC 1455/1457 will utilize a common service profile for bare-metal deployment

Network Design

Nexus Features

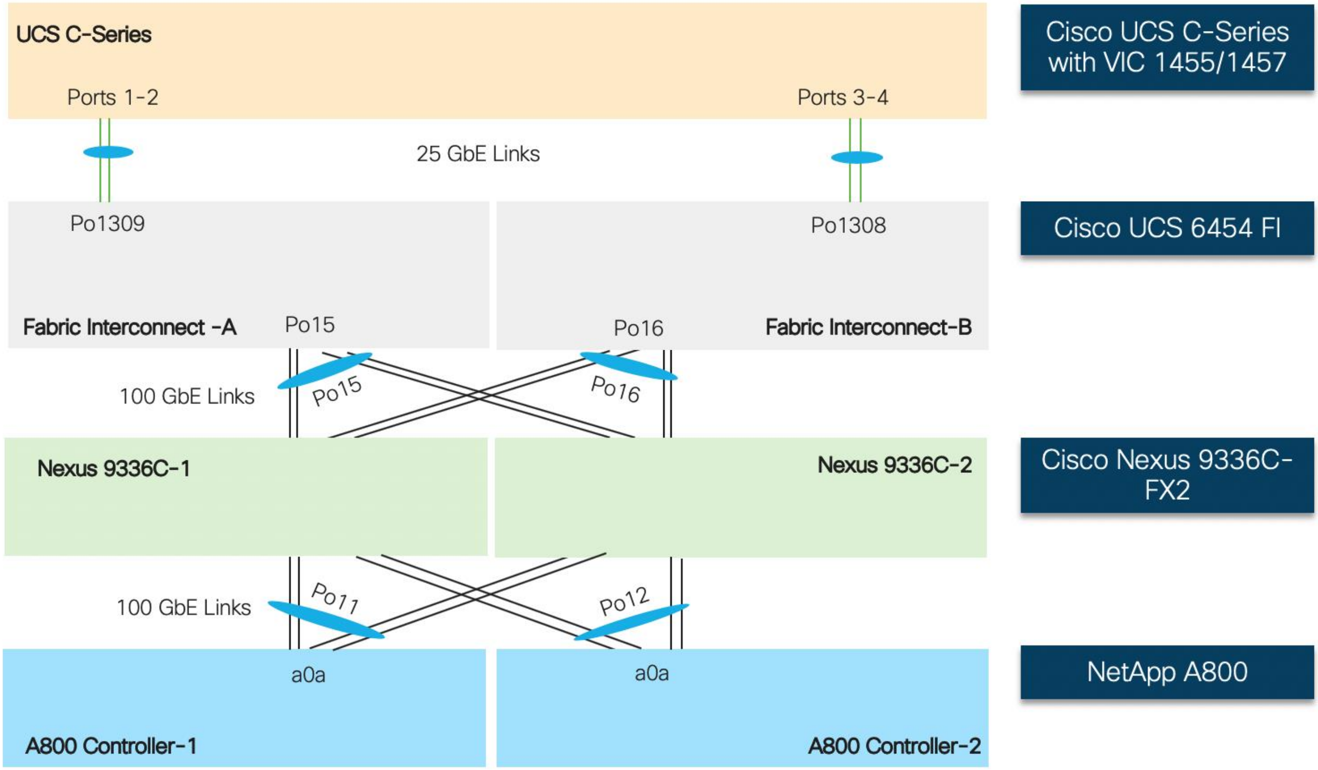
The Nexus 9336C-FX2 device configuration covers the core networking requirements for Layer 2 and Layer 3 communication. Some of the key NX-OS features implemented within the design are:

- Feature interface-vlan – Allows for VLAN IP interfaces to be configured within the switch as gateways.
- Feature HSRP – Allows for Hot Standby Routing Protocol configuration for high availability.
- Feature LACP – Allows for the utilization of Link Aggregation Control Protocol (802.3ad) by the port channels configured on the switch.
- Feature VPC – Virtual Port-Channel (vPC) presents the two Nexus switches as a single “logical” port channel to the connecting upstream or downstream device.
- Feature LLDP - Link Layer Discovery Protocol (LLDP), a vendor-neutral device discovery protocol, allows the discovery of both Cisco and non-Cisco devices.

Cisco UCS C-Series and NetApp A800 Logical Connectivity to Nexus Switches

Figure 23 shows the connectivity between GPU equipped UCS C-Series servers, UCS Fabric Interconnects (FI) and NetApp controllers. Each UCS C-Series server is connected to both the FIs using all 4 25 Gbps VIC interfaces. Port Channels and vPCs (as shown in the figure) are set up for effectively forwarding high speed data. If required, additional links from NetApp A800 and Cisco UCS Fabric Interconnect to Cisco Nexus switches can be deployed for increased bandwidth.

Figure 23 Logical Network Connectivity



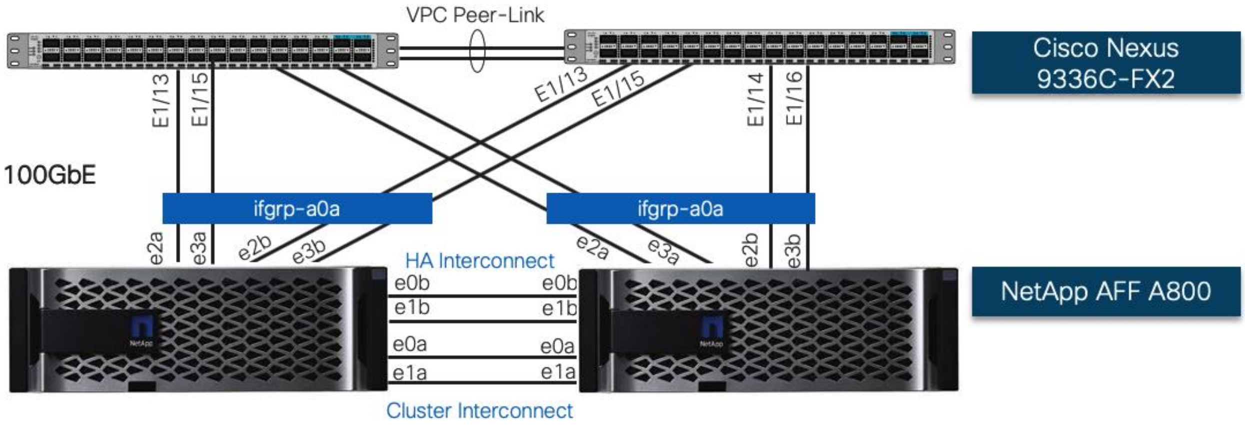
Storage Design

Physical Connectivity

NetApp A800 controllers are connected to Cisco Nexus 9336C-FX2 switches using 100GbE connections. Figure 24 depicts the physical connectivity design of the NetApp AFF A800 system running ONTAP 9.6.

In Figure 24 the two storage controllers in the high availability pair are drawn separately for clarity. Physically, the two controllers exist within a single chassis.

Figure 24 NetApp A800 storage Design



The storage controllers are deployed in a switchless cluster configuration using the onboard ports e0a and e1a. The AFF A800 systems do not have a backplane high availability interconnect and therefore the onboard ports e0b and e1b on both controller nodes were externally connected as the high availability interconnect.

This CVD setup only requires the 100GbE PCIe cards which were installed in slots 2 and 3 of both the controllers according to NetApp best practice recommendations. If additional PCIe cards (for example,10GbE or Fiber Channel) need to be installed in the system, refer to the NetApp Hardware Universe to get the recommended slot information.

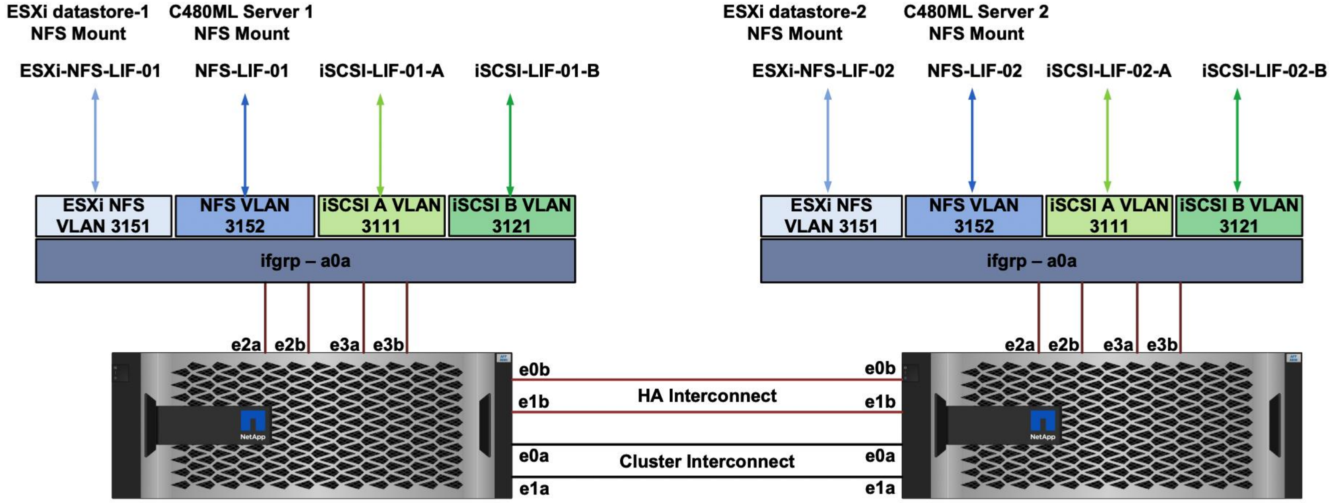
Network Connection to Cisco Nexus 9336C-FX2

A X1146A PCIe3 card is installed on each controller to provide 100GbE network connectivity option. Each controller node includes four 100GbE ports that are bundled together as an interface group (ifgrp) 'a0a' with multimode_lacp. All four ports are active at any given point and with 'multimode_lacp' they can instantly detect link failures and rebalance the traffic on the surviving links, enabling a highly available system with excellent performance.

Multiple VLAN interfaces are created on the ifgrp for NFS and iSCSI data traffic as shown in Figure 25 These interfaces are used as follows:

- iSCSI interfaces are used to provide redundant SAN boot path for stateless OS installation for both hosts in ESXi environment as well as Cisco UCS C480 ML servers.
- ESXi-NFS-LIF interfaces on each controller allow ESXi environment to mount datastores for Virtual Machines.
- NFS-LIF interfaces are used to present the AI/ML dataset FlexGroup volume (imagenet_dataset) to the Cisco UCS C480ML servers as an NFS mount point.

Figure 25   NetApp A800 ifgrp VLANs

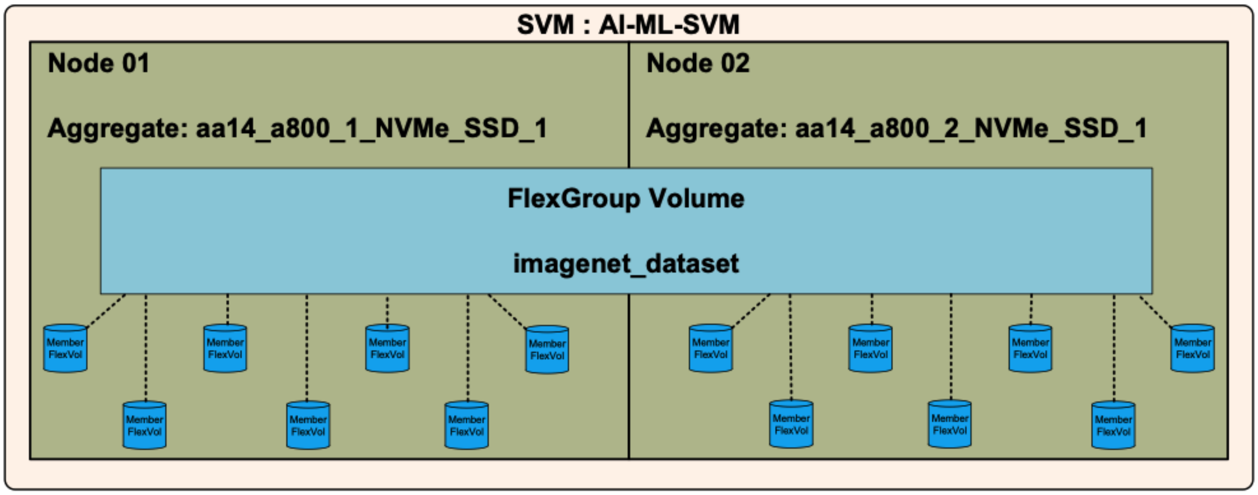


Utilizing separate NFS interfaces for traditional ESXi environment and AI/ML hosts provides path selection flexibility on NetApp controllers equipped with multiple NICs.

Storage Virtual Machine (SVM) Configuration

To provide secure separation of resources and achieve data integrity for the AI/ML environment, a dedicated SVM, AI-ML-SVM, was created, as shown in Figure 26

Figure 26 Storage Virtual Machine (SVM) AI-ML-SVM



The AI-ML-SVM maintained ownership of the storage volumes and networks that were used for the AI-ML workloads and the network interfaces. Customers can use the same SVM for hosting the traditional ESXi environment (as covered in this design) or can choose to define a new SVM for data and path segregation.

FlexGroup Volume

A FlexGroup volume with a capacity of 10TB was created to host the AI/ML (ImageNet) data used in this validation. The FlexGroup volume was created by ONTAP, by automatically creating an equal number of member FlexVol volumes on both the storage controller nodes.

NFS LIF Configuration for AI/ML Workloads

A single NFS LIF is created for each ifgrp on each controller to provide discrete mount points for the FlexGroup volume to both the bare-metal servers and the VMs. This configuration helps distribute the traffic load across the two controllers. During this validation, half the servers (bare metal and VMs) used the NFS LIF created with ‘home-node’ set as Storage Controller 01 to mount the ‘imagenet_dataset’ FlexGroup volume while the remaining half used the NFS LIF created with ‘home-node’ set as storage controller 02 to mount the same volume.

Software Setup and Configuration

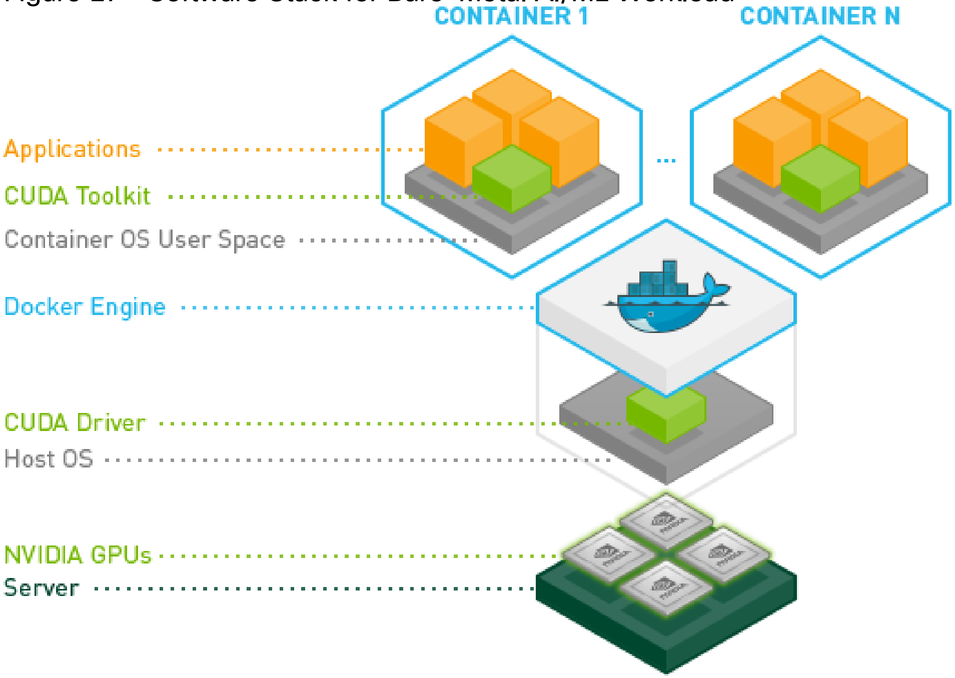
NVIDIA GPU Cloud

Nvidia GPU Cloud (NGC) is the hub for GPU-optimized software for deep learning, machine learning, and high-performance computing (HPC) that takes care of all the software setup and dependencies. NGC software containers can be deployed on bare metal servers or on virtualized environments, maximizing utilization of GPUs, portability, and scalability of applications and provide a range of AI framework container options that meet the needs of data scientists, developers, and researchers.

Bare Metal Server Setup

After setting up the necessary compute, storage and networking components, operating system and various software packages are installed on Cisco UCS C-Series servers to enable the customers to download and run NGC containers. Figure 27 provides a high-level overview of the software stack installation on a bare-metal server:

Figure 27 Software Stack for Bare-Metal AI/ML Workload



To enable the customers to download an AI/ML framework (TensorFlow) from NGC, following installation steps must be completed:

- Download and install the RHEL 7.6 on GPU equipped Cisco C-Series servers.
- Download and install the required Linux packages including gcc, kernel headers, development packages and so on.
- Download and install NVIDIA Driver and CUDA Toolkit
- Download and install NVIDIA Docker 2
- Download and run TensorFlow Container from NGC
- Download and execute the CNN Benchmark Script for ImageNet data hosted on NetApp FlexGroup Volume.

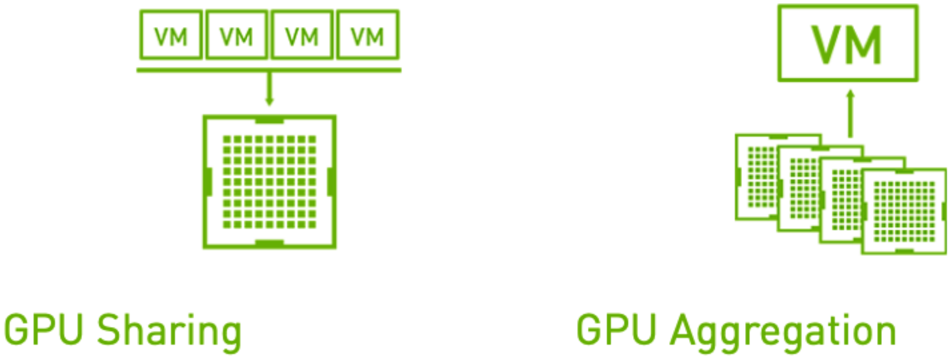
The TensorFlow CNN benchmarks contain implementations of several popular convolutional models such as ResNet, Inception, VGG16, and so on.

The installation instructions and software versions used are explained in detail in the deployment guide.

NVIDIA Virtual Compute Server

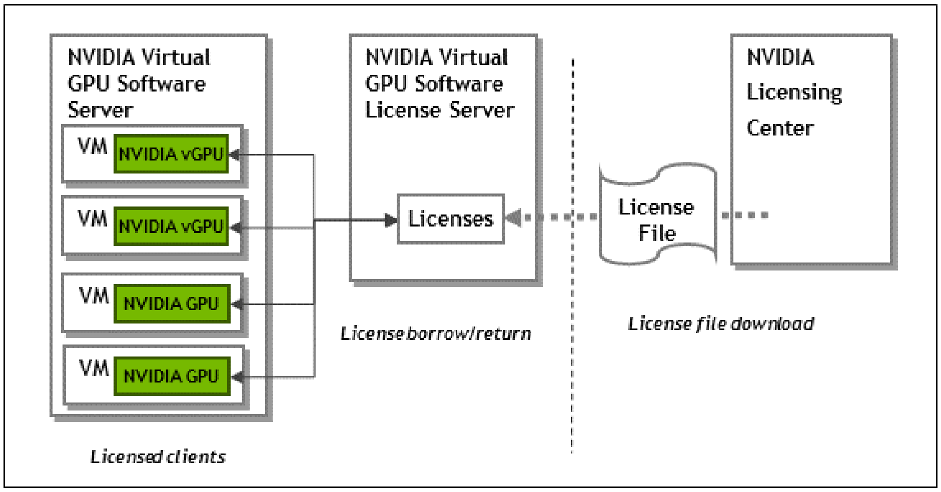
NVIDIA Virtual Compute Server (vComputeServer) enables the benefits of hypervisor-based server virtualization for GPU- accelerated servers so that the most compute-intensive workloads, such as AI and ML can be run in a VM. vComputeServer supports NVIDIA NGC GPU-optimized software and containers. With GPU sharing, multiple VMs can be powered by a single GPU, maximizing utilization. With GPU aggregation, a single VM can be powered by multiple virtual GPUs, making even the most intensive workloads possible.

Figure 28 GPU Sharing and GPU Aggregation



NVIDIA vComputeServer is a licensed product where Virtual GPU (vGPU) functionalities are activated during guest OS boot by the acquisition of a software license served over the network from an NVIDIA vGPU software license server. The license is returned to the license server when the guest OS shuts down.

Figure 29 NVIDIA vGPU Software Architecture



To utilized GPUs in a VM environment, the following configuration steps must be completed:

- Create an NVIDIA Enterprise Account and add appropriate product licenses.
- Deploy a Windows based VM as NVIDIA vGPU License Server and install license file.
- Download and install NVIDIA software on the hypervisor.
- Setup VMs to utilize GPUs.

The installation instructions and software versions used are explained in detail in the deployment guide.

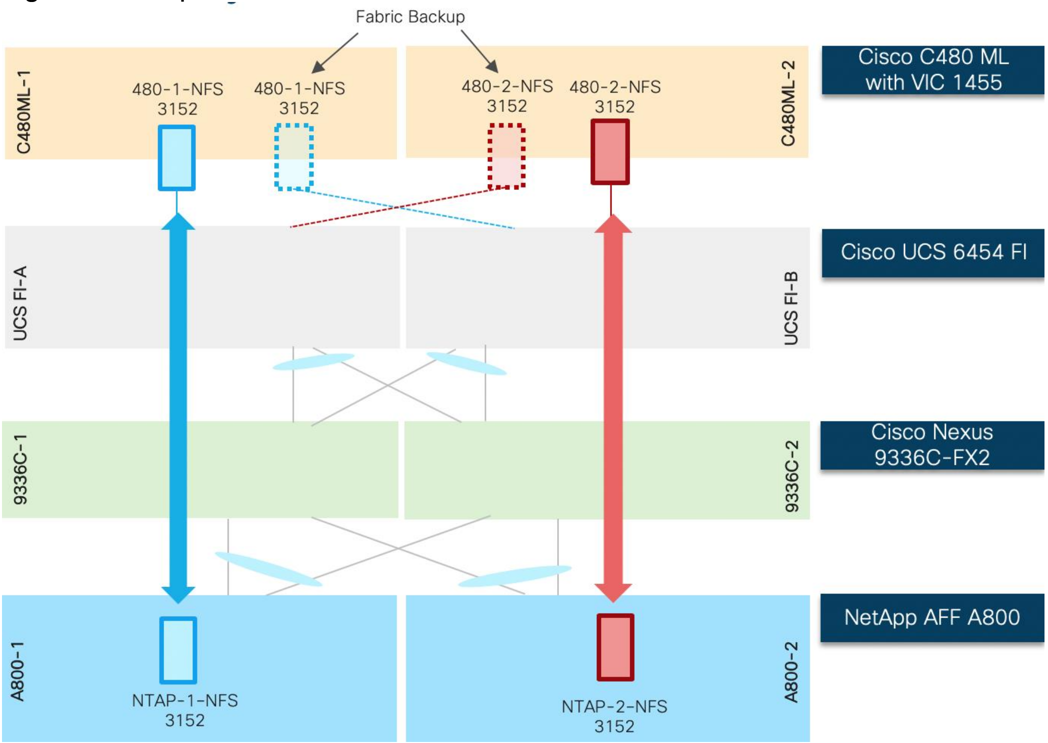
Deployment Considerations

Some of the important deployment considerations for compute, network and storage configurations are discussed below.

Compute Considerations

Typical AI/ML workloads require data access over high-speed network links at a low latency and no packet drops due to network congestion. Since NFS vNIC within a service profile template is tied to a single FI, depending on the number of GPU equipped Cisco UCS C-Series servers in the customer environment, it is possible to congest the network links between a Cisco UCS FI and Cisco Nexus 9336C switches. To mitigate this scenario, it is recommended to deploy two separate service profiles for the AI/ML hosts where both service profiles utilize separate path for NFS data access. The first service profile will utilize an NFS vNIC setup on Fabric-A while the second service profile will utilize the NFS vNIC setup on Fabric-B. By using the appropriate LIFs on the NetApp A800 controllers to mount NFS datastores, the NFS traffic will follow separate paths. This concept is illustrated in Figure 30 for a few Cisco UCS C480ML M5 servers.

Figure 30 Separate Service Profiles to Distribute Traffic Across Different Paths



Network Considerations

Quality of Service (QoS)

Network QoS can help mitigate packet forwarding issues when the interfaces get saturated and start dropping packets. If network congestion is observed in customer environment, implementing QoS policy for NFS and/or data traffic for Cisco UCS hosts is recommended. Implementing QoS in a Cisco UCS requires:

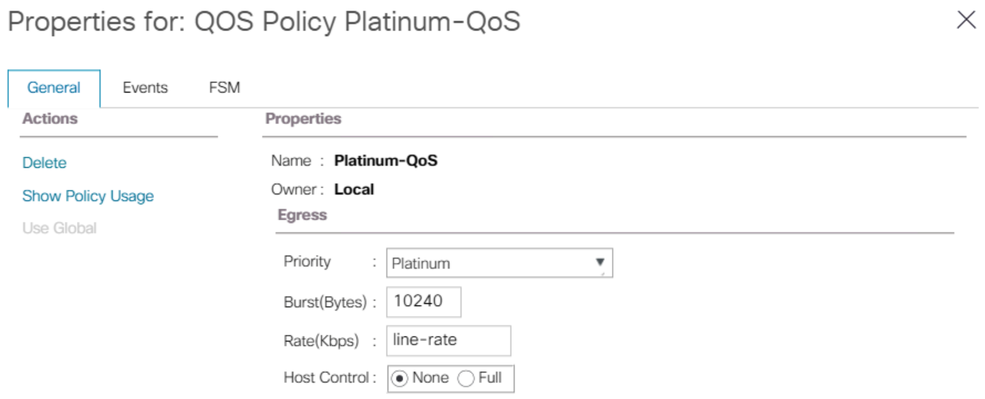
- Enabling the QoS System Class “Platinum” under LAN->Lan Cloud->System QoS Class
- Setting appropriate CoS, weight and MTU settings

- Defining a QoS policy in UCS (as shown in Figure 31
- Applying the QoS policy to appropriate vNIC (as shown in Figure 32

For setting up QoS for the return NFS traffic i.e. traffic from NetApp A800 to the Cisco UCS C480 ML M5, following configuration on the Nexus 9336C-FX2 switches should also be implemented:

- Defining and applying marking policies for NFS traffic from NetApp A800 controllers
- Implementing appropriate QoS configuration to prioritize the marked traffic

Figure 31 QoS Policy on Cisco UCS



Properties for: QOS Policy Platinum-QoS

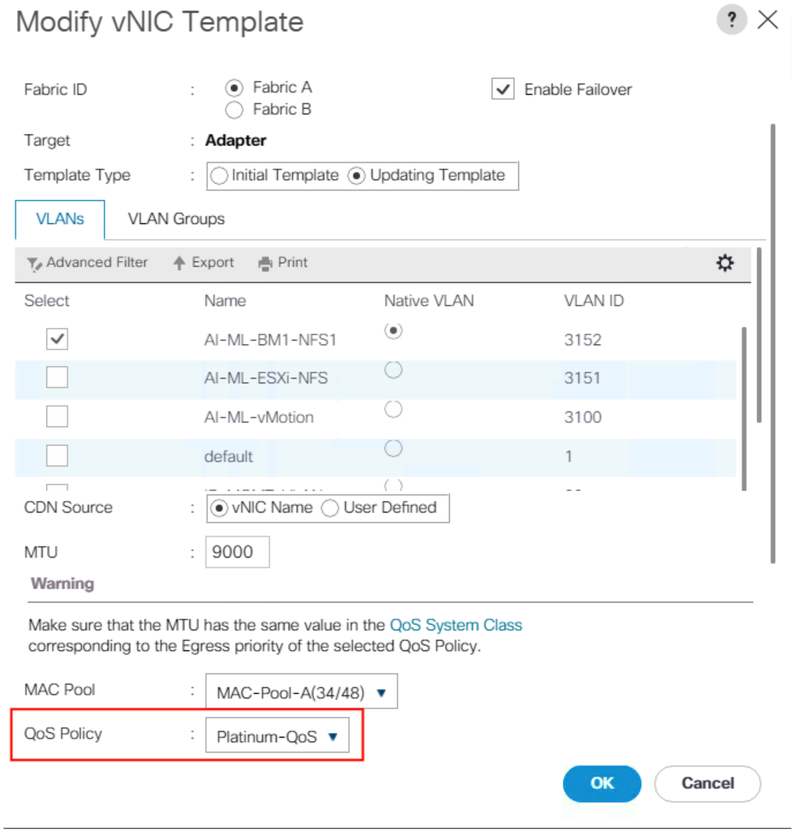
Name : **Platinum-QoS**

Owner : **Local**

Egress

Priority :	Platinum
Burst(Bytes) :	10240
Rate(Kbps) :	line-rate
Host Control :	● None ○ Full

Figure 32 Applying QoS Policy to a vNIC



Modify vNIC Template

Fabric ID :	● Fabric A ○ Fabric B
	☑ Enable Failover
Target :	**Adapter**
Template Type :	○ Initial Template ● Updating Template

Select	Name	Native VLAN	VLAN ID
☑	AI-ML-BM1-NFS1	●	3152
□	AI-ML-ESXi-NFS	○	3151
□	AI-ML-vMotion	○	3100
□	default	○	1

CDN Source :	● vNIC Name ○ User Defined
MTU :	9000

Warning

Make sure that the MTU has the same value in the QoS System Class corresponding to the Egress priority of the selected QoS Policy.

MAC Pool :	MAC-Pool-A(34/48)
QoS Policy :	Platinum-QoS

The QoS policy shown in Figure 31 just an example. Actual QoS policy in a customer environment may be different depending on the traffic profile. When defining the QoS policy on Cisco UCS, an equivalent policy must also be defined on the Cisco Nexus switches to enable end to end QoS.

Storage Considerations

Volume Auto-Grow

The size of the training/learning datasets can grow well beyond the 10TB capacity that is originally configured on the FlexGroup volume. In such a scenario, these volumes can be configured to grow automatically by using the 'volume autosize' command and setting the mode to 'grow'. This would make sure that the capacity of the volume increases when new data is written to it if the underlying aggregates can supply more space.

FlexGroup volumes can also be configured to auto shrink when used with autogrow. Automatic shrinking prevents a volume from being larger than required and frees up space in the aggregate for use by other volumes.

```
volume autosize -vserver vserver_name -volume vol_name -mode [grow | grow_shrink]
```

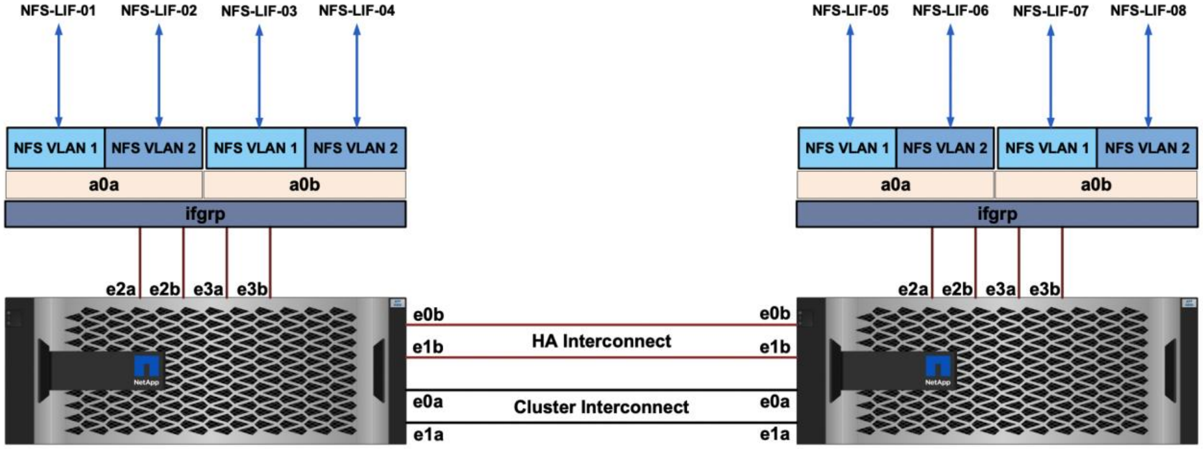
NFS LIFs

If the Cisco UCS GPU equipped Cisco UCS C-Series server count grows beyond a couple of servers, the storage network design can be modified as follows to provide additional discreet NFS mount points for the servers.

1. Create two ifgrps on each controller node (instead of one shown in an earlier design) by using two physical 100GbE ports per ifgrp.
2. Introduce an additional NFS VLAN by creating the VLAN interfaces on all the ifgrps.
3. Create another broadcast domain for the newly created VLAN and add the member VLAN interfaces to it.
4. Create NFS LIFs that use the newly created VLAN interfaces on the ifgrps as their home ports.

Figure 33 illustrates the new design after this modification.

Figure 33 NetApp AFF A800 Scale Considerations



With the proposed changes, eight NFS LIFs are now available. LIFs from both controllers across both ifgrps and VLANs can be chosen so that the data traffic is well balanced across all available paths.

NVIDIA Software Deployment Considerations

Licensing Server for vGPU support

To setup a standalone license server for vGPU licensing requirements, a windows server 2012 VM with the following software and hardware parameters was setup for this deployment:

- 2 vCPUs
- 4GB RAM
- 100GB HDD
- 64-bit Operating System
- Static IP address
- Internet access
- Latest version of Java Runtime Environment

Setting the ESXi Host Graphics to SharedPassthru

In a VMware environment, a GPU can be configured in shared virtual graphics mode or the vGPU (SharedPassthru) mode. For the AI/ML workloads, the NVIDIA card should be configured in the SharedPassthru mode.

VM Setup Requirements for vGPU support

Using the vGPUs for AI/ML workloads in a VM has some VM setup restrictions in an ESXi environment. The following VM considerations must be considered when deploying a vGPU enabled VM:

- The guest OS must be a 64-bit OS.
- 64-bit MMIO and EFI boot must be enabled for the VM.
- The guest OS must be able to be installed in EFI boot mode.
- The VM’s MMIO space must be increased to 64 GB (refer to VMware KB article: https://kb.vmware.com/s/article/2142307). When using multiple vGPUs with single VM, this value might need to be increased to match the total memory for all the vGPUs.
- To use multiple vGPUs in a VM, set the VM compatibility to vSphere 6.7 U2.

Deployment Hardware and Software

Hardware and Software Revisions

Table 2 Hardware and Software Revisions

Component		Software
Network	Nexus 9336C-FX2	7.0(3)I7(6)
Compute	Cisco UCS Fabric Interconnect 6454	4.0(4e)
	Cisco UCS C-Series M5 Servers	4.0(4e)
	VMware ESXi	6.7U3
	ESXi ENIC Driver	1.0.29.0
	VMware vCenter Appliance	6.7U3
	Red Hat Enterprise Linux (RHEL)	7.6
	RHEL ENIC driver	3.2.210.18-738.12
	NVIDIA driver for RHEL	418.40.04
	NVIDIA driver for ESXi	430.46
	NVIDIA CUDA Toolkit	10.1 Update 2
Storage	NetApp A800	9.6
	NetApp NFS Plugin for VMware VAAI	1.1.2-3
	NetApp Virtual Storage Console	9.6

Validation

FlexPod Datacenter for AI/ML with Cisco UCS 480 ML is validated for successful infrastructure configuration and availability using a wide variety of test cases and by simulating partial and complete device and path failure scenarios. The types of tests executed on the system (at a high level) are listed below:

- Use TensorFlow with the Imagenet dataset and execute various test models including ResNet-152, ResNet-50, VGG16, Inception v2 and v3 to observe the GPU and Storage performance.
- Validate Cisco UCS C480 ML platform can be successfully deployed and managed using Cisco UCS Manager in both bare-metal (RHEL) and virtualized (ESXi) configuration.
- Validate creation, deletion, and re-attachment of the service profiles for the Cisco UCS C480 ML platform.
- Validate the GPU functionality for bare-metal servers and vGPU functionality in the VMware environment.
- Validate vGPU functionality where multiple VMs use the same GPU (shared) as well as single VM uses multiple GPUs (performance).
- Validate vMotion for VMs with vGPUs.
- Generate traffic using IOMeter instances as well as AI/ML workloads in parallel to verify the connectivity, bandwidth utilization, and network usage.
- Validate path, network, compute and storage device failures while workloads are running.

Summary

Artificial Intelligence (AI) and Machine Learning (ML) initiatives have seen a tremendous growth due to the recent advances in GPU computing technology. The FlexPod Datacenter for AI/ML with Cisco UCS 480 ML solution aims to deliver a seamless integration of the Cisco UCS GPU enabled Cisco UCS C-Series platforms including Cisco UCS C480 ML M5 into the current FlexPod portfolio to enable the customers to easily utilize the platform's extensive GPU capabilities for their AI/ML workloads without requiring extra time and resources for a successful deployment.

The validated solution achieves the following core design goals:

- Optimized integration of Cisco UCS C-Series including Cisco UCS C480 ML M5 platform into the FlexPod design.
- Integration of a NetApp A800 NVMe based storage system into the FlexPod architecture for AI/ML.
- Showcase AI/ML workload acceleration using NVIDIA V100 32GB and NVIDIA T4 16GB GPUs.
- Support for Cisco UCS C220 M5 and Cisco C240 M5 with NVIDIA GPUs for inferencing and low intensity workloads.
- Showcasing NVIDIA vCompute Server functionality for the AI/ML workloads in VMware environment.
- NetApp FlexGroup volumes with ONTAP 9.6.

References

Products and Solutions

Cisco Unified Computing System:

http://www.cisco.com/en/US/products/ps10265/index.html

Cisco UCS 6454 Fabric Interconnects:

https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/datasheet-c78-741116.html

Cisco UCS 5100 Series Blade Server Chassis:

http://www.cisco.com/en/US/products/ps10279/index.html

Cisco UCS B-Series Blade Servers:

http://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-b-series-blade-servers/index.html

Cisco UCS C480 ML M5 Rack Server:

https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/datasheet-c78-741211.html

Cisco UCS VIC 1400 Adapters:

https://www.cisco.com/c/en/us/products/collateral/interfaces-modules/unified-computing-system-adapters/datasheet-c78-741130.html

Cisco UCS Manager:

http://www.cisco.com/en/US/products/ps10281/index.html

NVIDIA GPU Cloud

https://www.nvidia.com/en-us/gpu-cloud/

NVIDIA vComputeServer

https://www.nvidia.com/en-us/data-center/virtual-compute-server/

Cisco Nexus 9336C-FX2 Switch:

https://www.cisco.com/c/en/us/support/switches/nexus-9336c-fx2-switch/model.html

VMware vCenter Server:

http://www.vmware.com/products/vcenter-server/overview.html

NetApp Data ONTAP:

http://www.netapp.com/us/products/platform-os/ontap/index.aspx

NetApp AFF A800:

https://www.netapp.com/us/products/storage-systems/all-flash-array/aff-a-series.aspx

Interoperability Matrixes

Cisco UCS Hardware Compatibility Matrix:

https://ucshcltool.cloudapps.cisco.com/public/

VMware Compatibility Guide:

http://www.vmware.com/resources/compatibility

NetApp Interoperability Matric Tool:

http://mysupport.netapp.com/matrix/

About the Authors

Haseeb Niazi, Technical Marketing Engineer, Cisco Systems, Inc.

Haseeb Niazi has over 20 years of experience at Cisco in the Datacenter, Enterprise and Service Provider Solutions and Technologies. As a member of various solution teams and Advanced Services, Haseeb has helped many enterprise and service provider customers evaluate and deploy a wide range of Cisco solutions. As a technical marking engineer at Cisco UCS Solutions group, Haseeb focuses on network, compute, virtualization, storage and orchestration aspects of various Compute Stacks. Haseeb holds a master's degree in Computer Engineering from the University of Southern California and is a Cisco Certified Internetwork Expert (CCIE 7848).

Arvind Ramakrishnan, Solutions Architect, NetApp, Inc.

Arvind Ramakrishnan works for the NetApp Infrastructure and Cloud Engineering team. He focusses on development, validation and implementation of Cloud Infrastructure solutions that include NetApp products. Arvind has more than 10 years of experience in the IT industry specializing in Data Management, Security, Cloud and Datacenter technologies. Arvind holds a bachelor’s degree in Electronics and Communication.

Acknowledgements

For their support and contribution to the design, validation, and creation of this Cisco Validated Design, the authors would like to thank:

- John George, Technical Marketing Engineer, Cisco Systems, Inc.