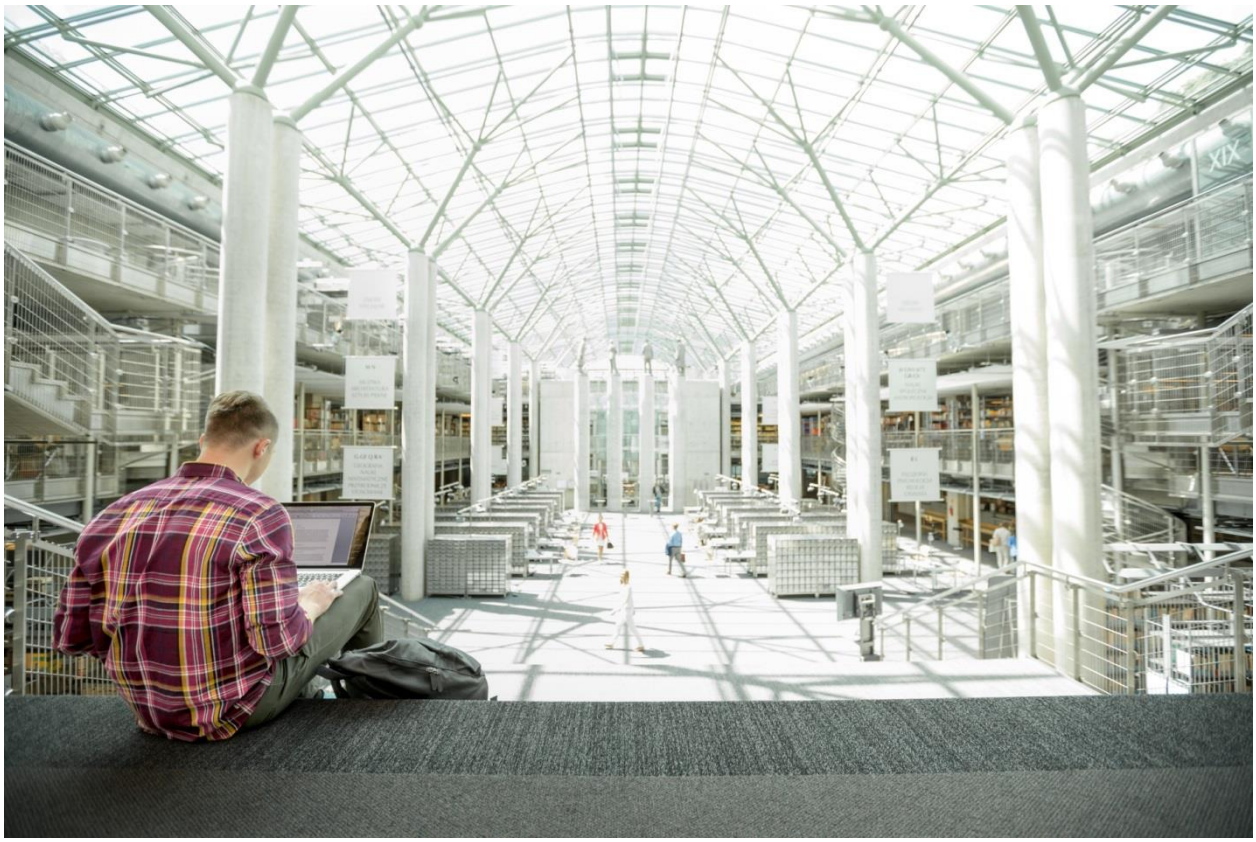


FlashStack® for AI: Powering the Data Pipeline Design Guide

Published: January 16, 2020

CISCO VALIDATED DESIGN

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.

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 data center 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 Pure Storage have partnered to deliver a series of FlashStack™ solutions that enable strategic data center platforms with the above characteristics. FlashStack solution delivers a modern converged infrastructure (CI) solution that is smarter, simpler, smaller, and more efficient. With FlashStack, customers can modernize their operational model, stay ahead of business demands, and protect and secure their applications and data, regardless of the deployment model on premises, at the edge, or in the cloud. FlashStack's fully modular and non-disruptive architecture abstracts hardware into software for non-disruptive changes which allow customers to seamlessly deploy new technology without having to re-architect their data center solutions.

Artificial Intelligence (AI) and Machine Learning (ML) initiatives have seen tremendous growth due to the recent advances in GPU computing technology. 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 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 infrastructure 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 GPU equipped Cisco UCS C-Series M5 platforms and Pure Storage FlashBlade into the FlashStack solution to deliver a unified approach for providing AI and ML capabilities within the converged infrastructure. By offering customers the ability to manage the AI/ML servers with the familiar tools they use to administer traditional FlashStack systems, the administrative overhead, as well as the cost of deploying deep learning platform, is significantly reduced. The design presented in this CVD covers Cisco UCS C480 ML M5 server with 8 NVIDIA V100 32GB SXM2 GPUs, Cisco UCS C220 M5 server with 2 NVIDIA T4 GPUs and Cisco UCS C240 M5 server with 2 NVIDIA V100 32GB PCIe GPUs as various options for AI/ML workloads.

Solution Overview

Introduction

Building an AI-platform with off-the-shelf hardware and software components leads to solution complexity and eventually stalled initiatives. Valuable months are lost in IT resources 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.

FlashStack is a pre-designed, integrated and validated architecture for data center that combines Cisco UCS servers, Cisco Nexus family of switches, Cisco MDS fabric switches and Pure Storage Arrays into a single, flexible architecture. FlashStack 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. The FlashStack for AI solution aims to deliver seamless integration of the Cisco UCS C480 ML M5 and other GPU equipped Cisco UCS C-Series platforms into the current FlashStack portfolio to enable the customers to successfully deploy and efficiently utilize the platforms' extensive GPU capabilities for their workloads. FlashStack design can support different hypervisor options, bare metal servers and can also be sized and optimized based on various workload requirements.

The FlashStack design discussed in this document has been validated for resiliency and fault tolerance during system upgrades, component failures, and partial as well as complete 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 FlashStack from previous models:

- Optimized integration of Cisco UCS C480 ML M5 platform into the FlashStack design
- Integration of Pure FlashBlade 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 T4 GPUs on Cisco UCS C220 M5 platform.
- Showcase NVIDIA Virtual Compute Servers (vComputeServer) and Virtual GPU (vGPU) capabilities on various UCS platforms.
- Support for Intel 2nd Gen Intel Xeon Scalable Processors (Cascade Lake) processors*.

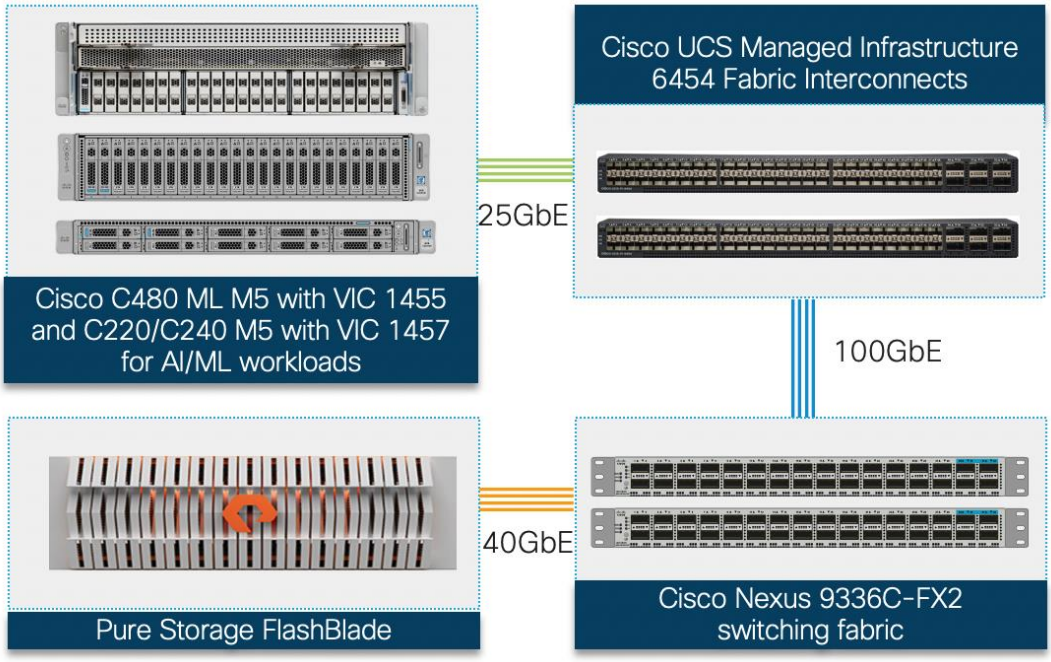
*** The Cisco UCS software version 4.0(4e) (explained in this validation) and RHEL 7.6 support Cascade Lake CPUs on Cisco UCS C220 M5 and C240 M5 servers. Support for Cisco UCS C480 ML M5 will be available in the upcoming Cisco UCS release.**

For more information about previous FlashStack designs, see: https://www.cisco.com/c/en/us/solutions/design-zone/data-center-design-guides/data-center-design-guides-all.html#FlashStack

Solution Summary

In the FlashStack for AI solution, Cisco UCS C480 ML M5 computing platform brings massive GPU acceleration close to the data stored within the FlashStack infrastructure. Just like the other Cisco UCS blade and rack servers in the FlashStack 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 Pure Storage FlashBlade providing storage access using high-speed redundant paths. With this integrated approach, customers reap the benefits of a consistent, easily managed architecture. Using the Cisco UCS Manager, 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 FlashStack for AI and Deep Learning



Like other FlashStack designs, FlashStack for AI solution 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 FlashStack system or the Cisco UCS servers) or scale-out (by adding more FlashStack instances). The validated design presented in this document combines a proven combination of technologies that allow customers to extract more intelligence out of all stages of their data lifecycle.

Technology Overview

FlashStack Solution Overview

FlashStack is a validated reference architecture developed by Cisco and Pure Storage to serve enterprise data centers. The infrastructure is built leveraging:

- Cisco Unified Computing System (Cisco UCS)
- Cisco Nexus and Cisco MDS* Switches
- Pure Storage Systems (FlashArray and FlashBlade)

* **Cisco MDS switches are only needed when the solutions require Fiber Channel connectivity**

These components are connected and configured according to the best practices of both Cisco and Pure Storage and provide an ideal platform for running a variety of workloads with confidence. As illustrated in Figure 1 , the current solution comprises of the following core components:

- Cisco UCS Manager on Cisco 4th generation 6454 Fabric Interconnects to support 10GbE, 25GbE and 100 GbE connectivity for various components.
- Cisco UCS C220 M5, C240 M5 and C480 ML M5 server with 8 NVIDIA Tesla V100-32GB GPUs for AI/ML applications.
- High-Speed Cisco NX-OS based Nexus 9336C-FX2 switching design supporting up to 100GbE connectivity.
- Pure Storage FlashBlade providing scale-out, all-flash storage purpose built for massive concurrency as needed for AI/ML workloads.
- (Optional) Pure Storage FlashArray for setting up FlashStack Virtual Server Infrastructure* as covered in: https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/ucs_flashstack_vsi_vm67_u1_design.html .

*** The VMware environment is required for deploying the NVIDIA vCompute Server functionality on the GPU enabled Cisco UCS C-series platforms configured as ESXi nodes.**

Cisco Unified Computing System

Cisco Unified Computing System™ (Cisco UCS) is a next-generation data center 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 in turn 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. 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 FIs integrate all components into a single, highly-available management domain controlled by the Cisco UCS Manager.

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. In this CVD, VIC 1455 was installed in the Cisco UCS C480 ML M5 and Cisco VIC 1457 was installed in Cisco UCS C220 and C240 M5.

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 vNICs or vHBAs.

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. Like Cisco VIC 1455, this card can also present PCIe standards-compliant interfaces to the host which can be dynamically configured as either vNICs or vHBAs.

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 3). 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 or 100G 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 3 Cisco UCS C480 ML M5 Server

Front View

Rear View

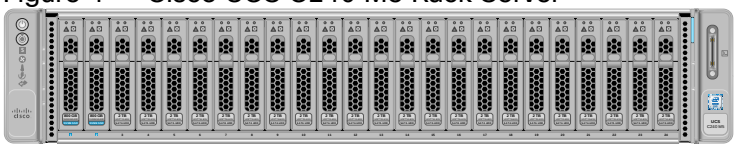
Cisco UCS C240 M5

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 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 5 Cisco UCS C220 M5 rack server

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 data centers deliver breakthrough performance with fewer servers at a lower cost. This validated design covers following NVIDIA GPUs:

NVIDIA Testa V100 32GB

NVIDIA Tesla V100 32GB, is an advanced data center 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. A single Tesla V100 SMX2 GPU supports up to 6 NVLINK connections which translates to a total

bandwidth of 300 GB/sec per GPU. The NVLINK connectivity is not supported on Cisco UCS C240 M5 and the platform 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 data centers 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, data center admins can use the same management tools for their GPU-accelerated servers as they do for the rest of their data center.

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 data center 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 data centers.

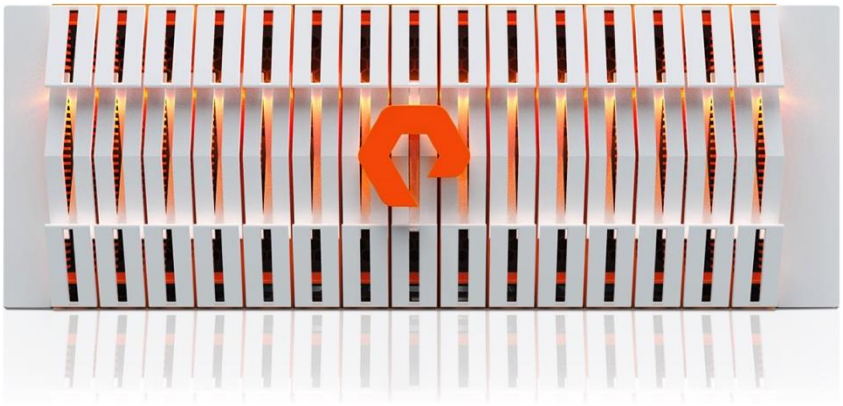
The Cisco Nexus 9336C-FX2 Switch is a 1RU switch that supports 36 ports, 7.2 Tbps of bandwidth and over 2.8 bpps. The switch can be configured to work as 1/10/25/40/100-Gbps offering flexible options in a compact form factor. Breakout is supported on all ports.

Figure 9 Nexus 9336C-FX2 Switch

Pure FlashBlade Systems

FlashBlade™ is a new, innovative scale-out storage system designed to accelerate modern analytics applications while providing best-of-breed performance in all dimensions of concurrency – including IOPS, throughput, latency, and capacity. FlashBlade is as simple as it is powerful, offering elastic scale-out storage services at every layer alongside DirectFlash™ technology for global flash management.

Figure 10 Pure Storage FlashBlade



Purpose-built for Modern Analytics

FlashBlade is the industry's first cloud-era flash purpose-built for modern analytics, delivering unprecedented performance for big data applications. Its massively distributed architecture enables consistent performance for all analytics applications using NFS, S3/Object, SMB, and HTTP protocols. Figure 11 and Figure 12 outline various advantages and differentiators of the FlashBlade system.

Figure 11 Pure Storage FlashBlade Advantage

FAST

- Elastic performance that grows with data, up to 17 GB/s
- Always-fast, from small to large files
- Massively parallel architecture from software to flash

BIG

- Petabytes of capacity
- Elastic concurrency, up to 10s of thousands of clients
- 10s of billions of objects and files

SIMPLE

- Evergreen™ – don't rebuy TBs you already own
- "Tuned for Everything" design, no manual optimizations required
- Scale-out everything instantly by simply adding blades

Figure 12 The Pure Storage FlashBlade Difference

BLADE

SCALE-OUT DIRECTFLASH + COMPUTE

Ultra-low latency, 8, 17, and 52TB capacity options that can be hot-plugged into the system for expansion and performance

PURITY//FB

SCALE-OUT STORAGE SOFTWARE

The heart of FlashBlade, implementing its scale-out storage capabilities, services, and management

FABRIC

SOFTWARE- DEFINED NETWORKING

Includes a built in 40Gb Ethernet fabric, providing a total network bandwidth of 320Gb/s for the chassis

FlashBlade delivers industry-leading throughput, IOPs, latency, and capacity with up to 20 times less space and 10 times less power and cooling.

Purity for FlashBlade (Purity//FB)

FlashBlade is built on the scale-out metadata architecture of Purity for FlashBlade, capable of handling 10s of billions of files and objects while delivering maximum performance, effortless scale, and global flash management. The distributed transaction database built into the core of Purity means storage services at every layer are elastic: simply adding blades grows system capacity and performance, linearly and instantly. Purity//FB supports S3-compliant object store, offering ultra-fast performance at scale. It also supports File protocol including NFSv3 and SMB, and offers a wave of new enterprise features, like snapshots, LDAP, network lock management (NLM), and IPv6, to extend FlashBlade into new use cases.

Pure1®

Pure1, the cloud-based management, analytics, and support platform, expands the self-managing, plug-n-play design of Pure all-flash arrays with the machine learning predictive analytics and continuous scanning of Pure1 Meta™ to enable an effortless, worry-free data platform.

Pure1 Manage

Pure1 Manage is SaaS-based, allowing customers to manage their array from any browser or from the Pure1 Mobile App – with nothing extra to purchase, deploy, or maintain. From a single dashboard, customers can manage all their arrays, with full visibility on the health and performance of their storage.

Pure1 Analyze

Pure1 Analyze delivers true performance forecasting – giving customers complete visibility into the performance and capacity needs of their arrays – now and in the future. Performance forecasting enables intelligent consolidation and unprecedented workload optimization.

Pure1 Support

Pure combines an ultra-proactive support team with the predictive intelligence of Pure1 Meta to deliver unrivaled support that's a key component in our proven FlashArray 99.9999% availability. Customers are often surprised and delighted when we fix issues they did not even know existed.

Pure1 META

The foundation of Pure1 services, Pure1 Meta is global intelligence built from a massive collection of storage array health and performance data. By continuously scanning call-home telemetry from Pure's installed base, Pure1 Meta uses machine learning predictive analytics to help resolve potential issues and optimize workloads. The result is both a white glove customer support experience and breakthrough capabilities like accurate performance forecasting.

Evergreen™ Storage

The Evergreen™ Storage ownership model operates like SaaS and the cloud. Deploy storage and benefit from a subscription to continuous innovation as you expand and improve performance, capacity, density, and/or features for 10 years or more - all without downtime, performance impact, or data migrations. Evergreen Storage provides expandability and upgradability for generations via its modular, stateless architecture, while FlashBlade's blade-based design delivers the linear scale of DirectFlash technology and compute simply by adding blades.

FlashBlade for Artificial Intelligence

In any large-scale deep learning full-stack solution, a storage system must be able to satisfy following three requirements:

- Diverse Performance: Deep learning often requires multi-gigabytes-per-second I/O rates but isn't restricted to a single data type or I/O size. Training deep neural network models for applications as diverse as machine vision, natural-language processing, and anomaly detection requires different data types and dataset sizes.
- Scalable Capacity: Successful machine learning projects often have ongoing data acquisition and continuous training requirements, resulting in a continued growth of data over time. Furthermore, enterprises that succeed with one AI project find ways to apply these powerful techniques to new application areas, resulting in further data expansion to support multiple use cases.
- Strong Resiliency: As the value of AI grows within an organization, so does the value of the infrastructure supporting its delivery. Storage systems that result in excessive downtime or require extensive administrative outages can cause costly project delays or service disruptions.

Pure Storage FlashBlade, with its scale-out, all-flash architecture and a distributed file system purpose-built for massive concurrency across all data types, is the leading storage system to deliver on all these dimensions, while keeping required configuration and management complexity to a bare minimum.

Solution Design

The FlashStack for AI solution focuses on the integration of the Cisco UCS GPU enabled platforms into FlashStack to provide GPU intensive artificial intelligence and machine learning capabilities in the converged infrastructure. The key requirements and design details to deliver this new design are outlined in this section.

Requirements

The solution closely aligns with latest NxOS based FlashStack 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.

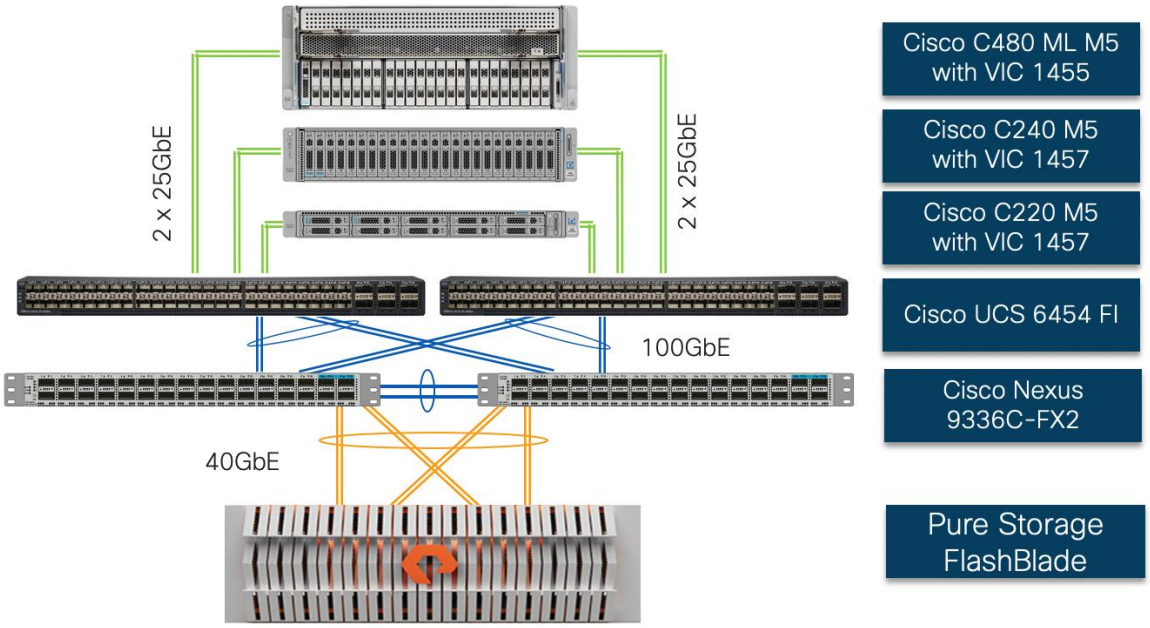
Additionally, for Cisco UCS C480 ML M5 platform integration into FlashStack, following specific design considerations are also observed:

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. 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 node failure.
3. Automatic load balancing and parallelized data access and scaling across switching architecture to enable AI/ML platform to efficiently access AI/ML training and inference dataset from the Pure Storage Flash-Blade.
4. Ability to utilize the GPU capabilities in the VMware environment where multiple Virtual Machines (VMs) can share a GPU as well as a single VM can utilize more than one GPU.

Physical Topology

The physical topology for the connecting Cisco UCS C-series M5 platforms to a Pure Storage FlashBlade using a Cisco UCS 6454 Fabric Interconnect and Nexus 9336C-FX2 switch is shown in Figure 13 :

Figure 13 FlashStack for Deep Learning - Physical Topology

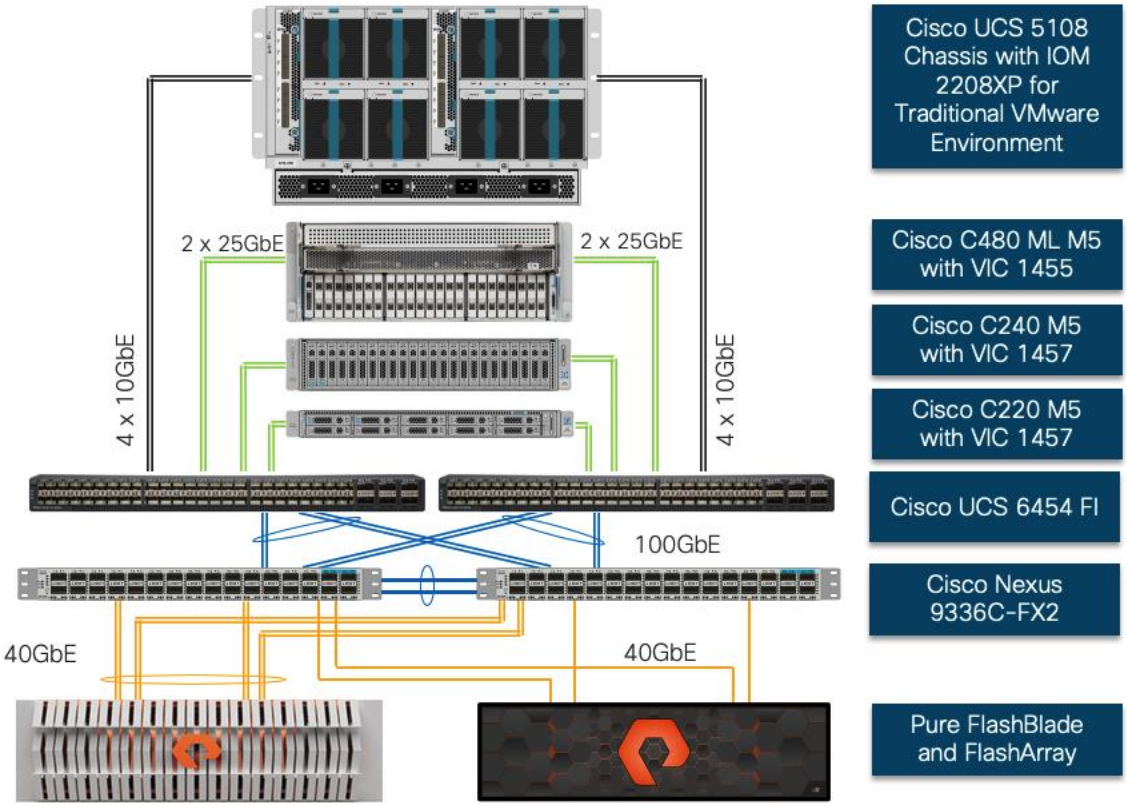


To validate the Cisco UCS C480 ML M5 and (GPU equipped) Cisco UCS C220/C240 M5 platforms addition to FlashStack, following components were setup to support both VM and bare metal AI/ML servers:

- Cisco UCS 6454 Fabric Interconnects (FI) is used to connect and manage Cisco UCS C-Series M5 servers.
- Cisco UCS C480 ML M5 is connected to each FI using Cisco VIC 1455. The server is connected to each FI using 2 x 25GbE connections configured as port-channels.
- Cisco UCS C220 M5 and UCS C240 M5 are connected to each FI using Cisco VIC 1457. The servers are connected to each FI using 2 x 25GbE connections configured as port-channels.
- Cisco Nexus 9336C configured in NxOS mode provides the switching fabric.
- Cisco UCS 6454 FI's 100GbE uplink ports, configured as port-channels, are connected to Nexus 9336C.
- Pure Storage FlashBlade is connected to Nexus 9336C switch using 40GbE ports, configured as a port-channel.

The design illustrated in Figure 13 allows customers to seamlessly integrate their traditional FlashStack for Virtual Machine Infrastructure with this new AI/ML configuration. The resulting physical topology is shown in Figure 14 where Cisco UCS 6454 FIs connects not only to GPU equipped Cisco UCS C-series servers, but also connect to Cisco UCS 5108 chassis containing Cisco UCS B200 M5 blades. The Nexus 9336C-FX2 platform provides additional connectivity to Pure Storage FlashArray for supporting the virtual machine infrastructure. The design shown in Figure 14 supports iSCSI connectivity option for the FlashStack Virtual Machine Infrastructure but can be extended to support FC connectivity design by utilizing Cisco MDS switches.

Figure 14 Integration of FlashStack for Virtual Machine Infrastructure and Deep Learning Platforms



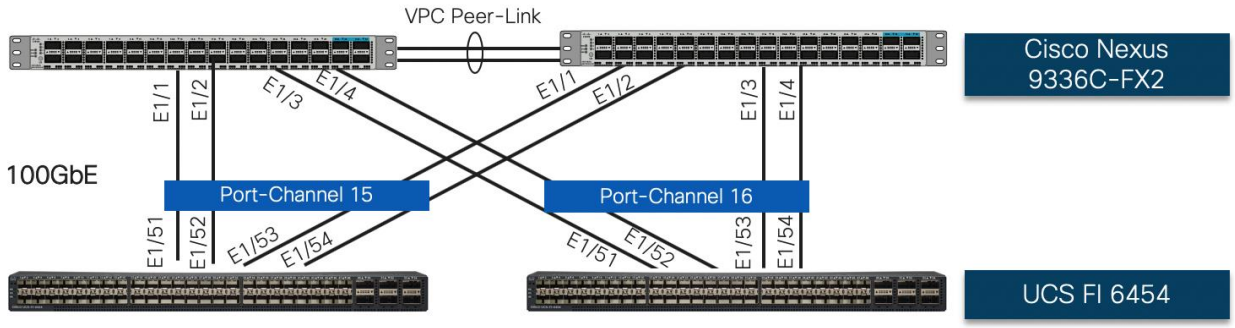
This design guide focuses on the components and design options for setting up the machine learning solution as shown in Figure 13 . For Virtual Machine Infrastructure FlashStack design details, refer to FlashStack CVDs at the following location: https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/ucs_flashstack_vsi_vm67_u1_design.html

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 15 . 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 with two separate vPCs, one for each FI.

Figure 15 Cisco UCS 6454 FI to Nexus 9336C Connectivity



Cisco UCS C220 M5 Connectivity

To manage the Cisco UCS C220 M5 platform with dual NVIDIA T4 GPUs using Cisco UCS Manager, the C220 M5 is connected to the Fabric Interconnects (FIs) using 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 16 . The ports connected to an FI form a port-channel providing an effective 50GbE bandwidth to each fabric interconnect.

Figure 16 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 Cisco UCS C240 M5 connects to the FIs using Cisco VIC 1457. Cisco VIC 1457 ports are connected to the Cisco UCS 6454 FI in pairs such that ports 1 and 2 are connected to the FI-A and the ports 3 and 4 are connected to the FI-B as shown in Figure 17 . The ports connected to an FI form a port-channel providing an effective 50GbE connectivity to each fabric interconnect.

Figure 17 Cisco UCS C240 M5 to Cisco UCS 6454 FI Connectivity

Manage 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 ML M5 is connected to the FIs using Cisco VIC 1455.

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

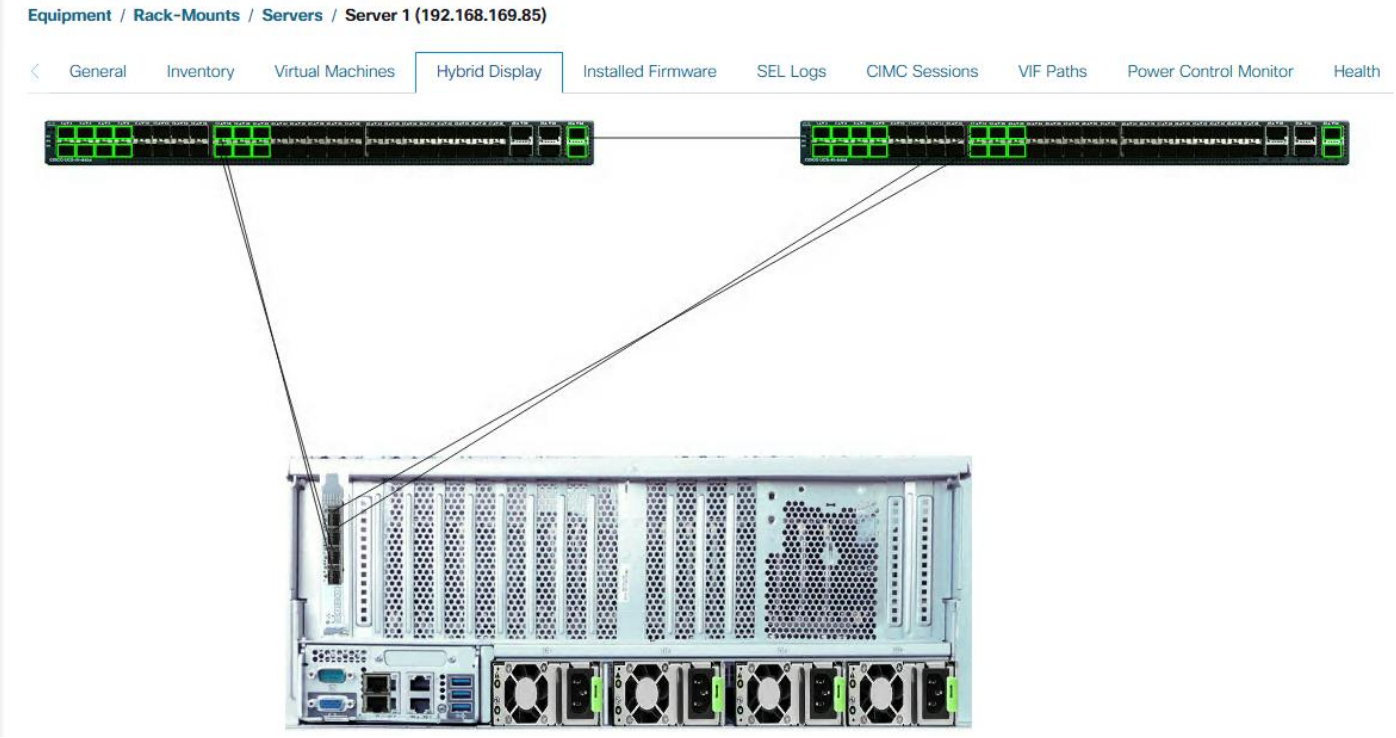
Cisco VIC 1455 has four 25GbE ports which are connected to the Cisco UCS 6454 FI in pairs such that ports 1 and 2 are connected to the FI-A and the ports 3 and 4 are connected to the FI-B as shown in Figure 18 . The ports connected to an FI form a port-channel providing an effective 50GbE bandwidth to each FI.

Figure 18 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 M5 will appear in the Cisco UCSM as shown in the figures below.

Figure 19 shows Cisco UCS C480 ML M5 connectivity to the Cisco UCS 6454 FIs:

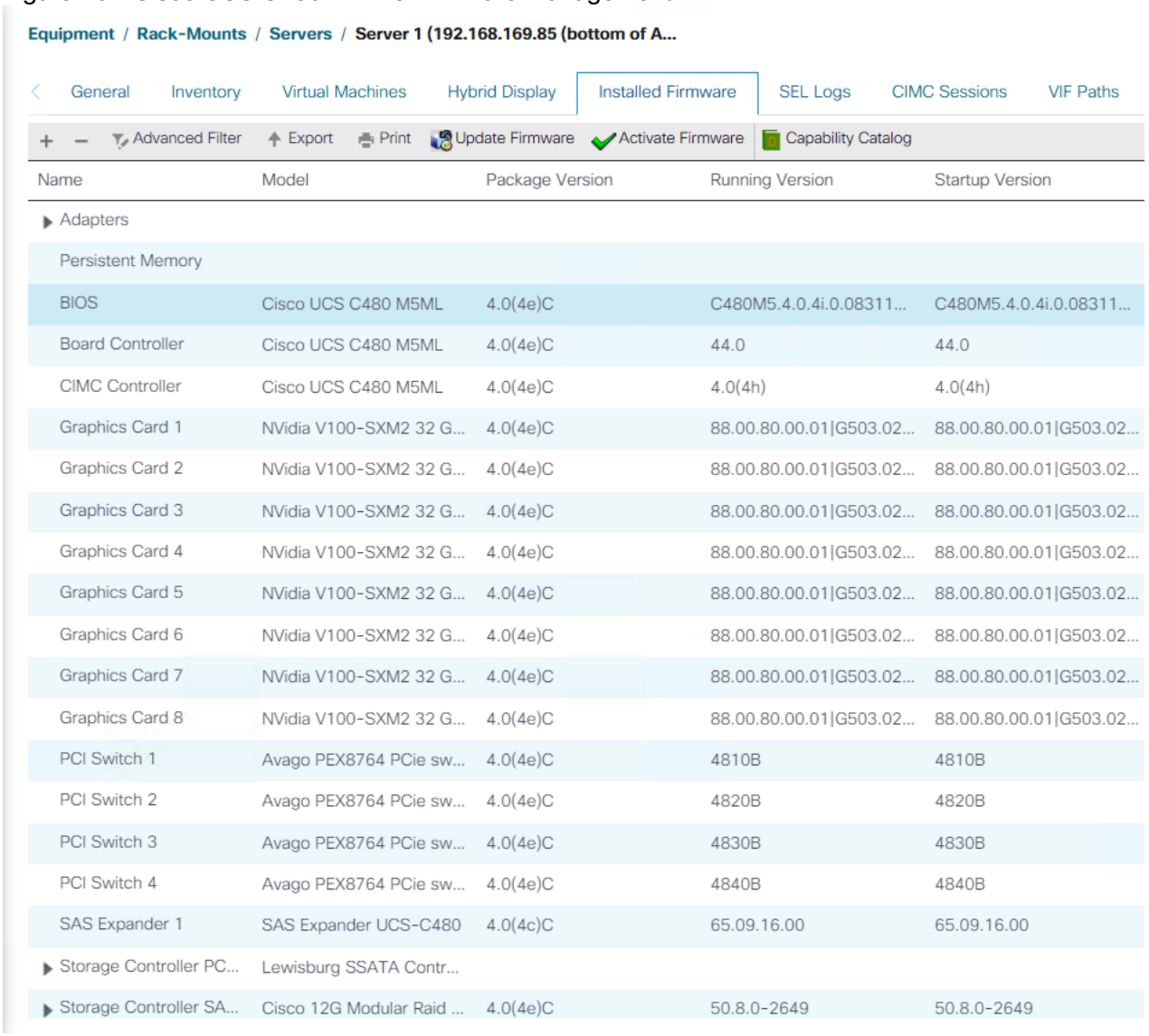
Figure 19 Cisco UCS C480 ML M5 Hybrid Display



Cisco UCS C480 ML M5 Inventory Information

Since the Cisco UCS C480 ML M5 is now being managed by Cisco UCS Manager, the firmware management for the platform is also handled by Cisco UCS Manager. Figure 20 shows various firmware versions and upgrade options for the platform, including the firmware installed on NVIDIA GPUs.

Figure 20 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 21) in Cisco UCS C480 ML M5 platforms under the inventory page and shows the information about the PCI switches (Figure 22).

Figure 21 Cisco UCS C480 ML M5 GPU Inventory

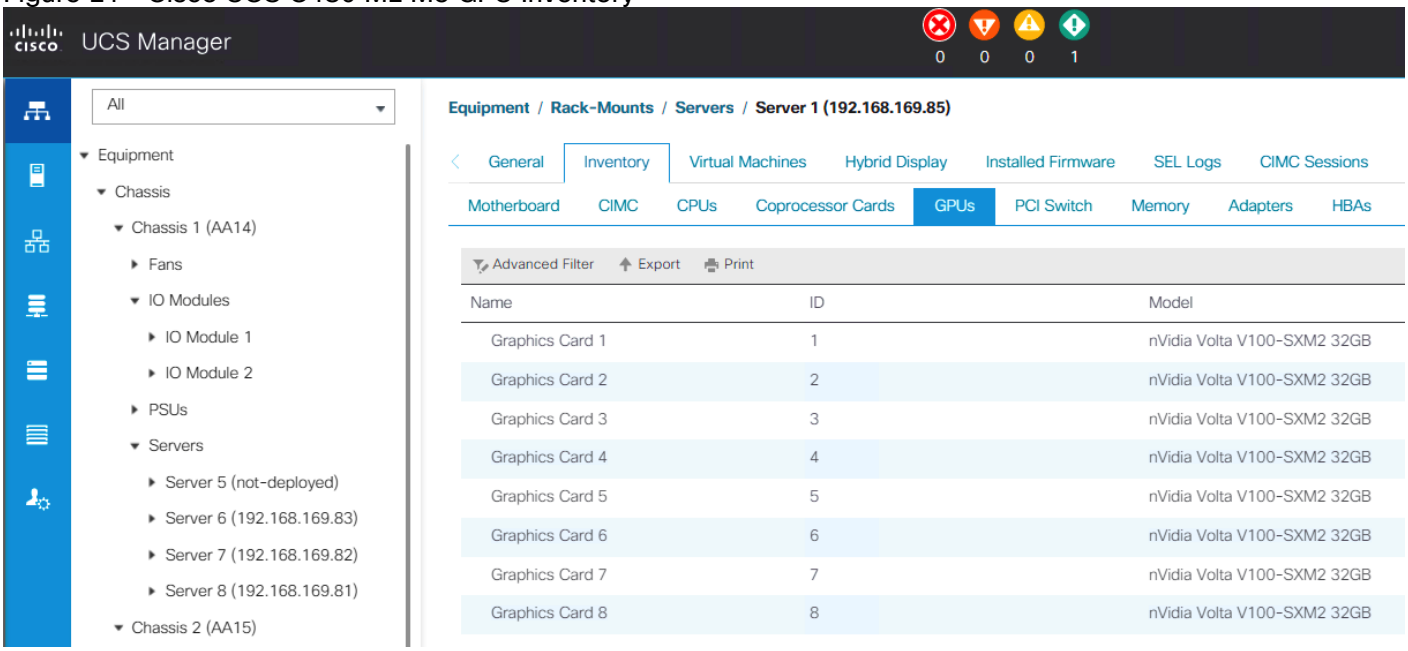
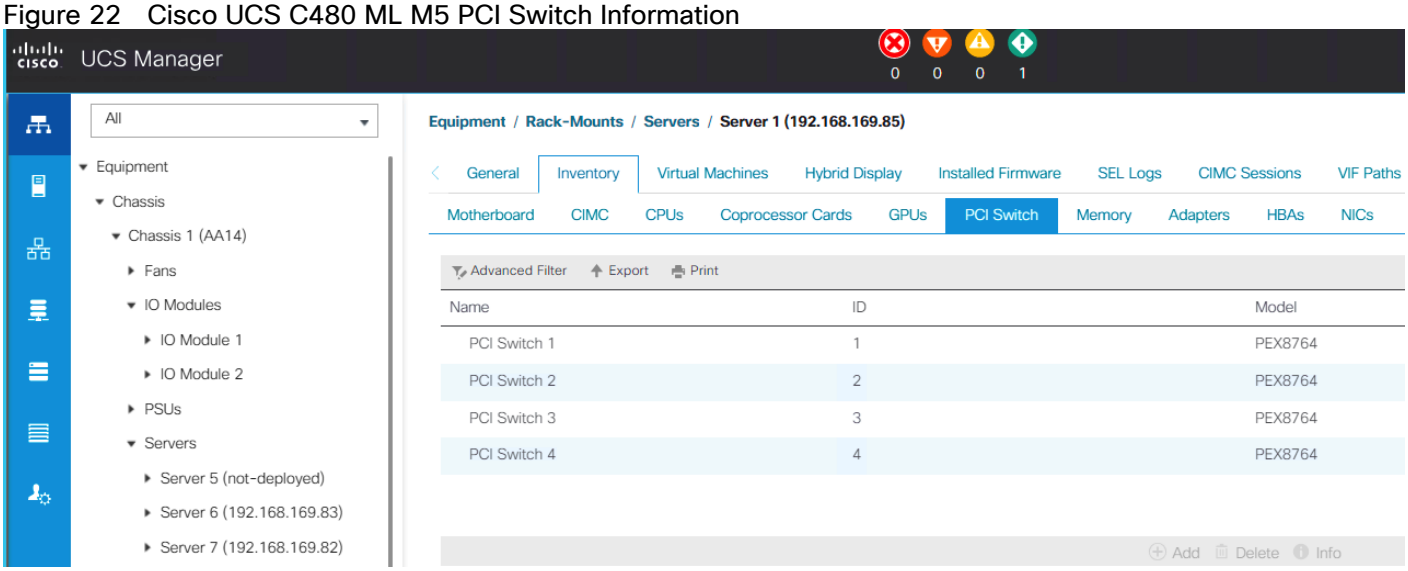


Figure 22 Cisco UCS C480 ML M5 PCI Switch Information



Service Profile Configuration

The 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 FlashStack environment for AI/ML including their specific usage.

Table 1 VLAN Usage

VLAN ID	Name	Usage

VLAN ID	Name	Usage
2	Native-VLAN	Use VLAN 2 as Native VLAN instead of default VLAN (1)
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
1110*	iSCSI-A-VLAN	iSCSI-A path for both B-Series and C-Series servers
1120*	iSCSI-B-VLAN	iSCSI-B path for both B-Series and C-Series servers
1130	vMotion	VLAN user for VM vMotion
3152	AI-ML-NFS	NFS VLAN to access AI/ML NFS volume on FlashBlade

*** For the FlashStack Virtual Server Infrastructure with iSCSI deployment guide, refer to:** https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/flashstack_vsi_iscsi_vm67_u1.html

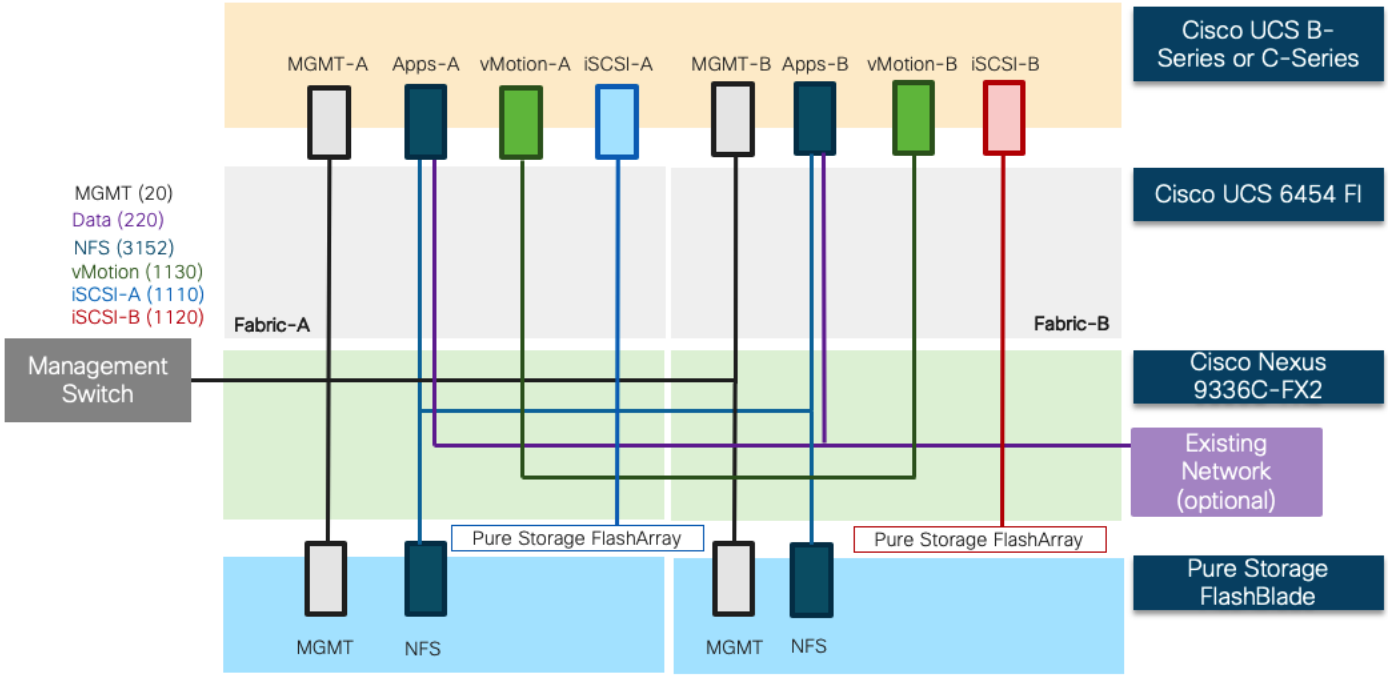
Some of the key highlights of VLAN usage are as follows:

- Both virtual machines and the bare-metal servers are managed using VLAN 20.
- An optional dedicated VLAN (220) is used for data communication; customers are encouraged to evaluate this VLANs usage according to their specific requirements.
- A dedicated NFS VLAN is defined to enables NFS data share access for AI/ML data residing on Pure Storage FlashBlade.
- A pair of iSCSI VLANs are utilized to access iSCSI LUNs for ESXi servers.
- A vMotion VLAN for VMs migration (in the VMware environment).

Service Profile for VMware Hosts

In FlashStack deployments, 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 23 :

Figure 23 ESXi Service Profile



Each ESXi service profile supports:

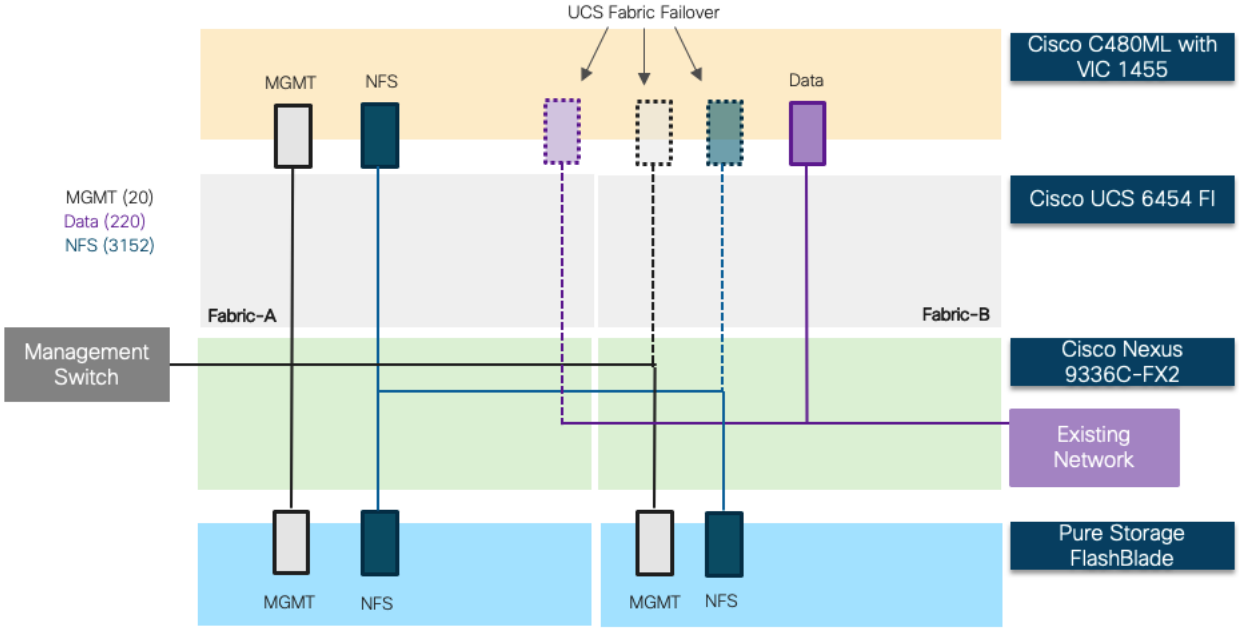
- Managing the ESXi hosts using a common management segment
- Diskless SAN boot using iSCSI with persistent operating system installation for true stateless computing
- Eight vNICs where:
  - 2 redundant vNICs (MGMT-A and MGMT-B) carry management traffic.
  - 2 redundant vNICs (Apps-A and Apps-B) carry application traffic including access to NFS volume on FlashBlade. The MTU value for this interface is set to 9000.
  - 2 redundant vNICs (vMotion-A and vMotion-B) provide vMotion capabilities. The MTU value for this interface is set to 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 to 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 to 9000.
- Each ESXi host allows VMs deployed on the host to access to ImageNet data using the NFS VLAN

The common ESXi service profile enables AI/ML NFS VLAN (3152) on Cisco UCS B-series ESXi hosts as well. However, the vGPU enabled VMs will only be deployed on GPU equipped Cisco UCS C-series ESXi hosts.

Service Profile for Bare-Metal Hosts

The service profile configuration for a Cisco UCS C-series host deployed as a bare-metal server is as shown in Figure 24

Figure 24 Cisco UCS C-Series 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.
- Three 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 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 setup 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).
- Each Cisco UCS C-Series host accesses NFS datastores configured on Pure Storage FlashBlade to be used for hosting AI/ML workload data (for example, Imagenet database).

- For handling the high-speed data efficiently, the NFS traffic and the data traffic vNIC (if needed) are configured on separate FIs.

All Cisco UCS C-Series M5 servers (including C220, C240 and C480 ML M5) equipped with Cisco VIC 1455/1457 will utilize a common service profile for bare-metal deployment

Network Design

Nexus Features

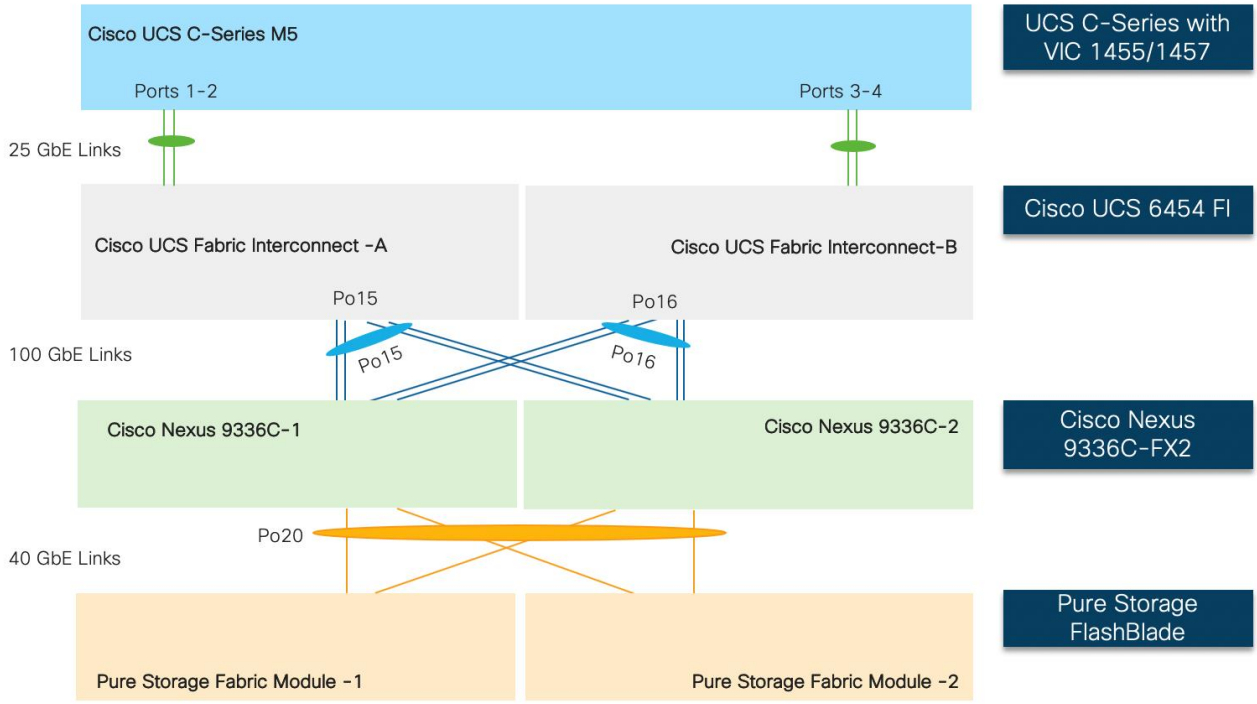
The Nexus 9336C-FX2 device configuration covers the core networking requirements for Layer 2 and Layer 3 communication. Some of the key NX-OS features implemented within the design are:

- 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 Pure FlashBlade Logical Connectivity to Nexus Switches

Figure 25 shows the connectivity between GPU equipped UCS C-series servers, UCS Fabric Interconnects (FI) and Pure Storage FlashBlade. 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 can be setup between Cisco Nexus and Cisco UCS FIs for increased bandwidth.

Figure 25 Logical Network Connectivity

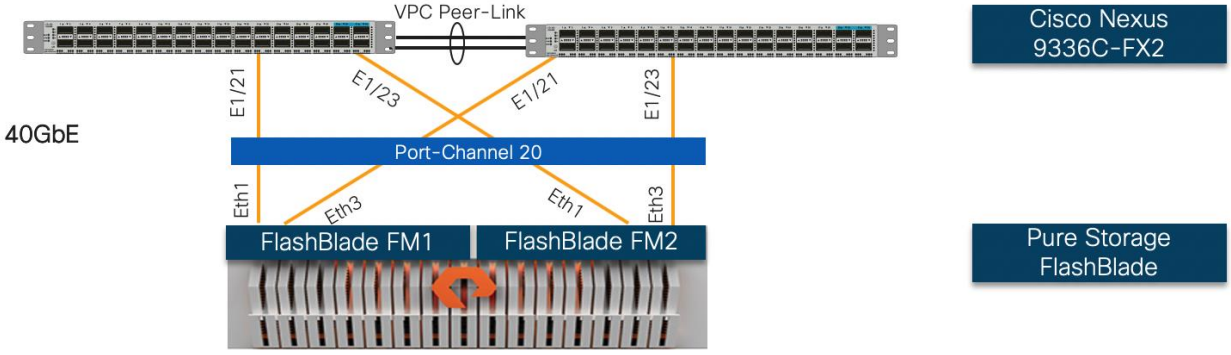


Storage Design

Physical Connectivity

Pure Storage FlashBlade External Fabric Modules are connected to Cisco Nexus 9336C-FX2 switches using 40GbE connections. Figure 26 depicts the physical connectivity design of the FlashBlade system.

Figure 26 Pure Storage FlashBlade Connectivity



The physical configuration of the FlashBlade system used during the validation is listed in Table 2 :

Table 2 Pure Storage FlashBlade Configuration

Storage Components	Description
FlashBlade	Pure Storage FlashBlade
Capacity	162.46 TB

Storage Components	Description
Connectivity	8 x 40 Gb/s redundant Ethernet port
Physical	4U

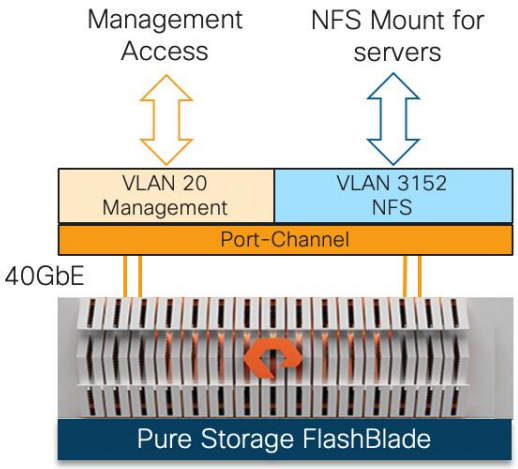
Network Connection to Cisco Nexus 9336C-FX2

In the FlashStack for AI and Deep Learning design, each Nexus 9336C-FX2 switch is connected to each of the External Fabric Modules using 2 x 40Gb connections. All the ports on the FlashBlade are configured as part of a single Port-Channel. On the Nexus switches, a Virtual Port-Channel (vPC) configuration is deployed so that FlashBlade sees both Nexus switches as a single virtual Switch.

The FlashBlade network settings are configured with 2 subnets across 2 VLANs as shown in Figure 27 . These interfaces are used as follows:

- Management-VLAN 20: This VLAN/subnet is used to manage the FlashBlade. The default gateway is configured in this subnet.
- NFS-VLAN 3152: This VLAN/subnet is used to mount NFS volume which contains the AI/ML dataset (Imagenet_dataset).

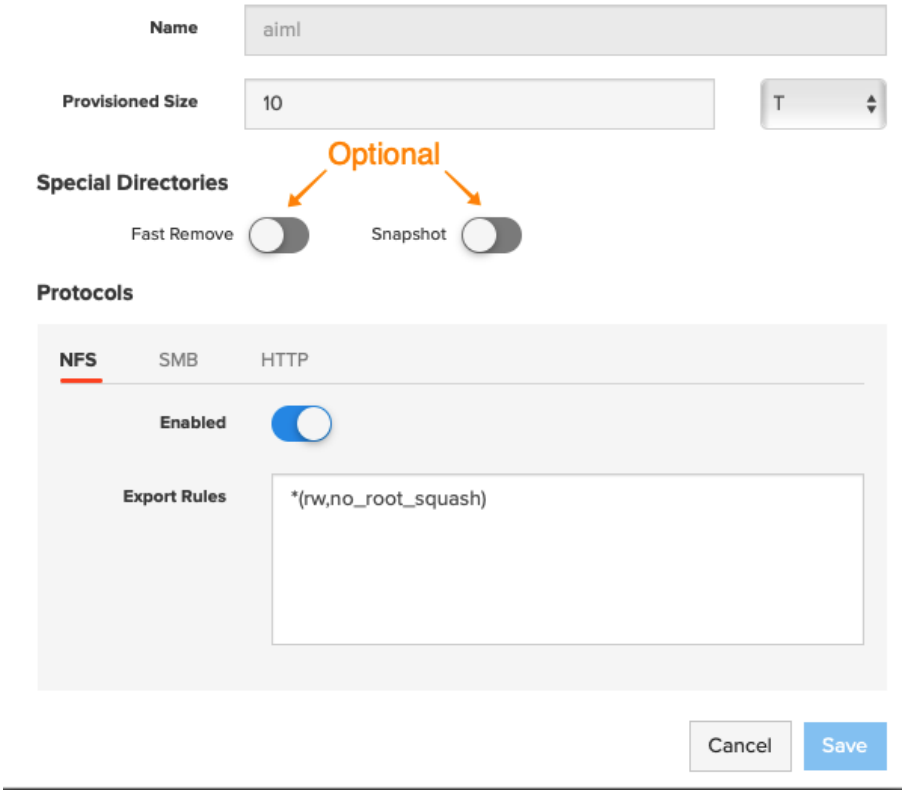
Figure 27 Pure Storage FlashBlade Subnets



NFS Configuration on Pure Storage FlashBlade

To add the NFS share in Pure Storage FlashBlade using Graphical User Interface (GUI), select Storage in the left pane and select File Systems in the main window. The NFS share can be added by clicking the '+' sign and providing various values in system dialog as shown in Figure 28 :

Figure 28 Setting up NFS Share on Pure Storage FlashBlade



Once the NFS share is successfully created and appropriate export rules are added, the NFS share will be mounted on the Cisco UCS C480 ML M5 servers by adding an entry in the Linux /etc/fstab file.

Software Setup and Configuration

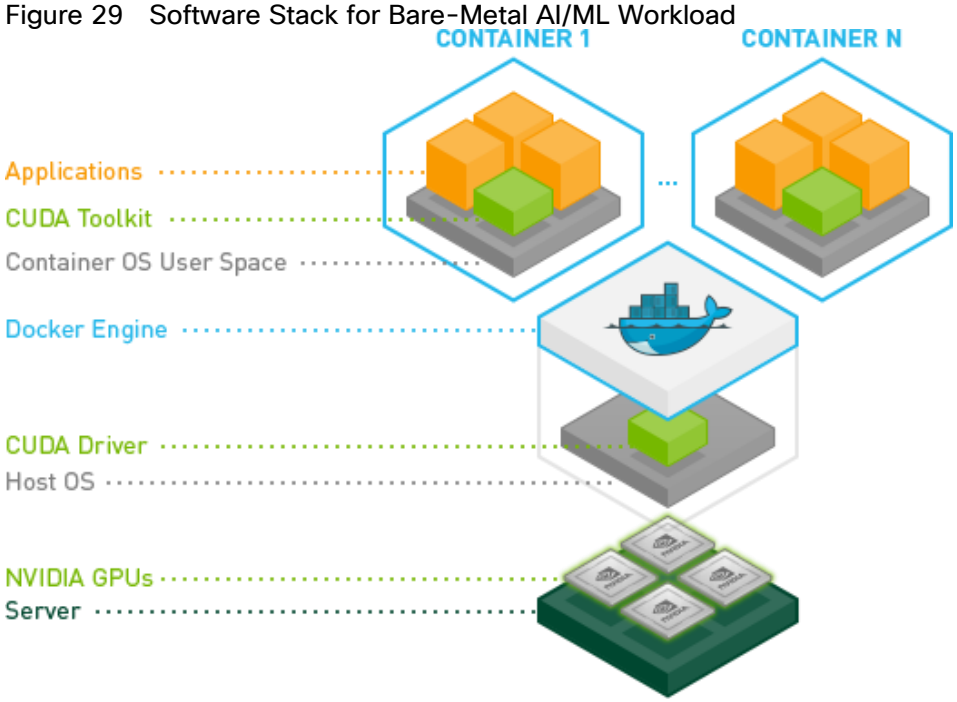
NVIDIA GPU Cloud

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 while providing 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 29 provides a high-level overview of the software stack installation on a bare-metal server:

Figure 29 Software Stack for Bare-Metal AI/ML Workload



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

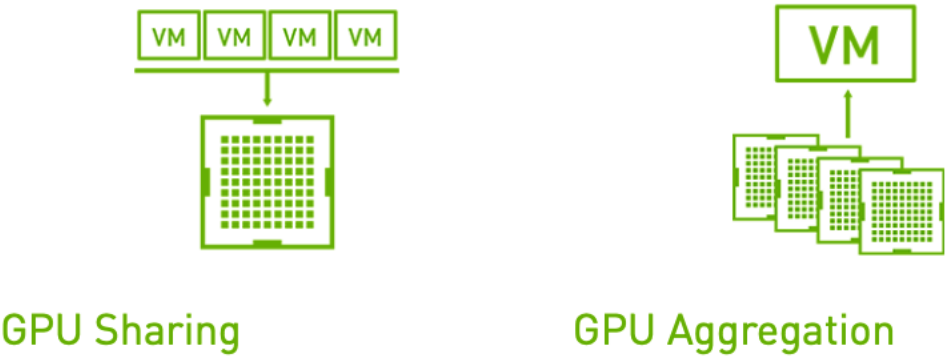
- Download and install the RHEL 7.6 on GPU equipped Cisco UCS C-series servers.
- Download and install Linux packages including gcc, kernel headers, development packages, and so on.
- Download and install NVIDIA Driver and CUDA Toolkit.
- Download and install NVIDIA Docker 2.
- Download and run a (TensorFlow) container from NGC.

To validate the installation, customers can download and execute the CNN Benchmark Script for ImageNet data hosted on Pure Storage FlashBlade. 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

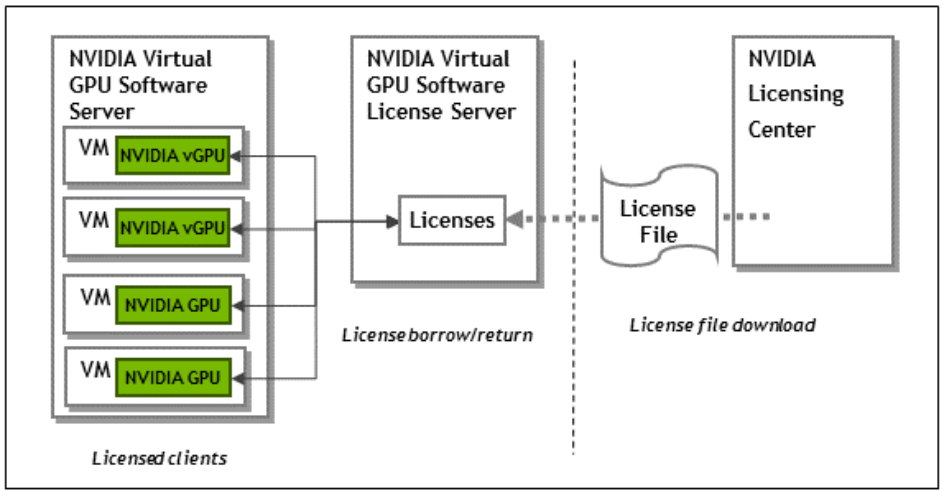
NVIDIA Virtual Compute Server (vComputeServer) enables the benefits of hypervisor-based server virtualization for GPU- accelerated servers so that the most compute-intensive workloads, such as AI and ML can be run in a VM. vComputeServer supports NVIDIA NGC GPU-optimized software and containers. With 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 30 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 31 NVIDIA vGPU Software Architecture



To utilized GPUs in a VM environment, 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

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 Cisco UCS (as shown in Figure 32)
- Applying the QoS policy to appropriate vNIC (as shown in Figure 33)

For setting up QoS for the return NFS traffic i.e. traffic from Pure Storage FlashBlade 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 FlashBlade
- Implementing appropriate QoS configuration to prioritize the marked traffic

Figure 32 QoS Policy on Cisco UCS

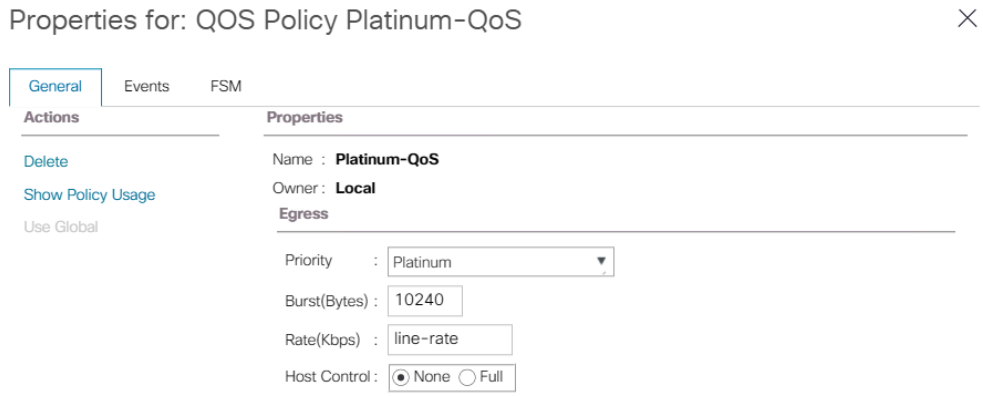
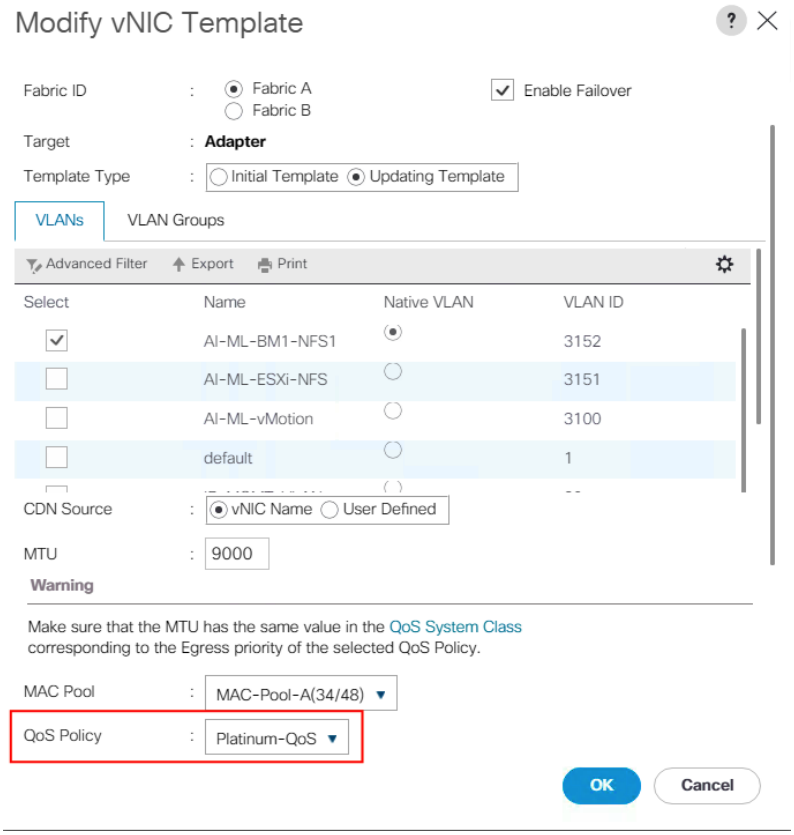


Figure 33 Applying QoS Policy to a vNIC



The QoS policy shown in Figure 32 is 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.

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 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

Set 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 3 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(23)
	VMware vSphere	6.7U3
	ESXi ENIC Driver	1.0.29.0
	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	Pure Storage FlashBlade (Purity//FB)	2.3.3

Validation

FlashStack for AI solution 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 to observe the GPU and Storage performance.
- Verify 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 GPU equipped Cisco UCS C-Series 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.
- Running 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.

Pure Storage and Cisco engineering teams have also worked on a FlashStack for AI Scale-Out solution. The following document provides the details of the validation:
https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/whitepaper-c11-742103.pdf

Summary

Artificial Intelligence (AI) and Machine Learning (ML) initiatives have seen a tremendous growth due to the recent advances in GPU computing technology. The FlashStack for AI solution aims to deliver a seamless integration of the GPU enabled Cisco UCS C-Series platforms including Cisco UCS C480 ML M5 into the current FlashStack 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 C480 ML M5 platform into the FlashStack design
- Integration of Pure Storage FlashBlade into the FlashStack architecture
- Showcase AI/ML workload acceleration using NVIDIA V100 32GB GPUs and NVIDIA T4 16GB GPUs.
- Support for Cisco UCS C220 M5 and C240 M5 with NVIDIA GPUs for inferencing and low intensity workloads.
- Support for Intel 2nd Gen Intel Xeon Scalable Processors (Cascade Lake) processors.
- Showcasing NVIDIA vCompute Server functionality for the AI/ML workloads in VMware environment.

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 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

Cisco UCS Hardware Compatibility Matrix:

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

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

Pure Storage FlashBlade:

https://www.purestorage.com/products/flashblade.html

FlashStack for AI: Scale-Out Infrastructure for Deep Learning

https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/whitepaper-c11-742103.pdf

About the Author

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

Haseeb Niazi has over 20 years of experience at Cisco in the Data Center, 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).

Acknowledgements

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

- Allen Clark, Technical Marketing Engineer, Cisco Systems, Inc.