White Paper

Build an AI ecosystem with FlexPod AI

Arvind Ramakrishnan, NetApp

June 2021 | WP-7345

In partnership with

Abstract

This white paper describes the enhancements to the FlexPod AI solution, including support for newer GPU acceleration options, a simplified user experience for data scientists, and how customers can build an end-to-end AI data lifecycle with FlexPod.

TABLE OF CONTENTS

LIST OF FIGURES

Executive summary

The FlexPod® Datacenter solution is a converged infrastructure solution jointly developed by NetApp® and Cisco. It provides a predesigned, integrated, and validated architecture for a data center to host a wide variety of applications and enterprise workloads. The FlexPod solution architecture incorporates compute, network, and storage design best practices and ensures compatibility between the various components, thereby minimizing IT risks.

Since its launch around 10 years ago, FlexPod has constantly evolved to meet the needs of the industry by supporting newer workloads, applications, and catering to infrastructure modernization requirements. FlexPod has also been addressing the needs of the emerging and cutting-edge technology areas, such as artificial intelligence (AI) and machine learning (ML).

AI and ML have seen a tremendous growth in the last few years due to their immense potential in helping businesses on several fronts from accelerating their growth, enhancing customer experience, reducing errors, improving logistics, automation, fraud detection and much more. The key to make all this happen is data and the ability to learn from the data that is on hand as quickly as possible.

Cisco Validated Designs (CVDs) on the design and deployment of a FlexPod solution for AI/ ML describe in detail the integration of the compute, network, storage, and the AI frameworks for performing AI training operations using a FlexPod platform.

This white paper is intended to provide an overview of the latest enhancements to the FlexPod AI solution, which includes support for newer GPU acceleration, workflow driven AI operations on a Kubernetes platform, building an AI lifecycle with FlexPod, and simplified data management that helps deliver a top-notch user experience.

Data center modernization and challenges for AI workloads

A few years back, the adoption of artificial intelligence was low, but today, every organization is looking for ways to extract insightful data from their businesses to grow further and faster. Data is a vital part of this journey and data centers are at the heart of this transformation.

No transformation is straightforward and the same holds good for data centers. Data centers need to meet new requirements from new user personas such as data scientists and data engineers and address the challenges that come with the resource-intensive nature of AI/ ML workloads.

Some of the requirements and challenges that data scientists must manage include the following:

- Gathering data from multiple sources to build a dataset for AI/ML operations
- Setting up the infrastructure to handle high volumes of data
- Accessing compute resources with acceleration support
- Having the ability to version control the trained model with the training dataset
- Minimizing exposure to DevOps and infrastructure management to focus on data science

On the infrastructure side, the AI/ML operations demand a lot of resources. There is a need for massive storage repositories to host historical data for AI/ML model training and a high velocity of incoming data for model inference and predictive analysis.

GPU based servers are a common choice because the AI/ML applications are compute intensive. The GPUs are fast due to their massive parallel processing power but need to ingest data in the form of millions of files from a dataset that can be in the range of a few hundred terabytes (TBs) to petabytes (PBs). The network and storage infrastructure have a critical role to play in ensuring that the GPUs ingest data at a rapid rate.

The storage systems must deliver capacity and performance elasticity to expand and maximize price and performance. The datasets grow quickly over time and the expansion of storage in the form of drives, shelves, or storage controllers should result in an equivalent increase in performance.

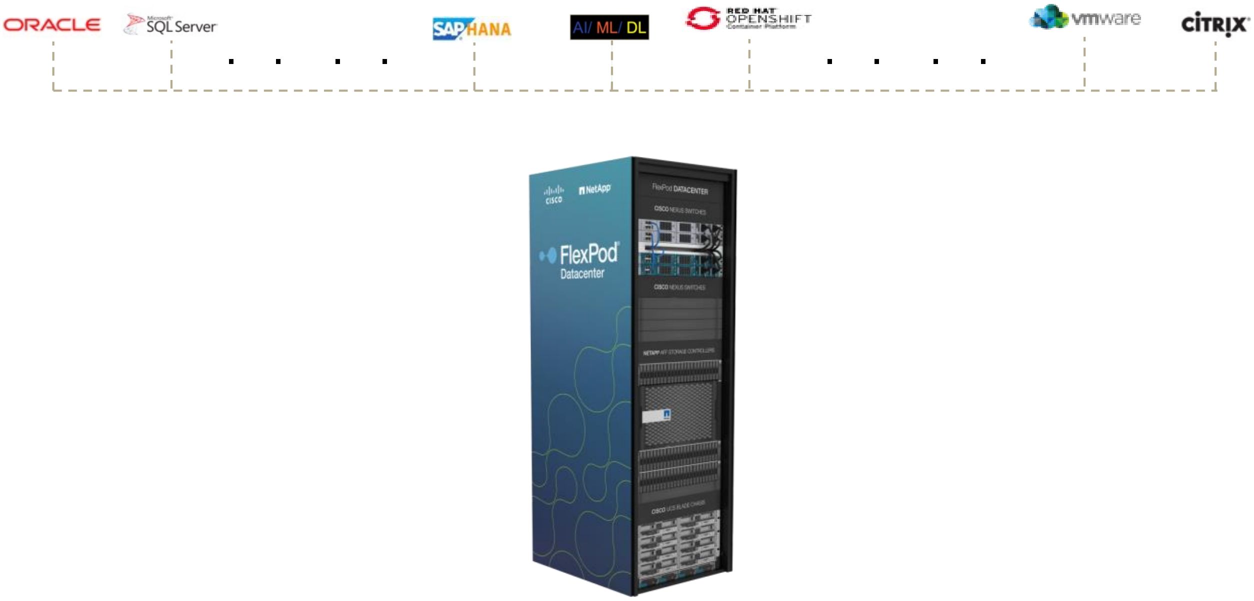
The data center should also be able to concurrently support all phases of the AI workflow. The ability to seamlessly share and schedule resources plays a big role in maintaining a cost-efficient data center. Data scientists should be able to work concurrently on different tasks such as data ingestion, training, and inferencing.

Most customers and/or businesses have had to rearchitect their data centers to set foot on their AI journey. The data centers that were serving their traditional workloads either lacked the features to transform and meet the needs of AI/ ML or needed a complete overhaul of the design. This resulted in a complex environment and increase in management and maintenance overhead, leading to a lower return on investment (ROI) and higher TCO.

FlexPod for AI and traditional workloads

There are more than 170 (and counting) validated reference architectures on FlexPod with design and deployment best practices for a wide variety of enterprise workloads ranging from databases, virtual desktop infrastructure, server virtualized workloads, SAP, to AI/ ML.

FlexPod provides a unique value proposition to its customers by highly simplifying their data center’s transformation to meet the needs of AI/ML workloads while continuing to serve their existing applications.



Irrespective of the workloads being hosted on a FlexPod, the solution fundamentals such as resilient architecture design, usage of standardized and interoperable components, and implementation of workload driven best-practices are always consistent. This has ensured that customers running a FlexPod solution in their data center need to deal with minimum to nil design changes to host newer workloads concurrently.

Every storage, network, and compute enhancement in the form of a software release or new hardware, integrates seamlessly into a FlexPod with very tight integration between the other components.

Listed following are some of the key features for AI /ML that have made their way into the FlexPod solution:

- NVMe-backed storage with NetApp AFF running NetApp ONTAP®
- Host millions of files in a single scalable logical container with NetApp ONTAP FlexGroup volumes
- Support for high-speed networking at 100GbE with Cisco Nexus switches
- Support for GPU acceleration options with Cisco UCS Compute
- Hybrid cloud connected data center with data fabric powered by NetApp
- Kubernetes based hosting of workloads with persistent storage through NetApp Trident

Each of the above-mentioned capabilities were introduced in FlexPod as and when they were generally available. FlexPod customers can easily extend their data centers to support the new features by just upgrading their software versions or scaling their FlexPod by procuring peripheral components as part of their operational expenditures (opex).

The continuous integration of new features and technologies in the FlexPod platform ever since its launch has ensured that the data center is always in the latest generation with the best of breed technology and opens up avenues for newer workloads such as AI/ ML to be run on the same platform.

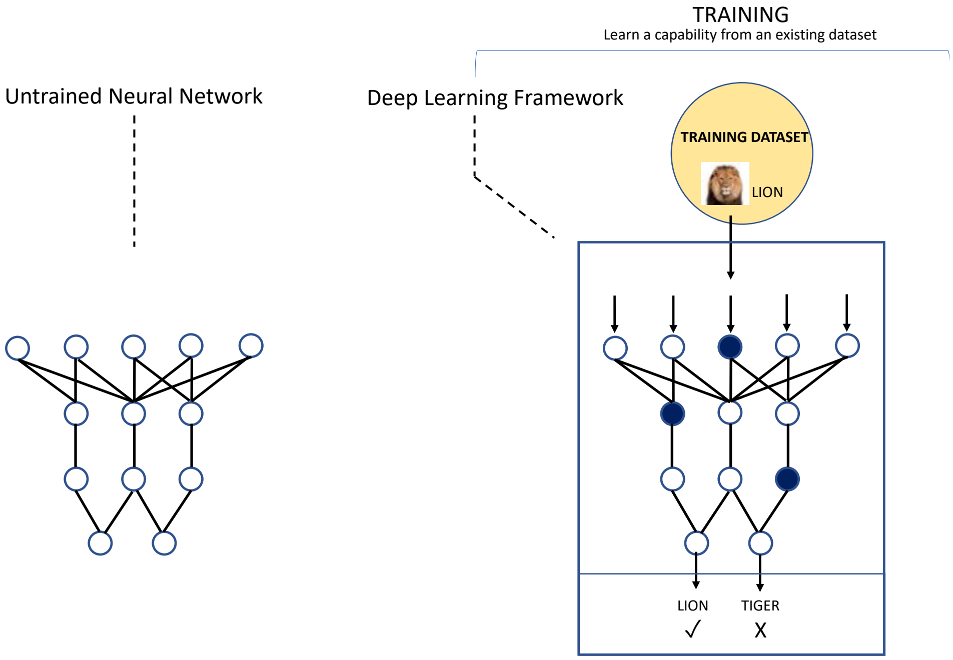
AI training and inference overview

The AI workloads can be broadly classified into two categories—training and inferencing.

Training

Training is the process of building an ML algorithm with the help of a deep learning (DL) framework and a dataset to train on, as shown in Figure 1.

Figure 1) AI training overview.



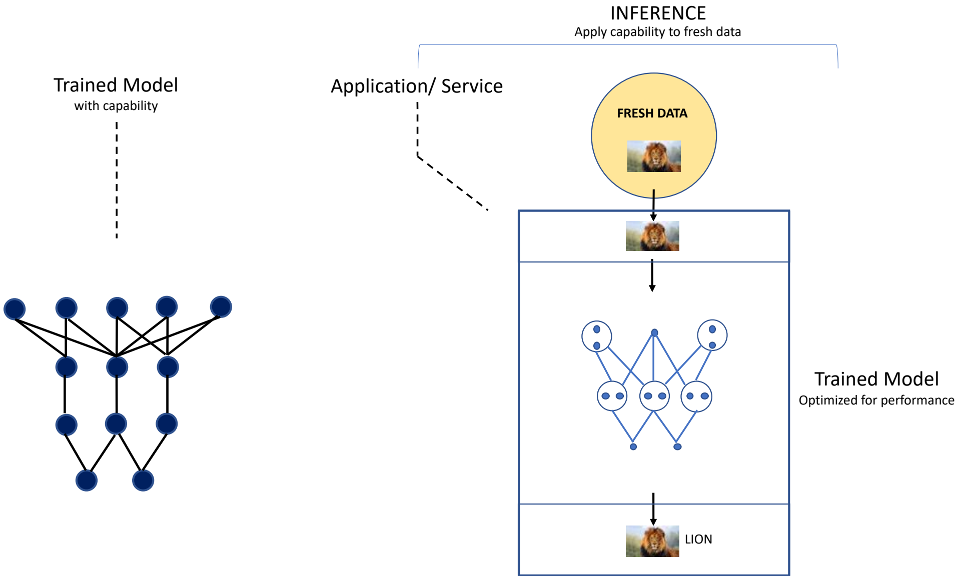
The dataset contains a known set of data and put together by the team that is trying to build a trained model. A known dataset is put through an untrained neural network. The results of the framework are compared with the known dataset results. Following the comparison, the error values are reevaluated by the framework, and it updates the weight of the data set in the layers of the neural network depending on how correct or incorrect the value is. This process of reevaluation is critical to training because it adjusts the neural network to improve the performance of the task it is learning.

This process is highly resource intensive; the size of the dataset can be huge, in the range of petabytes, and consisting of billions of files. The entire dataset needs to pass through the model multiple times in-order to build a well-trained model. This means that the network should be fast enough to pump all this data to the compute infrastructure as fast as possible and the compute layer should be able to munch all the data fed to it in a very short span of time.

Inferencing

Inferencing refers to the process of using a trained ML algorithm to make a prediction. As an example, you can use IoT data as an input to a trained machine learning model enabling predictions that can guide decision logic on the device, at the edge gateway, or anywhere else on the Internet of Things (IoT) system, as shown in the following figure.

Figure 2) AI inference overview.



Inferencing does not deal with reevaluation or adjustments to the layers of the neural network, based on the results. Instead, inferencing applies the knowledge gained by a trained neural network model on real-time data and tries to infer a result. When a new unknown dataset or production data is fed through a trained neural network, the output is a prediction based on predictive accuracy of the neural network.

Introduction to key products and features

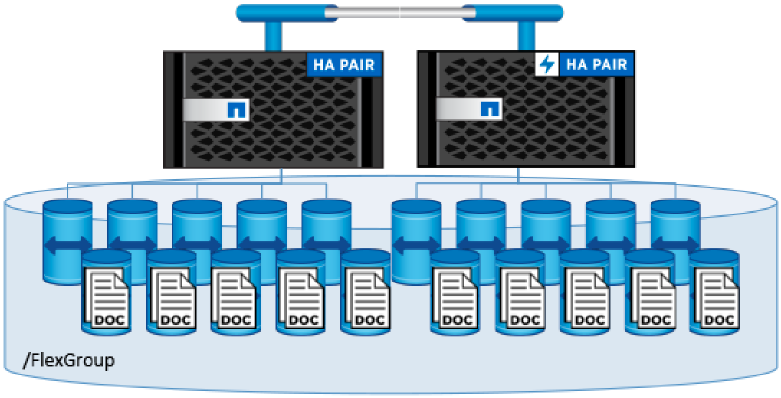
This section introduces the key products or features that play a significant role in building the FlexPod AI solution.

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 can 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 (ALB) by using all the resources available in the storage cluster and can scale to multiple petabytes of capacity, offering optimal performance. Figure 3 illustrates the architecture of a FlexGroup volume.

Figure 3) A FlexGroup volume.



Multiple FlexVol volumes are stitched together into a single namespace that behaves like a single FlexVol volume to clients and administrators. 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.

NetApp FlexCache

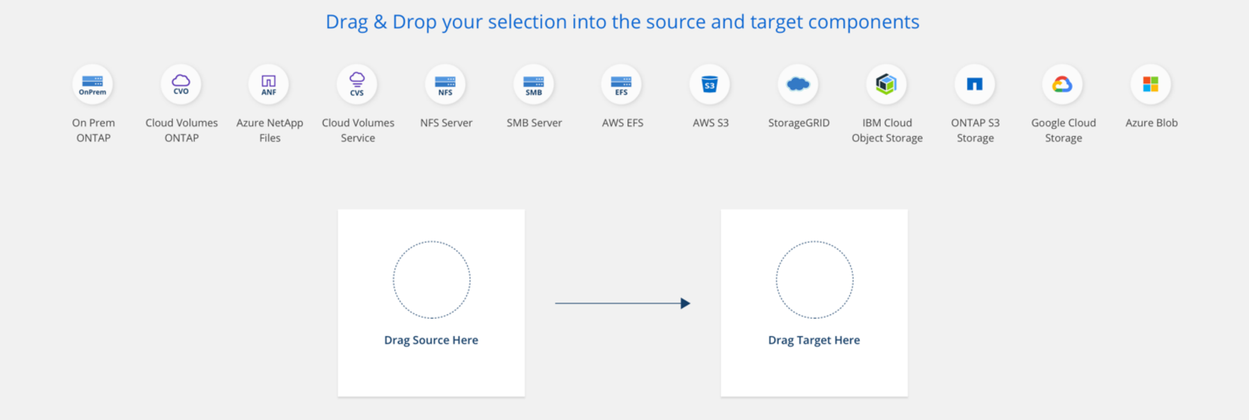
NetApp FlexCache® software provides a remote caching capability that simplifies file distribution, reduces WAN latency, and lowers WAN bandwidth costs. It provides the ability to cache only the actively-read data, rather than entire files or volumes, either locally within a data center or geographically dispersed at remote sites.

By serving hot data (data that needs to be accessed frequently), from multiple controllers in a cluster, you can increase the performance delivered to key applications. And by caching hot data local to users at multiple locations around the world, you can enhance their collaboration by enabling simultaneous access to centralized datasets while also reducing the response time they receive when accessing the hot data. AI inferencing models are developed to serve large scale inferencing requests. These models are accessed by multiple clients and applications concurrently from various endpoints. By caching the models at strategic locations, the inference performance can be significantly improved. The FlexCache volumes can also easily access newer versions of the model as and when they are available in the origin volume.

NetApp Cloud Sync

NetApp Cloud Sync offers a simple, secure, and automated way to migrate data to any target, in the cloud or on premises, as shown in the following figure. After the data is transferred, it is fully available for use on both source and target. Cloud Sync continuously synchronizes the data, based on a predefined schedule, moving only the deltas, so time and money spent on data replication is minimized.

Figure 4) NetApp Cloud Sync source and target.



Drag & Drop your selection into the source and target components

NetApp SnapMirror

The NetApp SnapMirror® feature is a feature of ONTAP that enables data replication. You can replicate data from specified source volumes or qtrees to specified destination volumes or qtrees, respectively. You can use the SnapMirror feature to replicate data within the same storage system or with different storage systems.

You can configure SnapMirror to operate in any of the three modes listed below:

- **Asynchronous mode**. Replicates NetApp Snapshot™ copies to the destination at specified, regular intervals.
- **Synchronous mode**. Replicates data to the destination as soon as the data is written to the source volume.
- **Semi-synchronous mode**. Replication at the destination volume lags behind the source volume by 10 seconds.

Kubernetes

Kubernetes, also known as K8s, is an open-source system for automating deployment, scaling, and management of containerized applications. It groups containers that make up an application into logical units for easy management and discovery. Originally designed by Google, it is now maintained by the Cloud Native Computing Foundation (CNCF). In recent years, Kubernetes has emerged as the dominant container orchestration platform. Although other container packaging formats and run times are supported, Kubernetes is most often used as an orchestration system for Docker containers. For more information, visit https://kubernetes.io.

NetApp Trident

Trident is an open-source dynamic storage orchestrator that simplifies the consumption of persistent volumes (PV) in Kubernetes. It is a Kubernetes-native application and runs directly within a Kubernetes cluster. With Trident, Kubernetes users (developers, data scientists, Kubernetes administrators, and so on) can create, manage, and interact with persistent storage volumes in the standard Kubernetes format that they are already familiar with. At the same time, they can take advantage of NetApp advanced data management capabilities and a data fabric that is powered by NetApp technology. Trident abstracts away the complexities of persistent storage and makes it simple to consume. For more information, see https://netapp.io/persistent-storage-provisioner-for-kubernetes/.

NVIDIA DeepOps

DeepOps is an open-source project from NVIDIA that uses Ansible to automate the deployment of GPU server clusters according to best practices. DeepOps is modular and you can use it for various deployment tasks. In the context of this paper, DeepOps is used to deploy a Kubernetes cluster that consists of GPU server worker nodes in the form of Cisco UCS servers. For more information, see https://github.com/NVIDIA/deepops.

Kubeflow

Kubeflow is an open-source AI and ML toolkit for Kubernetes that was originally developed by Google. The Kubeflow project is dedicated to making deployments of AI and ML workflows on Kubernetes simple, portable, and scalable. Kubeflow has been gaining significant traction as enterprise IT departments have increasingly standardized on Kubernetes. For more information, see https://www.kubeflow.org.

Kubeflow Pipelines are a key component of Kubeflow. Kubeflow Pipelines are a platform and standard for defining and deploying portable and scalable AI and ML workflows.

Jupyter Notebook Server

A Jupyter Notebook Server is an open-source web application that enables data scientists to create Wiki-like documents called Jupyter Notebooks that contain live code, equations, visualizations, and descriptive test. Jupyter Notebooks are widely used in the AI and ML community as a means of documenting, storing, and sharing AI and ML projects. Kubeflow simplifies the provisioning and deployment of Jupyter Notebook Servers on Kubernetes. For more information about Jupyter Notebooks, see https://jupyter.org. For more information about Jupyter Notebooks within the context of Kubeflow, see the official Kubeflow documentation.

AI on FlexPod

The following FlexPod AI/ML CVDs cover in detail the step-by-step design and deployment guidance to set up a FlexPod solution for running AI training operations.

- FlexPod Datacenter for AI/ML with Cisco UCS 480 ML for Deep Learning Design Guide https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/flexpod_c480m5l_aiml_design.html
- FlexPod Datacenter for AI/ML with Cisco UCS 480 ML for Deep Learning Deployment Guide https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/flexpod_480ml_aiml_deployment.html

The subsequent sections of this white paper focus on the other technical aspects of FlexPod that are equally important for hosting AI/ML workloads.

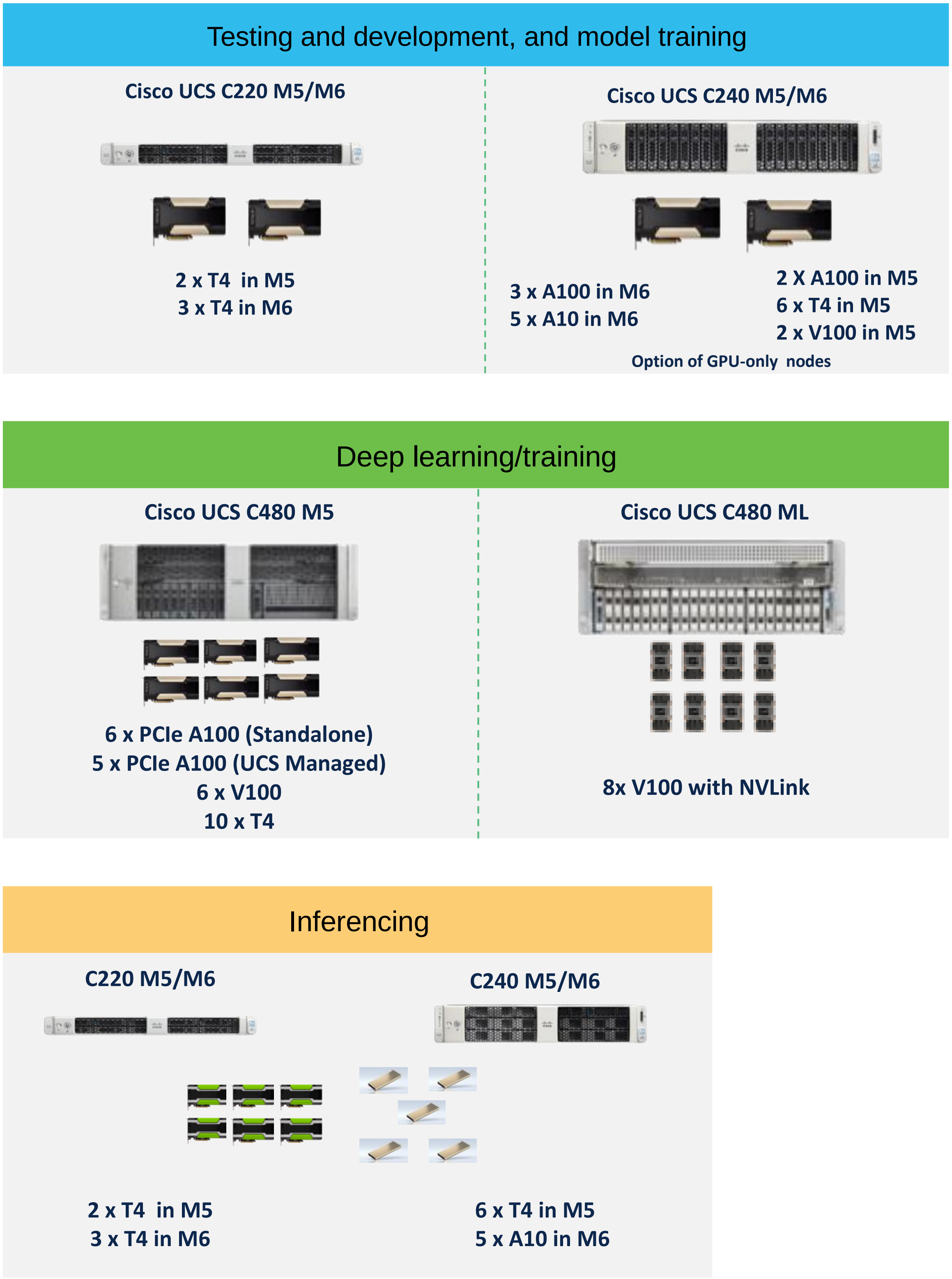
The FlexPod portfolio for AI: Acceleration options

GPUs are the most commonly used accelerators for AI/ ML training and inferencing. GPUs can process multiple computations simultaneously; they have a large number of cores, which facilitates the computation of multiple parallel processes. GPUs also have a high memory bandwidth when compared to CPUs and can ingest a lot of data at a given point, further reducing the time taken to complete a job.

FlexPod supports a variety of GPUs at different price points and capabilities that you can use for AI/ ML operations and other workloads, such as virtual desktop infrastructure (VDI).

Figure 5 illustrates the different GPU options available with FlexPod and the associated Cisco UCS servers with which you can use them.

Figure 5) FlexPod GPU acceleration options.



Testing and development, and model training

Cisco UCS C220 M5/M6

2 x T4 in M5
3 x T4 in M6

Cisco UCS C240 M5/M6

3 x A100 in M6
5 x A10 in M6

2 X A100 in M5
6 x T4 in M5
2 x V100 in M5

Option of GPU-only nodes

Deep learning/training

Cisco UCS C480 M5

6 x PCIe A100 (Standalone)
5 x PCIe A100 (UCS Managed)
6 x V100
10 x T4

Cisco UCS C480 ML

8x V100 with NVLink

Inferencing

C220 M5/M6

2 x T4 in M5
3 x T4 in M6

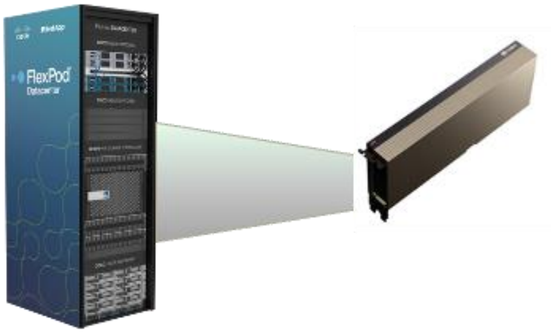
C240 M5/M6

6 x T4 in M5
5 x A10 in M6

NVIDIA A100 Tensor Core GPU

The NVIDIA A100 Tensor Core GPU, as shown in the following figure, is based on the NVIDIA Ampere GPU architecture and delivers unprecedented acceleration at every scale for AI, data analytics, and HPC. The third-generation tensor cores with Tensor Float (TF32) precision provide up to 20 times faster performance with sparsity when compared to the earlier generation.

Figure 6) NVIDIA A100 GPU.NVIDIA A100 Tensor Core GPU



The A100 supports PCIe Express Gen 4, which doubles the bandwidth of PCIe 3.0/ 3.1 by providing 31.5 GBps versus 15.75 GBps for x16 connections. This increase in speed is beneficial for A100 GPUs connecting to PCIe 4.0-capable CPUs and to support fast network interfaces.

The Multi-Instance GPU (MIG) feature expands the performance and value of each A100 GPU. It can securely partition the A100 GPU into as many as seven separate GPU Instances, each fully isolated with its own high-bandwidth memory, cache, and compute cores for CUDA applications, providing multiple users with separate GPU resources to accelerate their applications and development projects. This feature enables multiple networks to operate concurrently on a single A100 GPU for optimal utilization of compute resources. Each of the seven instance’s streaming multiprocessors have separate and isolated paths through the entire memory system. This ensures that an individual’s workload can run with predictable throughput and latency even if other tasks are saturating their limits.

MIG provides the flexibility to choose from many different instance sizes for each workload. This promotes optimal utilization of the GPU and maximizes ROI. You can use the GPU Instances on a single GPU for different purposes, such as training, inference, and high-performance computing (HPC) all at the same time, with guaranteed quality of service (QoS) around latency and throughput.

The A100 also supports Single Root Input/Output Virtualization (SR-IOV), which enables sharing and virtualizing a single PCIe connection for multiple processes or virtual machines (VMs)

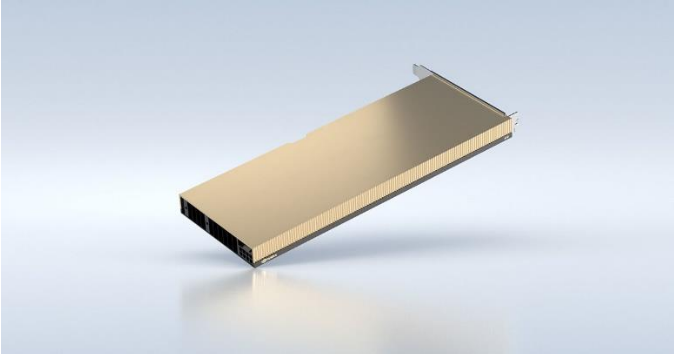
In a FlexPod solution, you can use the PCIe variant of the A100 GPUs with the Cisco UCS C240 and C480 M5 and Cisco UCS C240 M6 generation servers.

For more information, see the NVIDIA A100 product page.

NVIDIA A10 Tensor Core GPU

The NVIDIA A10 Tensor Core GPU, as shown in the following figure, is built on the latest NVIDIA Ampere architecture. The A10 combines second-generation RT Cores, third-generation Tensor Cores, and new streaming microprocessors with 24GB of GDDR6 memory—all in a 150W power envelope—for versatile graphics, rendering, AI, and compute performance.

Figure 7) NVIDIA A10 GPU.



.

NVIDIA A10 builds on the rich ecosystem of AI frameworks from the NVIDIA NGCTM catalog, CUDA-XTM libraries, over 2.3 million developers, and over 1,800 GPU-optimized applications, to help enterprises solve the most critical challenges in their business.

With FlexPod, the NVIDIA A10 GPU is available as an acceleration option for Cisco UCS C240 M6 Gen servers.

NVIDIA V100 Tensor Core GPU

The V100 GPU is the first Tensor Core GPU. It is powered by the NVIDIA Volta architecture and comes in 16GB and 32GB configurations. It has 640 Tensor Cores and was the first GPU to break the 100 TFLOP barrier of DL performance.

When paired with NVIDIA NVLink, as shown in the following figure, multiple V100 GPUs can connect with each other at speeds of 300GBps.

Figure 8) NVIDIA V100 GPU—NVLink.

Figure 9) NVIDIA V100 GPU—PCIe.

The V100 claims to deliver 24 times higher throughput when compared to a CPU for AI inferencing and 32 times higher throughput for AI training.

FlexPod customers can use the V100 GPUs in both the PCIe form factor and NVLink. The PCIe option is available with Cisco UCS C240 M5/ M6 Generation and C480 M5 Generation servers. The NVLink option is available with the Cisco UCS C480 ML servers.

For more information, see the NVIDIA V100 product page.

NVIDIA T4 GPU

The NVIDIA T4 GPU is based on Turing architecture and is packaged in an energy efficient small PCIe form factor optimized for maximum utility in enterprise data centers. You can use it for a wide range of modern applications that include ML, DL, and virtual desktops.

Figure 10) NVIDIA T4 GPU.

The NVIDIA T4 GPU is a great fit for data centers that are deployed at edge locations owing to its lower power consumption and smaller size. The T4 GPUs deliver best performance when used with servers that have been qualified to meet thermal and airflow requirements.

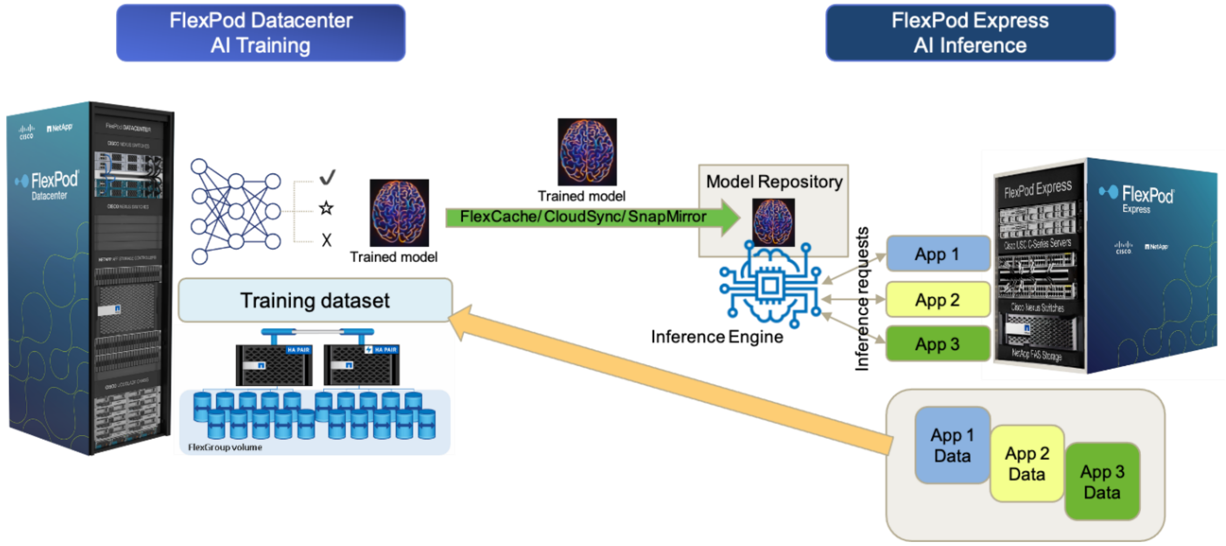
FlexPod customers can use the T4 GPUs on three qualified server models—Cisco UCS C220 M5, Cisco UCS C240 M5, and Cisco UCS C480 M5. In addition, the C220 M6 generation servers also support the T4 GPUs.

For more information, see the NVIDIA T4 product page.

AI lifecycle: Training and inference

Figure 11 shows how you can use the FlexPod portfolio to host the AI lifecycle of training and inference.

Figure 11) FlexPod for the AI lifecycle.



The FlexPod Datacenter is considered the core infrastructure where resource intensive AI model training is carried out using the AI training dataset that is presented through a FlexGroup volume. After the model has been trained and is ready for deployment, you can deliver the model to the edge locations that are running FlexPod Express. The FlexPod solutions running at both the core and the edge are powered by ONTAP and data fabric powered by NetApp.

There are multiple ways to deliver the trained model to the edge locations from the core. You can set up a FlexCache volume at the edge with the trained model being cached for read operations by the inference engine and stored in a model repository. Cloud Sync is another option to maintain a copy of the trained model at the edge locations. You can set up sync relationships between the core and edge locations with a desired sync schedule as often as every 1 minute. You can use SnapMirror to push a copy of the trained model to the edge locations. After the model is copied over at the destination volume, it can be promoted to production and be used as the model repository for the inference engine.

You can equip the FlexPod Datacenter at the core with the A100 GPUs for AI training and the FlexPod Express at the edge locations with the NVIDIA T4 GPUs to keep energy consumption low. In addition to delivering the inferencing capabilities for the applications running on the edge, the FlexPod Express also stores the application data that can be used for retraining the model. You can move this data from the FlexPod Express systems to the FlexPod Datacenter at the core by using the same data transfer capabilities offered by data fabric powered by NetApp. In the core, you can undertake the necessary data preprocessing/ cleaning/ massaging activities across all the data that has been received from various sources.

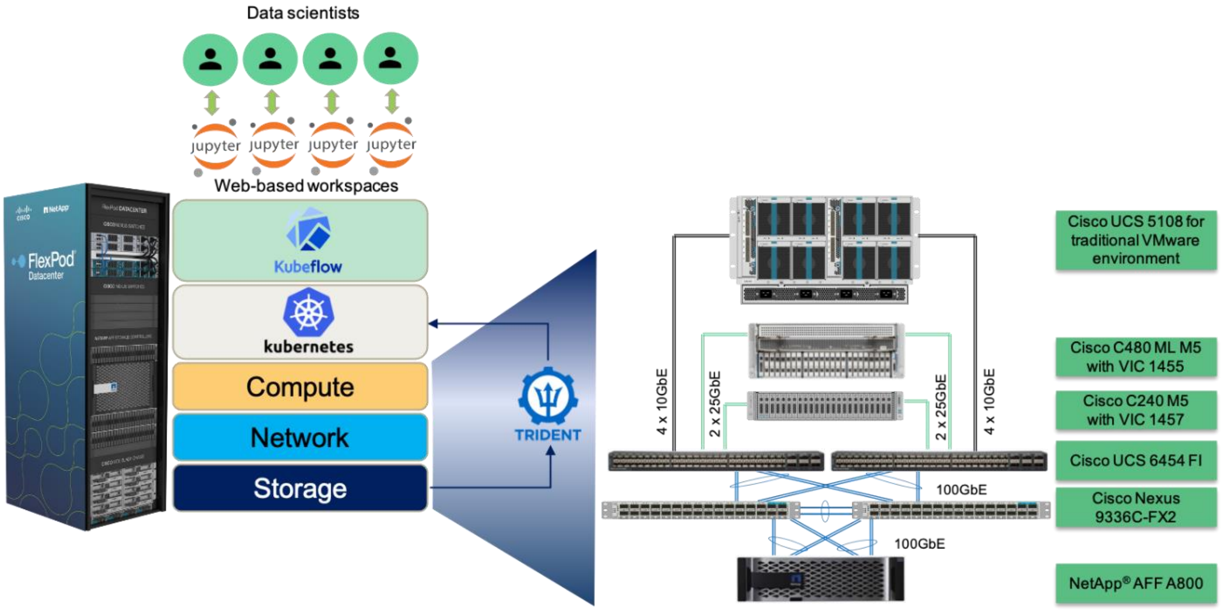
This entire process of model training, model deployment, and data gathering are typically repeated as and when there is more data available to further fine tune the model and increase its accuracy.

FlexPod customers who have been running their applications and storing the application data in the same platform can not only can extend their infrastructures to host AI training operations and use the data for training but take it a level further to support inference and the complete AI lifecycle.

FlexPod AI reference architecture

The following figure shows the reference architecture for FlexPod AI as the core infrastructure for training. The figure showcases the connectivity between the storage, network, and compute layers, which is no different from the standard connectivity being employed for all other FlexPod solutions. Each layer of the stack connects in a redundant fashion to the immediate next layer to ensure high availability and increased bandwidth for data flow. It is evident that FlexPod customers need not rearchitect the connectivity within the stack to host AI workloads.

Figure 12) FlexPod AI reference architecture.



With ONTAP, FlexPod customers can create a separate storage virtual machine (SVM) for their AI workloads. Aggregates can be assigned to this SVM, from which you can provision FlexGroup volumes to host the dataset. You can create a dedicated set of LIFs for data traffic and map them to data VLAN interfaces reserved for AI operations.

On the Cisco Nexus switches, you can configure Priority Flow Control (PFC), which enables the sending of pause frames for each specific Class of Service (CoS) and limits specific network traffic, while allowing other traffic to flow freely. Another feature called Enhanced Transmission Selection (ETS) provides the ability to allocate a specific bandwidth for each CoS and enable a tighter control over network usage. You can use these features to prioritize the AI traffic on the network for optimized performance.

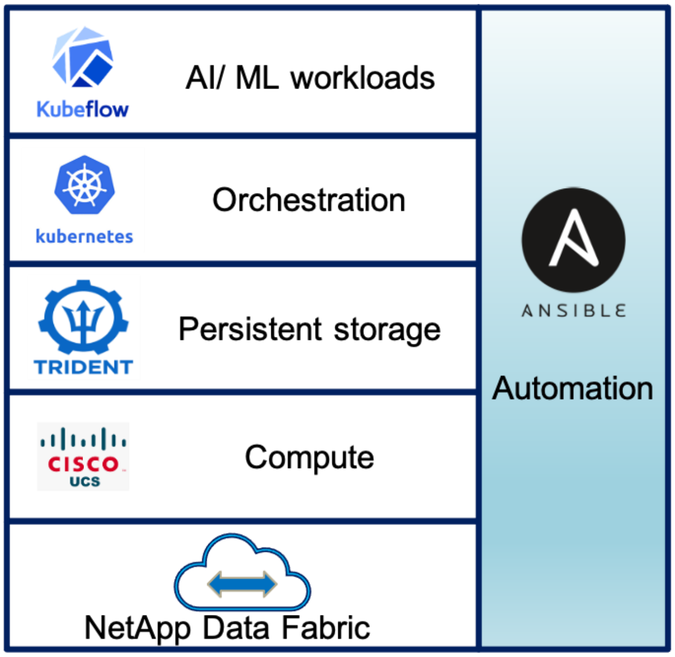
You can set up the Cisco UCS servers in a FlexPod platform that have GPU acceleration as VMware ESXi hosts with vGPU capabilities to serve AI workloads through VMs running a Linux distribution with docker, or setup as a Kubernetes worker VM. Another option is to deploy a supported Linux distribution on the UCS servers in a bare metal fashion and configure them as Kubernetes worker nodes.

NetApp Trident with Kubernetes and Kubeflow—NetApp AI Control Plane

The NetApp AI Control Plane provides a well-defined and tested approach to set up a platform for AI workloads and aims at abstracting the infrastructure and platform complexities from the application layer. It is targeted towards data scientists and machine learning operations (MLOps) Engineers and is built using open source and free components. It simplifies data management by presenting the capabilities through familiar tools and interfaces. It is a full stack solution, as shown in Figure 13, that pairs popular open-source MLOps tools with NetApp technology to rapidly manage AI data.

Kubernetes with NetApp Trident and Kubeflow are two of the three major components of the NetApp AI Control Plane; the third is the data fabric powered by NetApp, which is covered later in this paper.

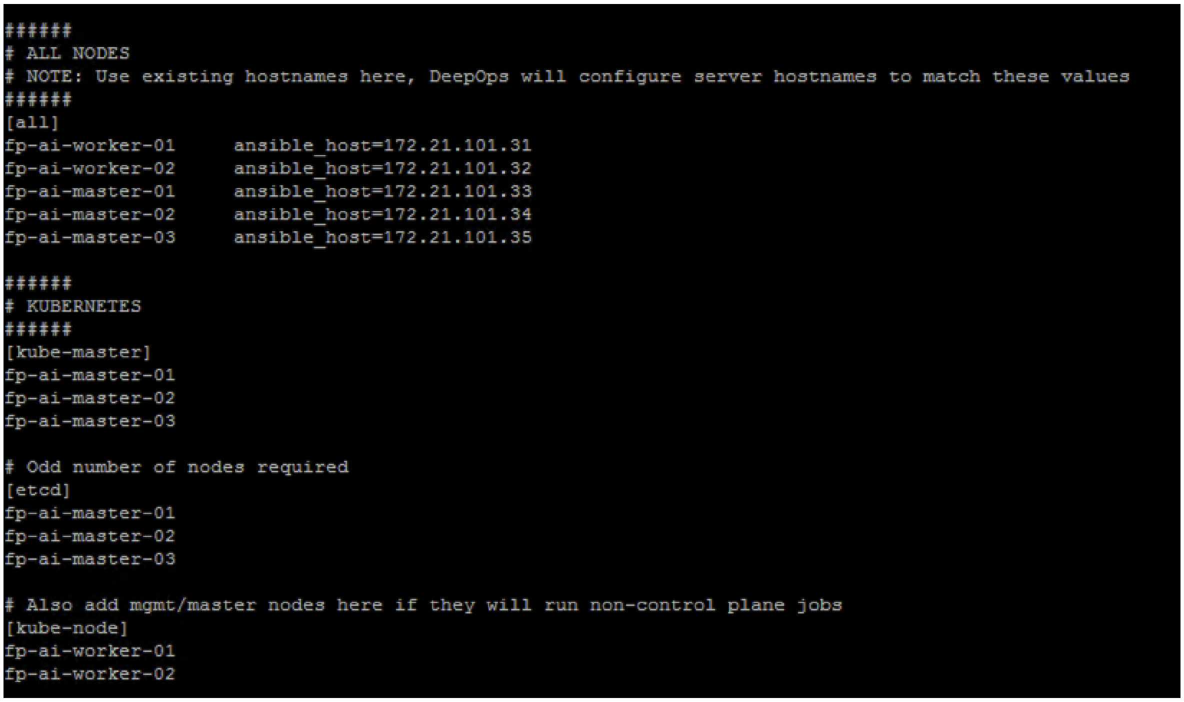
Figure 13) NetApp AI Control Plane—FlexPod.



Containers are among the most popular methods of packaging and deploying AI applications and workloads and leveraging Kubernetes to host containerized applications is the preferred choice for most customers.

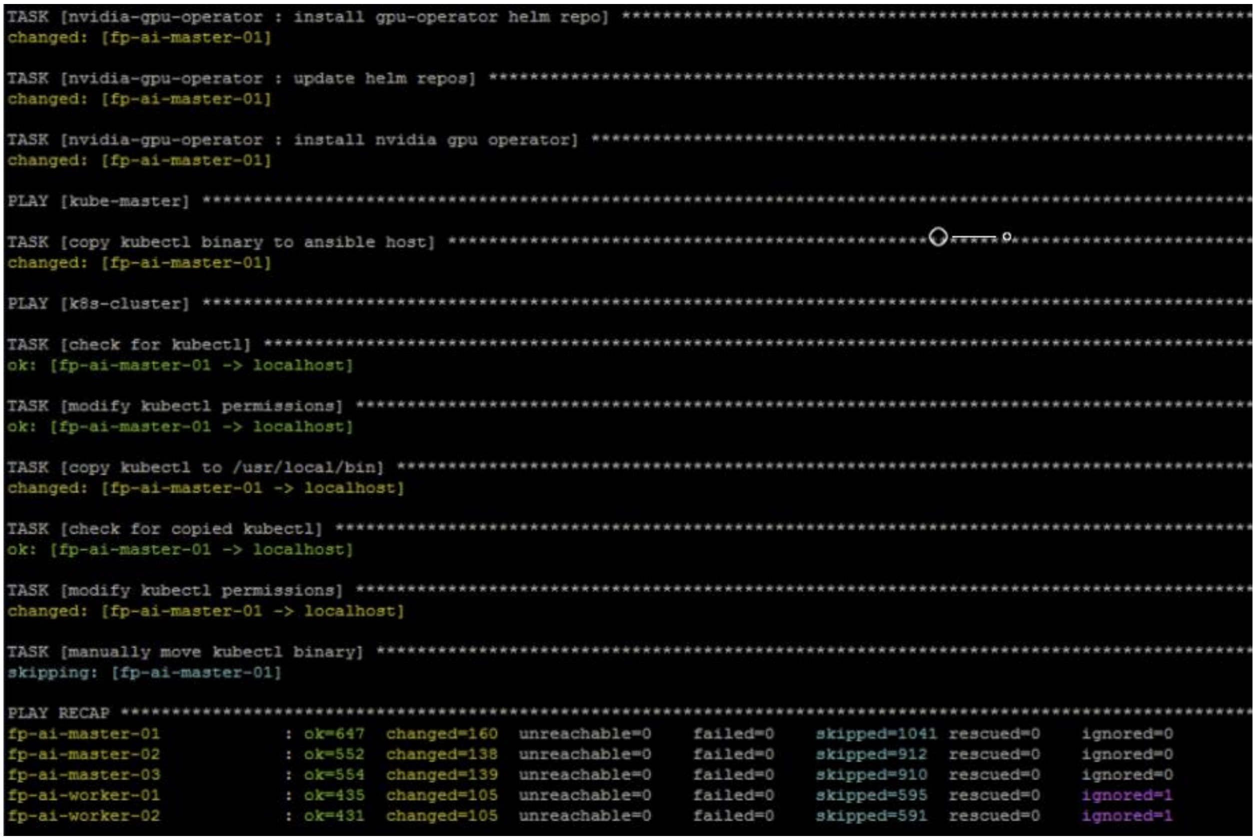
NVIDIA DeepOps provides an Ansible playbook that you can use to set up a Kubernetes cluster with GPU-based worker nodes. In addition to deploying Kubernetes, DeepOps also configures the GPU drivers, NVIDIA docker container runtime, and GPU device plugins, resulting in a Kubernetes cluster that is ready to host deployments that will need GPU acceleration.

The master and worker nodes to be used for the setup are declared in an inventory file as shown following image.



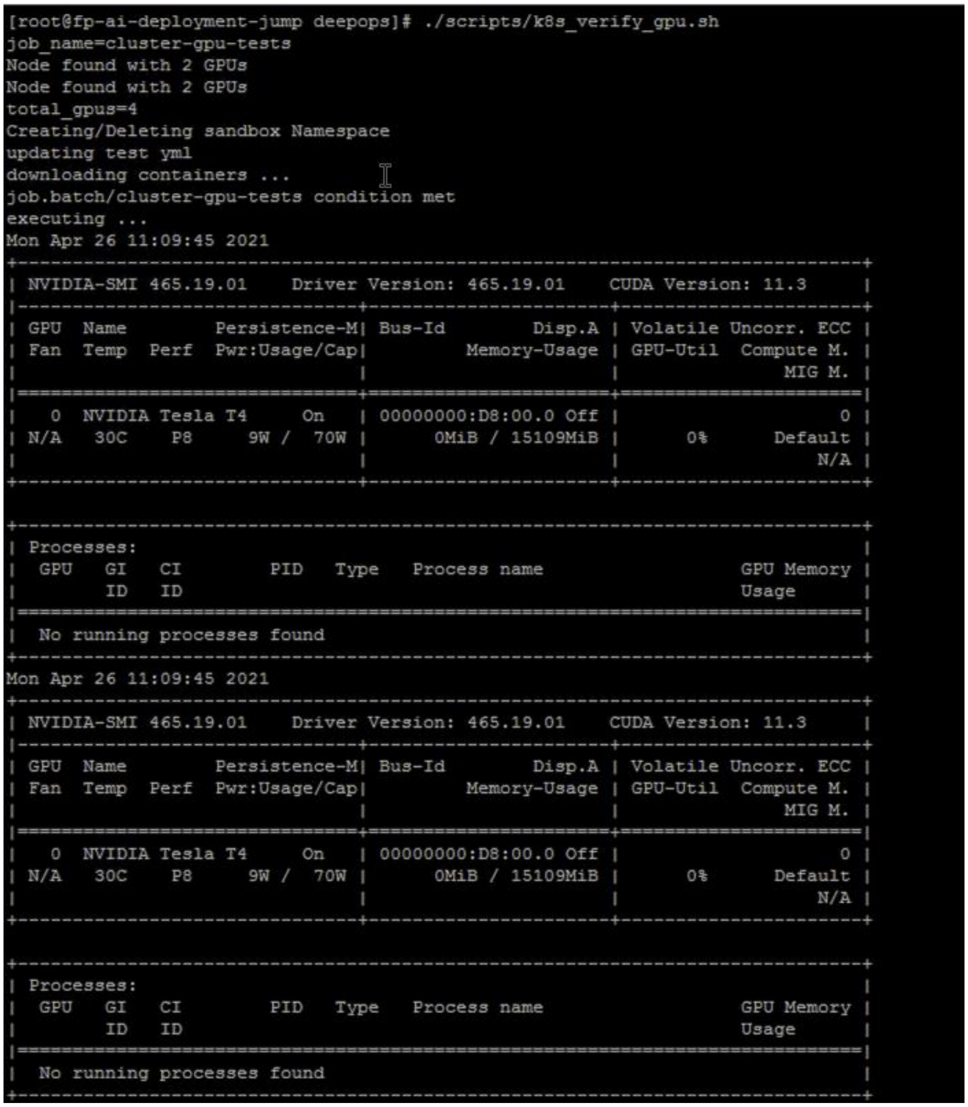
```
######
# ALL NODES
# NOTE: Use existing hostnames here, DeepOps will configure server hostnames to match these values
######
[all]
fp-ai-worker-01     ansible_host=172.21.101.31
fp-ai-worker-02     ansible_host=172.21.101.32
fp-ai-master-01     ansible_host=172.21.101.33
fp-ai-master-02     ansible_host=172.21.101.34
fp-ai-master-03     ansible_host=172.21.101.35

######
# KUBERNETES
######
[kube-master]
fp-ai-master-01
fp-ai-master-02
fp-ai-master-03

# Odd number of nodes required
[etcd]
fp-ai-master-01
fp-ai-master-02
fp-ai-master-03

# Also add mgmt/master nodes here if they will run non-control plane jobs
[kube-node]
fp-ai-worker-01
fp-ai-worker-02
```

The following image shows the status of the playbook execution after the Kubernetes cluster is set up.



```
TASK [nvidia-gpu-operator : install gpu-operator helm repo] **************************************************************
changed: [fp-ai-master-01]

TASK [nvidia-gpu-operator : update helm repos] ***************************************************************************
changed: [fp-ai-master-01]

TASK [nvidia-gpu-operator : install nvidia gpu operator] *****************************************************************
changed: [fp-ai-master-01]

PLAY [kube-master] *******************************************************************************************************

TASK [copy kubectl binary to ansible host] *******************************************************************************
changed: [fp-ai-master-01]

PLAY [k8s-cluster] *******************************************************************************************************

TASK [check for kubectl] *************************************************************************************************
ok: [fp-ai-master-01 -> localhost]

TASK [modify kubectl permissions] ****************************************************************************************
ok: [fp-ai-master-01 -> localhost]

TASK [copy kubectl to /usr/local/bin] ************************************************************************************
changed: [fp-ai-master-01 -> localhost]

TASK [check for copied kubectl] ******************************************************************************************
ok: [fp-ai-master-01 -> localhost]

TASK [modify kubectl permissions] ****************************************************************************************
changed: [fp-ai-master-01 -> localhost]

TASK [manually move kubectl binary] **************************************************************************************
skipping: [fp-ai-master-01]

PLAY RECAP ***************************************************************************************************************
fp-ai-master-01            : ok=647  changed=160  unreachable=0    failed=0    skipped=1041 rescued=0    ignored=0
fp-ai-master-02            : ok=552  changed=138  unreachable=0    failed=0    skipped=912  rescued=0    ignored=0
fp-ai-master-03            : ok=554  changed=139  unreachable=0    failed=0    skipped=910  rescued=0    ignored=0
fp-ai-worker-01            : ok=435  changed=105  unreachable=0    failed=0    skipped=595  rescued=0    ignored=1
fp-ai-worker-02            : ok=431  changed=105  unreachable=0    failed=0    skipped=591  rescued=0    ignored=1
```

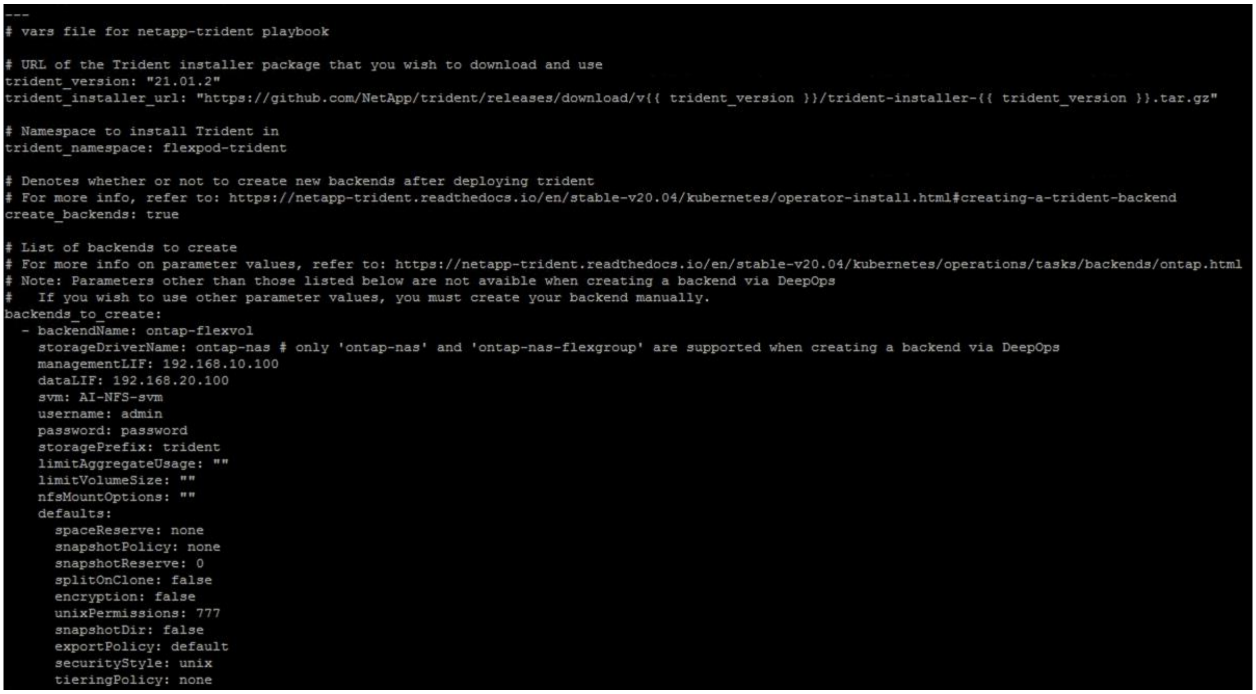
Access to GPUs in the worker nodes can be verified by running a shell script.



```
[root@fp-ai-deployment-jump deepops]# ./scripts/k8s_verify_gpu.sh
job_name=cluster-gpu-tests
Node found with 2 GPUs
Node found with 2 GPUs
total_gpus=4
Creating/Deleting sandbox Namespace
updating test yml
downloading containers ...
job.batch/cluster-gpu-tests condition met
executing ...
Mon Apr 26 11:09:45 2021
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 465.19.01    Driver Version: 465.19.01    CUDA Version: 11.3     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  NVIDIA Tesla T4     On   | 00000000:D8:00.0 Off |                    0 |
| N/A   30C    P8     9W /  70W |      0MiB / 15109MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
Mon Apr 26 11:09:45 2021
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 465.19.01    Driver Version: 465.19.01    CUDA Version: 11.3     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  NVIDIA Tesla T4     On   | 00000000:D8:00.0 Off |                    0 |
| N/A   30C    P8     9W /  70W |      0MiB / 15109MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

NetApp Trident integrates tightly with Kubernetes to enable users to request and manage persistent volumes using native Kubernetes interfaces and constructs. It is designed in such a way that users such as data scientists and system administrators can take advantage of the underlying capabilities of the storage infrastructure without having to know anything about it. Trident enables microservices and containerized applications to use enterprise-class storage services such as QoS, storage efficiencies, and cloning to meet the persistent storage demands of applications.

An Ansible playbook to set up Trident is available in the DeepOps package. You can define the ONTAP storage parameters in a variable file, as shown in the following image:



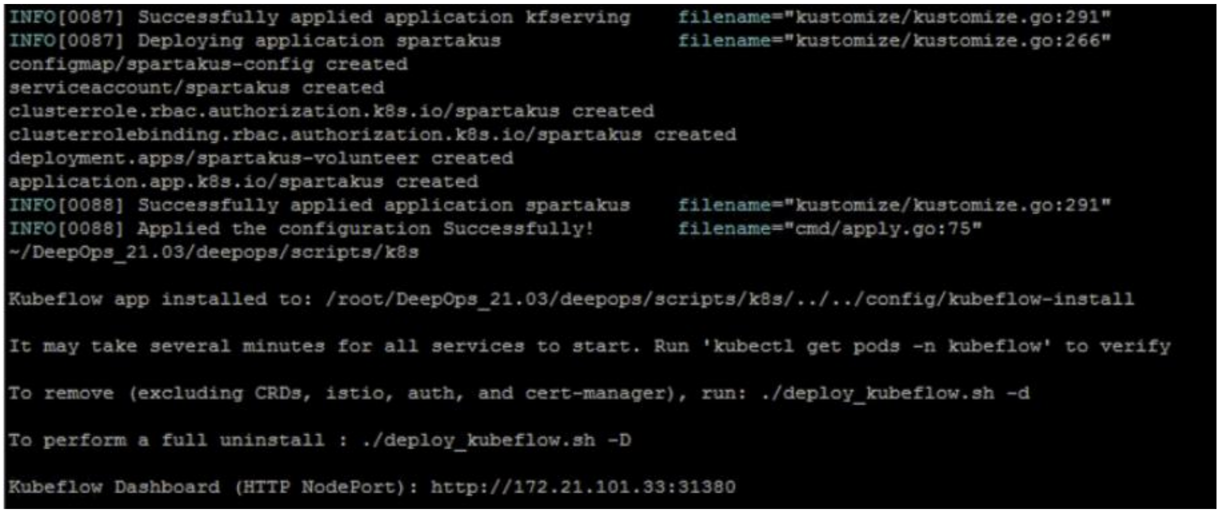
```
---
# vars file for netapp-trident playbook

# URL of the Trident installer package that you wish to download and use
trident_version: "21.01.2"
trident_installer_url: "https://github.com/NetApp/trident/releases/download/v{{ trident_version }}/trident-installer-{{ trident_version }}.tar.gz"

# Namespace to install Trident in
trident_namespace: flexpod-trident

# Denotes whether or not to create new backends after deploying trident
# For more info, refer to: https://netapp-trident.readthedocs.io/en/stable-v20.04/kubernetes/operator-install.html#creating-a-trident-backend
create_backends: true

# List of backends to create
# For more info on parameter values, refer to: https://netapp-trident.readthedocs.io/en/stable-v20.04/kubernetes/operations/tasks/backends/ontap.html
# Note: Parameters other than those listed below are not avaible when creating a backend via DeepOps
#   If you wish to use other parameter values, you must create your backend manually.
backends_to_create:
  - backendName: ontap-flexvol
    storageDriverName: ontap-nas # only 'ontap-nas' and 'ontap-nas-flexgroup' are supported when creating a backend via DeepOps
    managementLIF: 192.168.10.100
    dataLIF: 192.168.20.100
    svm: AI-NFS-svm
    username: admin
    password: password
    storagePrefix: trident
    limitAggregateUsage: ""
    limitVolumeSize: ""
    nfsMountOptions: ""
    defaults:
      spaceReserve: none
      snapshotPolicy: none
      snapshotReserve: 0
      splitOnClone: false
      encryption: false
      unixPermissions: 777
      snapshotDir: false
      exportPolicy: default
      securityStyle: unix
      tieringPolicy: none
```

After NetApp Trident has been set up, you can install Kubeflow.

The features of Trident when combined with Kubeflow further simplify the deployment of AI workloads and enhance the user experience for data scientists. Kubeflow abstracts away the intricacies of Kubernetes, enabling data scientists to focus on what they know best—data science.

Data scientists no longer need to double up as Kubernetes administrators. They do not need to know how to define Kubernetes deployments in YAML or execute `kubectl` commands. With Kubeflow, they can define end-to-end AI/ML/DL workflows using a simple Python SDK. Because most data scientists are already familiar with Python through the use of AI frameworks such as TensorFlow and PyTorch, the learning curve is not steep.

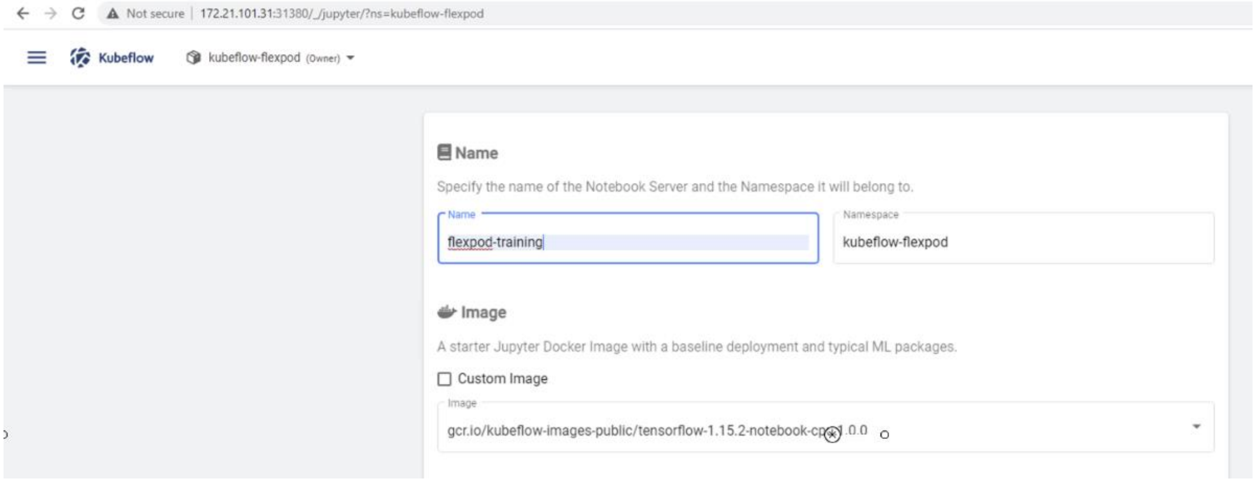
DeepOps provides a shell scrip to deploy Kubeflow. After the setup process is complete, it provides a link to the Kubeflow dashboard, as shown in the following image:



```
INFO[0087] Successfully applied application kfserving    filename="kustomize/kustomize.go:291"
INFO[0087] Deploying application spartakus               filename="kustomize/kustomize.go:266"
configmap/spartakus-config created
serviceaccount/spartakus created
clusterrole.rbac.authorization.k8s.io/spartakus created
clusterrolebinding.rbac.authorization.k8s.io/spartakus created
deployment.apps/spartakus-volunteer created
application.app.k8s.io/spartakus created
INFO[0088] Successfully applied application spartakus    filename="kustomize/kustomize.go:291"
INFO[0088] Applied the configuration Successfully!       filename="cmd/apply.go:75"
~/DeepOps_21.03/deepops/scripts/k8s

Kubeflow app installed to: /root/DeepOps_21.03/deepops/scripts/k8s/../../config/kubeflow-install

It may take several minutes for all services to start. Run 'kubectl get pods -n kubeflow' to verify

To remove (excluding CRDs, istio, auth, and cert-manager), run: ./deploy_kubeflow.sh -d

To perform a full uninstall : ./deploy_kubeflow.sh -D

Kubeflow Dashboard (HTTP NodePort): http://172.21.101.33:31380
```

Jupyter Notebooks are included with Kubeflow out of the box facilitating on-demand provisioning and destroying of Jupyter Notebook Servers / Workspaces. With Kubeflow being deployed as part of the NetApp AI Control Plane, data volumes, potentially containing petabytes worth of data, can be presented to data scientists as simple folders within a Jupyter Workspace. The data scientist is given instant access to all of their data from within a familiar interface. Data scientists never even need to know that the data resides on NetApp storage.

You can create a notebook server with a custom container image or a standard image by logging into the Kubeflow dashboard through a browser.


Not secure | 172.21.101.31:31380/_/jupyter/?ns=kubeflow-flexpod
Kubeflow
kubeflow-flexpod (Owner)

Name

Specify the name of the Notebook Server and the Namespace it will belong to.

Name: flexpod-training

Namespace: kubeflow-flexpod

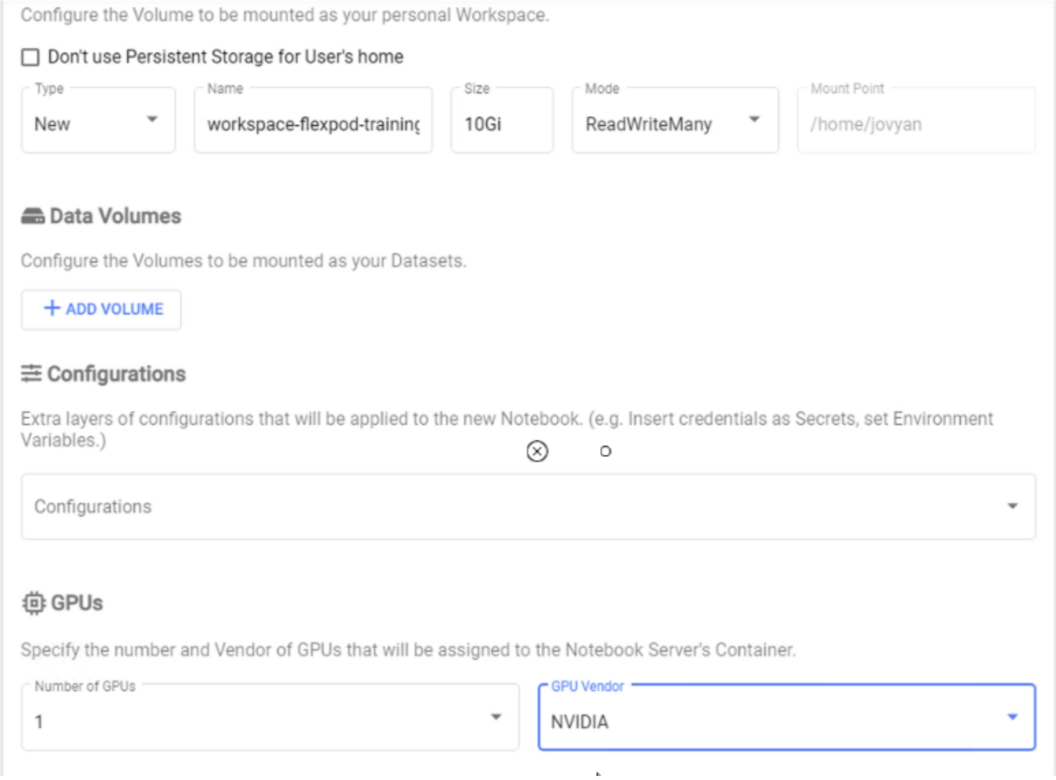
Image

A starter Jupyter Docker Image with a baseline deployment and typical ML packages.

□ Custom Image

Image: gcr.io/kubeflow-images-public/tensorflow-1.15.2-notebook-cpu:1.0.0

You can provision and mount a persistent volume to the notebook using NetApp Trident at the backend. You can also assign the notebook the desired number of GPUs.



Configure the Volume to be mounted as your personal Workspace.

□ Don't use Persistent Storage for User's home

Type	Name	Size	Mode	Mount Point
New	workspace-flexpod-training	10Gi	ReadWriteMany	/home/jovyan

Data Volumes

Configure the Volumes to be mounted as your Datasets.

Configurations

Extra layers of configurations that will be applied to the new Notebook. (e.g. Insert credentials as Secrets, set Environment Variables.)

Configurations

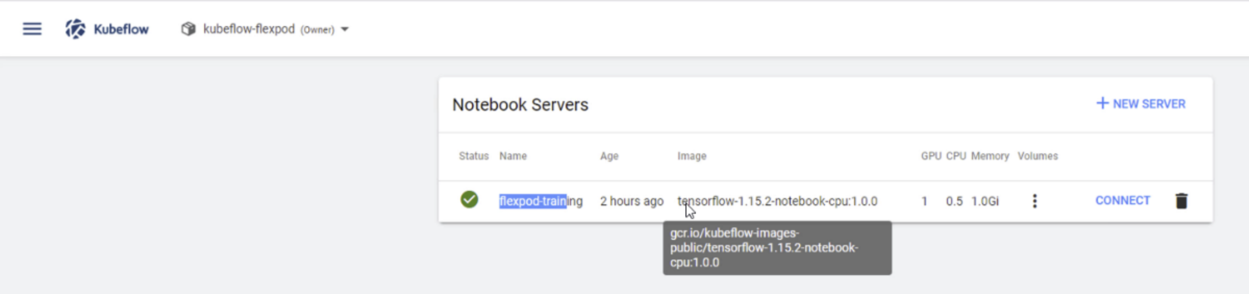
GPUs

Specify the number and Vendor of GPUs that will be assigned to the Notebook Server's Container.

Number of GPUs: 1

GPU Vendor: NVIDIA

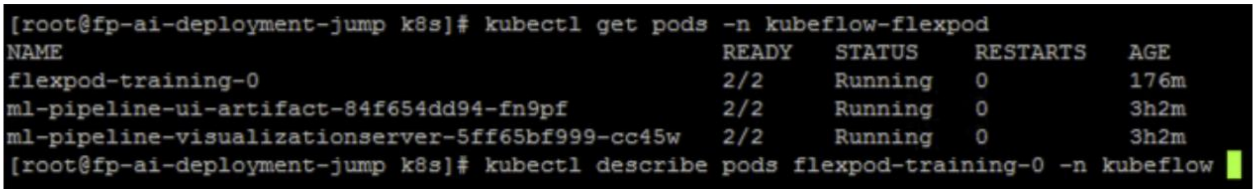
The notebook server is created with the desired image and resources.



Status	Name	Age	Image	GPU	CPU	Memory	Volumes
✓	flexpod-training	2 hours ago	tensorflow-1.15.2-notebook-cpu:1.0.0	1	0.5	1.0Gi	⋮

gcr.io/kubeflow-images-public/tensorflow-1.15.2-notebook-cpu:1.0.0

The notebook server corresponds to a pod/container in the Kubernetes environment.



```
[root@fp-ai-deployment-jump k8s]# kubectl get pods -n kubeflow-flexpod
NAME                                            READY   STATUS    RESTARTS   AGE
flexpod-training-0                              2/2     Running   0          176m
ml-pipeline-ui-artifact-84f654dd94-fn9pf        2/2     Running   0          3h2m
ml-pipeline-visualizationserver-5ff65bf999-cc45w   2/2     Running   0          3h2m
[root@fp-ai-deployment-jump k8s]# kubectl describe pods flexpod-training-0 -n kubeflow
```

With Kubeflow Pipelines, users can build portable and scalable AI/ML workflows that promote the reuse of components and pipelines to quickly build end-to-end solutions.

While there is no fixed approach on how a pipeline needs to be defined or what steps it needs to execute, some typical examples for AI/ML include the following:

- Execute end-to-end AI training with traceability and versioning.
- Clone training datasets for workspaces to be used by data scientists.
- Replicate newly acquired data for training from an edge location to the core by using SnapMirror technology.
- Replicate data from a variety of file and object storage platforms on-premises or in the cloud to the FlexPod AI Datacenter.

The NetApp AI Control Plane provides the following benefits:

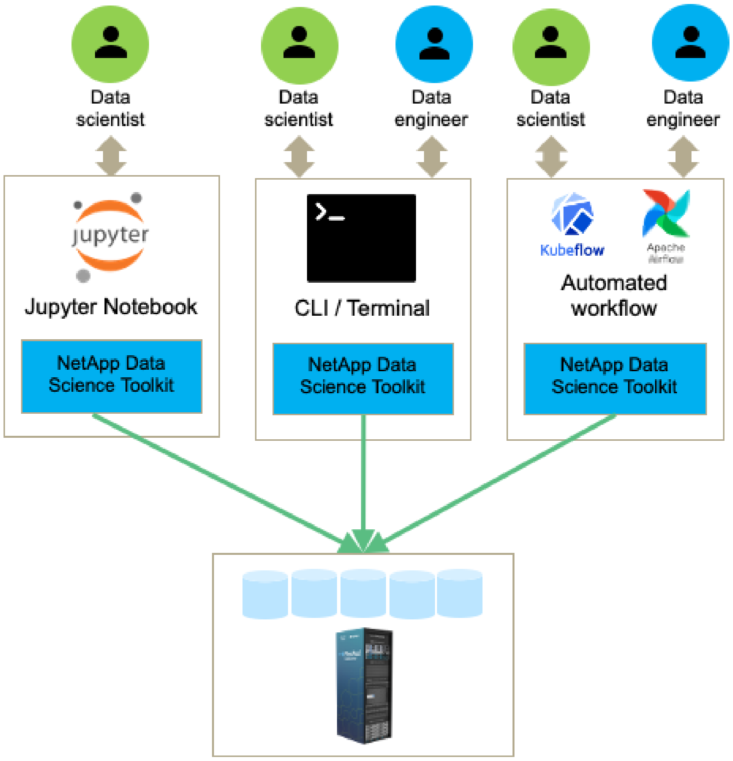
- Automated processes to set up the AI infrastructure by using NVIDIA DeepOps, NetApp Trident, and Kubeflow.
- Microservices friendly architecture.
- Data scientists need not know the technical details of Kubernetes and ONTAP storage but still stand to benefit from both by consuming the resources in a simplified approach.
- Ability to operate in a hybrid cloud environment with data fabric powered by NetApp.
- Improved collaboration between data scientists and their associated data with Kubeflow Pipelines.
- Simplified data management with ONTAP, Trident, and data fabric powered by NetApp.

NetApp Data Science Toolkit

The NetApp Data Science Toolkit is a Python program that makes it simple for data scientists and engineers to perform advanced data management tasks. The program can function as a CLI utility or be presented as a library of functions that can be imported into any Python program or Jupyter Notebook.

The NetApp Data Science Toolkit enhances the NetApp AI Control Plane solution by further simplifying data management. Data scientists who have a Jupyter Notebook that was provisioned through the NetApp AI Control Plane can use this toolkit to implement a data management task in one simple line of Python code. You can also add these tasks as a step in the Kubeflow Pipelines automated workflow.

Figure 14) NetApp Data Science Toolkit.



While the NetApp AI Control Plane is built on top of Kubernetes, you can use the Data Science Toolkit with traditional environments as well as Kubernetes. In cases where Kubernetes is not being used, you must use the toolkit in conjunction with a NetApp data storage system.

Below are some of the capabilities of the toolkit that can be executed individually or added to a pipeline:

- Clone a data volume
- Create a Snapshot copy for a data volume
- Pull the contents of a bucket from Simple Storage Service (S3) to a data volume on-premises
- Prepopulate specific files or directories on a FlexCache volume
- Clone a JupyterLab workspace
- Create a Snapshot copy for a JupyterLab workspace
- Create a new persistent volume
- Create a Snapshot copy for a persistent volume

More operations and capabilities are available and documented in GitHub.

Users can clone the toolkit to their environment from GitHub and use it from Linux or macOS hosts.

Extend the data fabric to FlexPod AI

Data centers have grown well beyond the physical constraints of a single site; organizations have their data residing at multiple sites (on-premises and cloud) across the globe in a hybrid cloud model. It is imperative that organizations have a well-defined standard to enable data mobility across the different locations, including the public cloud. Especially for workloads such as AI/ML/DL, it is critical for organizations and businesses to have the capability to bring the desired data to a designated location at a specific time in the most secure manner possible and make it available for the applications. With the portfolio of products for data fabric powered by NetApp, customers can build their own data fabric with their data endpoints across the globe and ensure that there are well-defined data pathways that help them in building a robust data management framework for their business.

FlexPod customers can start building their data fabric by integrating the ONTAP instance that is running in the FlexPod platform with the product portfolio of your data fabric powered by NetApp.

Building the dataset

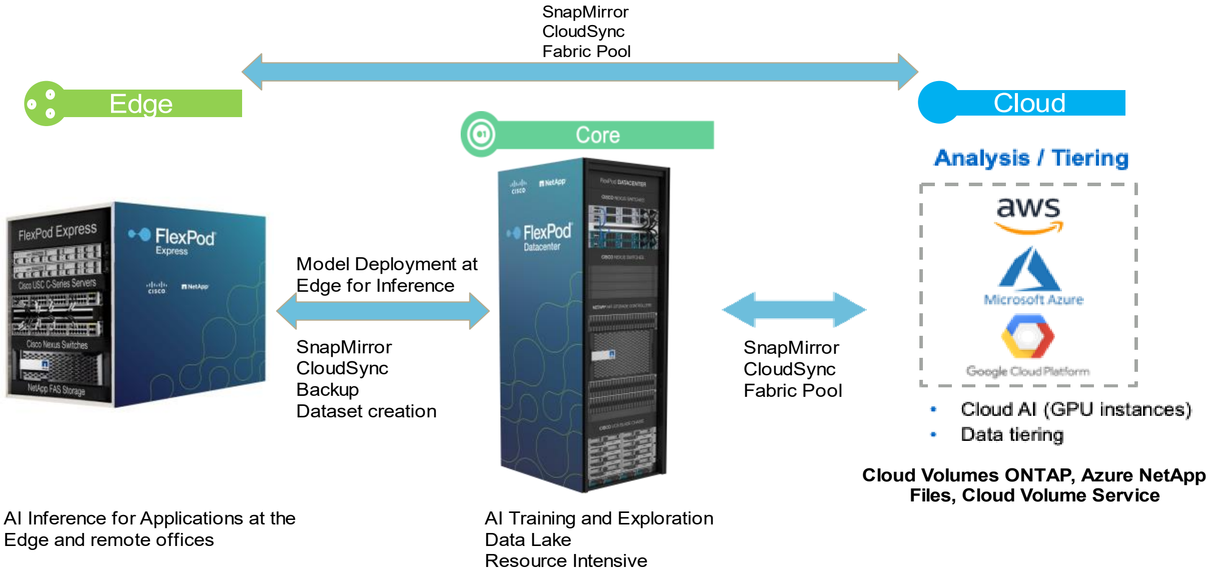
The training dataset holds the key to a successful implementation of an AI project. An entire AI project can fail if the dataset is not good enough despite having skilled data scientists and a compelling use case. Most organizations plan to use AI to enhance their products and services to eventually add more value to their business. To accomplish this, they usually need access to data that is specific to their line of business, and there is no better source than their own data. Although organizations might have the data they need, often that data is hard to access and locked out, making data collection more difficult. In order to gain insights, the data must be diverse in nature and gathered from multiple sources. This is not a one-time operation and is part of the larger AI data lifecycle where the dataset must be enriched with data that has the potential to further improve the insights and the associated products. Organizations should build a data collection strategy that enables them to funnel the data that is being generated by their products and services to build a dataset that is unique to them and holds the insights that they need and differentiates them from their competitors.

It is evident that there is a critical need to create data connections between the silos and foster a data-driven culture in every organization. Data fabric powered by NetApp addresses this challenge by helping organizations to build a global data resource that can be accessed from anywhere at any time.

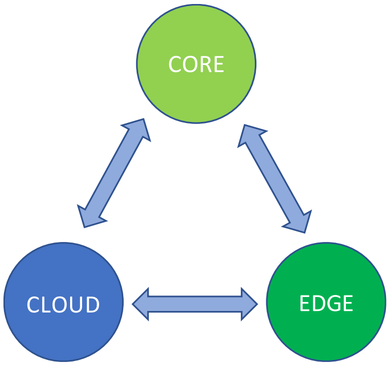
Data mobility: Edge to cloud to core

The following figure shows how a data fabric for AI/ML/DL can be built with FlexPod platforms running at the core and edge locations, coupled with the public cloud service providers. The applications running in edge locations feed data to the AI dataset that is used for continuous DL training. While most of these high resource-intensive operations occur at the core on-premises data center, the data pathways must span between these three tiers.

Figure 15) FlexPod data fabric.



The data/storage consumption pattern at each of these three tiers varies considerably in an AI lifecycle. The edge locations are subject to a lot of inferencing and data generation by the applications. There is a need for ultra-low latency and concurrency that can be managed by FlexPod Express running NetApp AFF arrays and Cisco compute with GPU acceleration. If the data created at the edge is not going to be used immediately, you can tier it to low-cost storage in the cloud by using FabricPool or back it up or replicate it to the cloud by using features such as Cloud Backup, SnapMirror, and Cloud Sync. When the time is appropriate to use the data, you can restore the data directly from the cloud to the FlexPod Datacenter ONTAP system running in the core for AI operations.



For training operations, the datasets can reach enormous sizes and there is a demand for a scale-out storage system that can host multiple petabytes of data under a single namespace. The FlexPod Datacenter in the core powered by ONTAP and AFF arrays deliver this capability through FlexGroup volumes. The data required to build the dataset can be mobilized from the edge and cloud by using SnapMirror and Cloud Sync. NetApp Public Cloud services such as NetApp Cloud Volumes ONTAP, NetApp Cloud Volumes Service, and Azure NetApp Files are available across all the major public cloud service providers, thereby enabling a common language and medium for building data pathways across the edge, core, and cloud.

Conclusion

FlexPod is a proven converged infrastructure solution for enterprises and small and midsized businesses. Customers rely on the FlexPod solution to host their traditional workloads/applications, and as part of their IT transformation need not look beyond their FlexPod platforms to host their AI operations. The CVDs on FlexPod for AI/ML/DL showcase the infrastructure capabilities and architecture design. This white paper showcases how FlexPod delivers a simplified user experience for data scientists by abstracting the infrastructure layer and providing them with a simplified usage model through automation and orchestration. FlexPod also provides the unique value add of being able to support the AI lifecycle of training and inferencing across edge and core locations while seamlessly integrating with the public cloud through data fabric powered by NetApp, thereby enabling customers to build their FlexPod AI data fabric.

Where to find additional information

To learn more about the information that is described in this document, review the following documents and/or websites:

- NetApp product documentation
  https://www.netapp.com/support-and-training/documentation/
- NetApp Trident
  https://netapp.io/persistent-storage-provisioner-for-kubernetes/

- Data fabric powered by NetApp
https://www.netapp.com/data-fabric/
- NVIDIA Ampere
https://developer.nvidia.com/blog/nvidia-ampere-architecture-in-depth/
- NVIDIA Multi Instance GPU
https://www.nvidia.com/en-in/technologies/multi-instance-gpu/
- Cisco UCS
https://www.cisco.com/c/en_in/products/servers-unified-computing/index.html
- NVIDIA DeepOps
https://github.com/NVIDIA/deepops
- Kubeflow
https://www.kubeflow.org

Version history

Version	Date	Document version history
Version 1.0	June 2021	Initial release.

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

Copyright Information

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

Trademark Information

WP-7345-0621

NetApp