Overview

HCI

NetApp
October 01, 2020

Table of Contents

Overview

This section describes the steps required to deploy the AI inferencing platform using NetApp HCI. The following list provides the high-level tasks involved in the setup:

1. Configure network switches
2. Deploy the VMware virtual infrastructure on NetApp HCI using NDE
3. Configure the H615c compute nodes to be used as K8 worker nodes
4. Set up the deployment jump VM and K8 master VMs
5. Deploy a Kubernetes cluster with NVIDIA DeepOps
6. Deploy ONTAP Select within the virtual infrastructure
7. Deploy NetApp Trident
8. Deploy NVIDIA Triton inference Server
9. Deploy the client for the Triton inference server
10. Collect inference metrics from the Triton inference server

Next: Configure Network Switches