



# Architecture

## NetApp Solutions

Kevin Hoke  
April 20, 2021

This PDF was generated from [https://docs.netapp.com/us-en/netapp-solutions/ai/hciai\\_edge\\_architecture.html](https://docs.netapp.com/us-en/netapp-solutions/ai/hciai_edge_architecture.html) on September 24, 2021. Always check docs.netapp.com for the latest.

# Table of Contents

- Architecture ..... 1
  - Solution Technology ..... 1
  - Architectural Diagram ..... 1
  - Hardware Requirements ..... 2
  - Software Requirements ..... 3

# Architecture

## Solution Technology

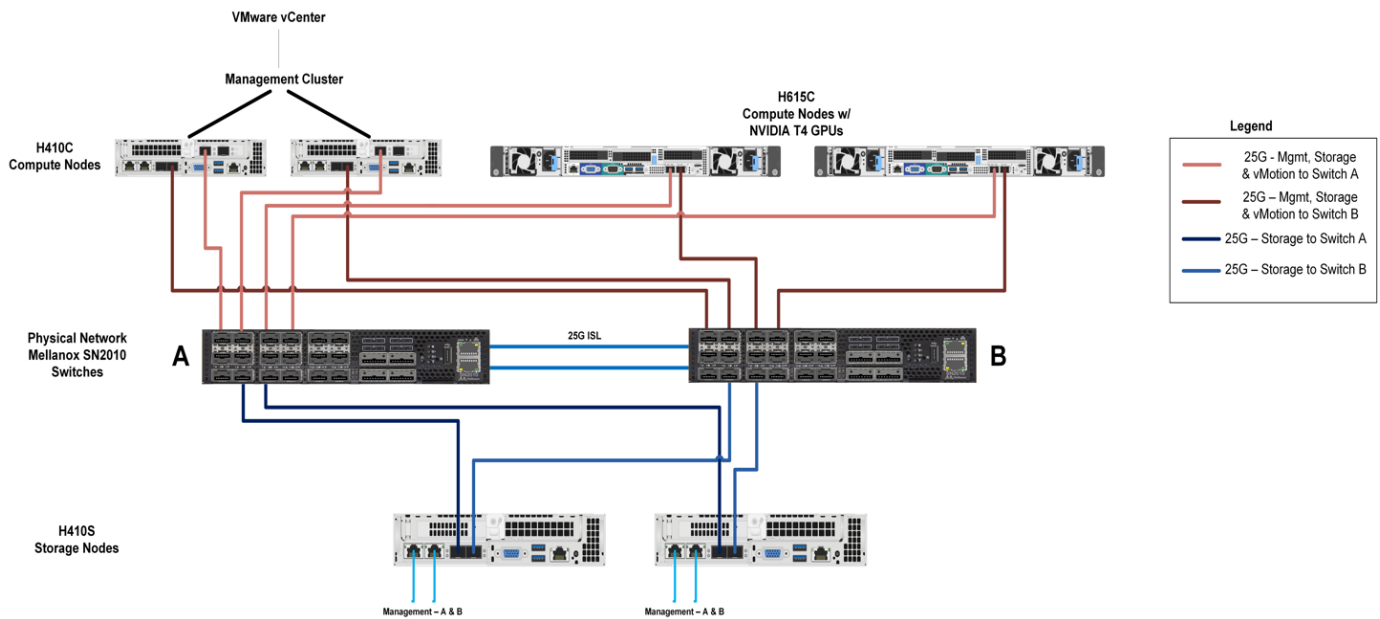
This solution is designed with a NetApp HCI system that contains the following components:

- Two H615c compute nodes with NVIDIA T4 GPUs
- Two H410c compute nodes
- Two H410s storage nodes
- Two Mellanox SN2010 10GbE/25GbE switches

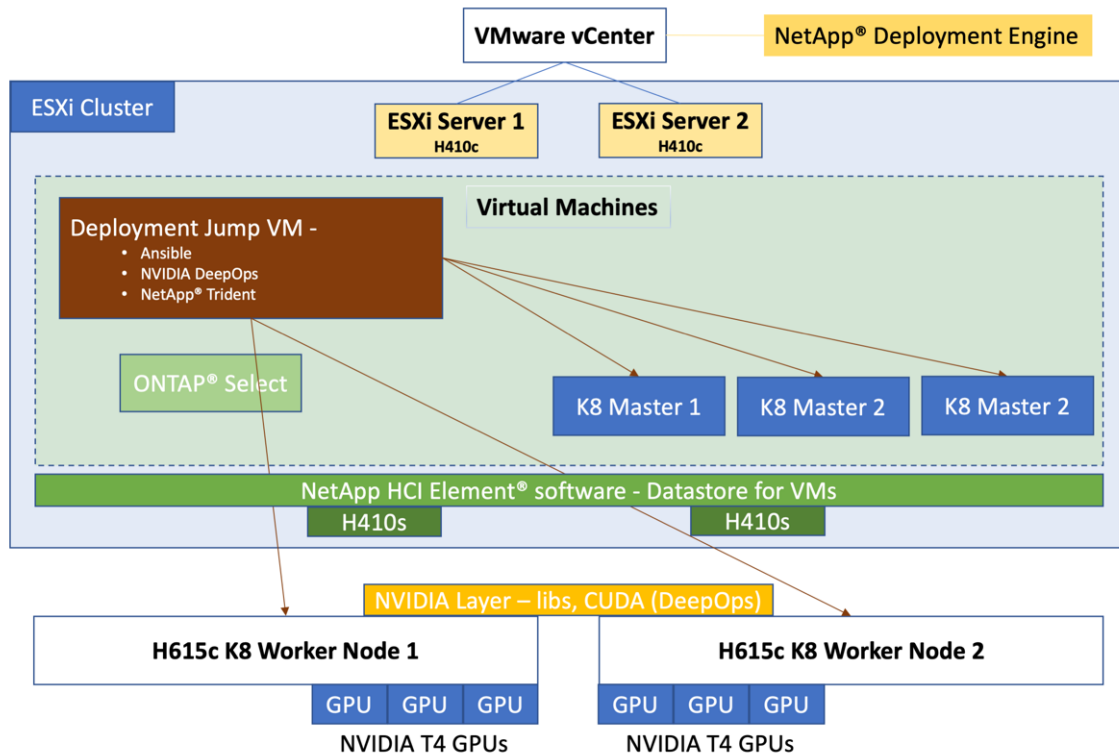
## Architectural Diagram

The following diagram illustrates the solution architecture for the NetApp HCI AI inferencing solution.

NetApp HCI Architecture design for AI Inferencing



The following diagram illustrates the virtual and physical elements of this solution.



A VMware infrastructure is used to host the management services required by this inferencing solution. These services do not need to be deployed on a dedicated infrastructure; they can coexist with any existing workloads. The NetApp Deployment Engine (NDE) uses the H410c and H410s nodes to deploy the VMware infrastructure.

After NDE has completed the configuration, the following components are deployed as VMs in the virtual infrastructure:

- **Deployment Jump VM.** Used to automate the deployment of NVIDIA DeepOps. See [NVIDIA DeepOps](#) and storage management using NetApp Trident.
- **ONTAP Select.** An instance of ONTAP Select is deployed to provide NFS file services and persistent storage to the AI workload running on Kubernetes.
- **Kubernetes Masters.** During deployment, three VMs are installed and configured with a supported Linux distribution and configured as Kubernetes master nodes. After the management services have been set up, two H615c compute nodes with NVIDIA T4 GPUs are installed with a supported Linux distribution. These two nodes function as the Kubernetes worker nodes and provide the infrastructure for the inferencing platform.

## Hardware Requirements

The following table lists the hardware components that are required to implement the solution. The hardware components that are used in any particular implementation of the solution might vary based on customer requirements.

| Layer   | Product Family | Quantity | Details                         |
|---------|----------------|----------|---------------------------------|
| Compute | H615c          | 2        | 3 NVIDIA Tesla T4 GPUs per node |

| Layer   | Product Family  | Quantity | Details                                     |
|---------|-----------------|----------|---|
|         | H410c           | 2        | Compute nodes for management infrastructure |
| Storage | H410s           | 2        | Storage for OS and workload                 |
| Network | Mellanox SN2010 | 2        | 10G/25G switches                            |

## Software Requirements

The following table lists the software components that are required to implement the solution. The software components that are used in any particular implementation of the solution might vary based on customer requirements.

| Layer                | Software                       | Version   |
|----------------------|--------------------------------|---|
| Storage              | NetApp Element software        | 12.0.0.333  |
|                      | ONTAP Select                   | 9.7   |
|                      | NetApp Trident                 | 20.07   |
| NetApp HCI engine    | NDE                            | 1.8   |
| Hypervisor           | Hypervisor                     | VMware vSphere ESXi 6.7U1   |
|                      | Hypervisor Management System   | VMware vCenter Server 6.7U1   |
| Inferencing Platform | NVIDIA DeepOps                 | 20.08   |
|                      | NVIDIA GPU Operator            | 1.1.7   |
|                      | Ansible                        | 2.9.5   |
|                      | Kubernetes                     | 1.17.9  |
|                      | Docker                         | Docker CE 18.09.7   |
|                      | CUDA Version                   | 10.2  |
|                      | GPU Device Plugin              | 0.6.0   |
|                      | Helm                           | 3.1.2   |
|                      | NVIDIA Tesla Driver            | 440.64.00   |
|                      | NVIDIA Triton Inference Server | 2.1.0 – NGC Container v20.07  |
| K8 Master VMs        | Linux                          | Any supported distribution across NetApp IMT, NVIDIA DeepOps, and GPUOperator<br><br>Ubuntu 18.04.4 LTS was used in this solution<br>Kernel version: 4.15 |

| Layer                    | Software | Version   |
|--------------------------|----------|---|
| Host OS/ K8 Worker Nodes | Linux    | Any supported distribution across NetApp IMT, NVIDIA DeepOps, and GPUOperator<br><br>Ubuntu 18.04.4 LTS was used in this solution<br>Kernel version: 4.15 |

[Next: Design Considerations](#)

## Copyright Information

Copyright © 2021 NetApp, Inc. All rights reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means-graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system-without prior written permission of the copyright owner.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:

THIS SOFTWARE IS PROVIDED BY NETAPP "AS IS" AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WHICH ARE HEREBY DISCLAIMED. IN NO EVENT SHALL NETAPP BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

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.

RESTRICTED RIGHTS LEGEND: Use, duplication, or disclosure by the government is subject to restrictions as set forth in subparagraph (c)(1)(ii) of the Rights in Technical Data and Computer Software clause at DFARS 252.277-7103 (October 1988) and FAR 52-227-19 (June 1987).

## Trademark Information

NETAPP, the NETAPP logo, and the marks listed at <http://www.netapp.com/TM> are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.