NVIDIA AI Enterprise with NetApp and VMware
NetApp Solutions

NetApp
April 03, 2024

This PDF was generated from https://docs.netapp.com/us-en/netapp-solutions/ai/nvaie_introduction.html on April 03, 2024. Always check docs.netapp.com for the latest.
# Table of Contents

NVIDIA AI Enterprise with NetApp and VMware ............................................. 1
  NVIDIA AI Enterprise with NetApp and VMware ............................................. 1
  Technology Overview ......................................................................................... 1
  Architecture ....................................................................................................... 3
  Initial Setup ......................................................................................................... 4
  Utilize NVIDIA NGC Software ......................................................................... 5
  Where to Find Additional Information .............................................................. 9
NVIDIA AI Enterprise with NetApp and VMware

For IT architects and admins, AI tooling can be complicated and unfamiliar. Additionally, many AI platforms are not enterprise-ready. NVIDIA AI Enterprise, powered by NetApp and VMware, was created to deliver a streamlined, enterprise-class AI architecture.

NVIDIA AI Enterprise is an end-to-end, cloud-native suite of AI and data analytics software that is optimized, certified, and supported by NVIDIA to run on VMware vSphere with NVIDIA-Certified Systems. This software facilitates the simple and rapid deployment, management, and scaling of AI workloads in the modern hybrid cloud environment. NVIDIA AI Enterprise, powered by NetApp and VMware, delivers enterprise-class AI workload and data management in a simplified, familiar package.

Technology Overview

NVIDIA AI Enterprise

NVIDIA AI Enterprise is an end-to-end, cloud-native suite of AI and data analytics software that is optimized, certified, and supported by NVIDIA to run on VMware vSphere with NVIDIA-Certified Systems. This software facilitates the simple and rapid deployment, management, and scaling of AI workloads in the modern hybrid cloud environment.
cloud environment.

**NVIDIA GPU Cloud (NGC)**

NVIDIA NGC hosts a catalog of GPU-optimized software for AI practitioners to develop their AI solutions. It also provides access to various AI services including NVIDIA Base Command for model training, NVIDIA Fleet Command to deploy and monitor models, and the NGC Private Registry for securely accessing and managing proprietary AI software. Also, NVIDIA AI Enterprise customers can request support through the NGC portal.

**VMware vSphere**

VMware vSphere is VMware’s virtualization platform, which transforms data centers into aggregated computing infrastructures that include CPU, storage, and networking resources. vSphere manages these infrastructures as a unified operating environment, and provides administrators with the tools to manage the data centers that participate in that environment.

The two core components of vSphere are ESXi and vCenter Server. ESXi is the virtualization platform where administrators create and run virtual machines and virtual appliances. vCenter Server is the service through which administrators manage multiple hosts connected in a network and pool host resources.

**NetApp ONTAP**

ONTAP 9, the latest generation of storage management software from NetApp, enables businesses to modernize infrastructure and transition to a cloud-ready data center. Leveraging industry-leading data management capabilities, ONTAP enables the management and protection of data with a single set of tools, regardless of where that data resides. You can also move data freely to wherever it is needed: the edge, the core, or the cloud. ONTAP 9 includes numerous features that simplify data management, accelerate, and protect critical data, and enable next generation infrastructure capabilities across hybrid cloud architectures.

**Simplify data management**

Data management is crucial to enterprise IT operations and data scientists so that appropriate resources are used for AI applications and training AI/ML datasets. The following additional information about NetApp technologies is out of scope for this validation but might be relevant depending on your deployment.

ONTAP data management software includes the following features to streamline and simplify operations and reduce your total cost of operation:

- Inline data compaction and expanded deduplication. Data compaction reduces wasted space inside storage blocks, and deduplication significantly increases effective capacity. This applies to data stored locally and data tiered to the cloud.

- Minimum, maximum, and adaptive quality of service (AQoS). Granular quality of service (QoS) controls help maintain performance levels for critical applications in highly shared environments.

- NetApp FabricPool. Provides automatic tiering of cold data to public and private cloud storage options, including Amazon Web Services (AWS), Azure, and NetApp StorageGRID storage solution. For more information about FabricPool, see TR-4598: FabricPool best practices.

**Accelerate and protect data**

ONTAP delivers superior levels of performance and data protection and extends these capabilities in the following ways:

- Performance and lower latency. ONTAP offers the highest possible throughput at the lowest possible
latency.

• Data protection. ONTAP provides built-in data protection capabilities with common management across all platforms.

• NetApp Volume Encryption (NVE). ONTAP offers native volume-level encryption with both onboard and External Key Management support.

• Multitenancy and multifactor authentication. ONTAP enables sharing of infrastructure resources with the highest levels of security.

**Future-proof infrastructure**

ONTAP helps meet demanding and constantly changing business needs with the following features:

• Seamless scaling and nondisruptive operations. ONTAP supports the nondisruptive addition of capacity to existing controllers and to scale-out clusters. Customers can upgrade to the latest technologies, such as NVMe and 32Gb FC, without costly data migrations or outages.

• Cloud connection. ONTAP is the most cloud-connected storage management software, with options for software-defined storage (ONTAP Select) and cloud-native instances (NetApp Cloud Volumes Service) in all public clouds.

• Integration with emerging applications. ONTAP offers enterprise-grade data services for next generation platforms and applications, such as autonomous vehicles, smart cities, and Industry 4.0, by using the same infrastructure that supports existing enterprise apps.

**NetApp DataOps Toolkit**

The NetApp DataOps Toolkit is a Python-based tool that simplifies the management of development/training workspaces and inference servers that are backed by high-performance, scale-out NetApp storage. Key capabilities include:

• Rapidly provision new high-capacity JupyterLab workspaces that are backed by high-performance, scale-out NetApp storage.

• Rapidly provision new NVIDIA Triton Inference Server instances that are backed by enterprise-class NetApp storage.

• Near-instantaneously clone high-capacity JupyterLab workspaces in order to enable experimentation or rapid iteration.

• Near-instantaneously save snapshots of high-capacity JupyterLab workspaces for backup and/or traceability/baselining.

• Near-instantaneously provision, clone, and snapshot high-capacity, high-performance data volumes.

**Architecture**

This solution builds upon a proven and familiar architecture featuring NetApp, VMware, and NVIDIA-Certified Systems. See the following table for details.

<table>
<thead>
<tr>
<th>Component</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI and Data Analytics Software</td>
<td>NVIDIA AI Enterprise for VMware</td>
</tr>
<tr>
<td>Virtualization Platform</td>
<td>VMware vSphere</td>
</tr>
<tr>
<td>Compute Platform</td>
<td>NVIDIA-Certified Systems</td>
</tr>
</tbody>
</table>
Initial Setup

This section describes the initial setup tasks that need to be performed in order to utilize NVIDIA AI Enterprise with NetApp and VMware.

Prerequisites

Before you perform the steps that are outlined in this section, we assume that you have already deployed VMware vSphere and NetApp ONTAP. Refer to the NVIDIA AI Enterprise Product Support Matrix for details on
supported vSphere versions. Refer to the NetApp and VMware solution documentation for details on deploying VMware vSphere with NetApp ONTAP.

**Install NVIDIA AI Enterprise Host Software**

To install the NVIDIA AI Enterprise host software, follow the instructions outlined in sections 1-4 in the NVIDIA AI Enterprise Quick Start Guide.

**Utilize NVIDIA NGC Software**

This section describes the tasks that need to be performed in order to utilize NVIDIA NGC enterprise software within an NVIDIA AI Enterprise environment.

**Setup**

This section describes the initial setup tasks that need to be performed in order to utilize NVIDIA NGC enterprise software within an NVIDIA AI Enterprise environment.

**Prerequisites**

Before you perform the steps that are outlined in this section, we assume that you have already deployed the NVIDIA AI Enterprise host software by following the instructions outlined on the Initial Setup page.

**Create an Ubuntu Guest VM with vGPU**

First, you must create an Ubuntu 20.04 guest VM with vGPU. To create an Ubuntu 20.04 guest VM with vGPU, follow the instructions outlined in the NVIDIA AI Enterprise Deployment Guide.

**Download and Install NVIDIA Guest Software**

Next, you must install the required NVIDIA guest software within the guest VM that you created in the previous step. To download and install the required NVIDIA guest software within the guest VM, follow the instructions outlined in sections 5.1-5.4 in the NVIDIA AI Enterprise Quick Start Guide.

> When performing the verification tasks outlined in section 5.4, you may need to use a different CUDA container image version tag as the CUDA container image has been updated since the writing of the guide. In our validation, we used 'nvidia/cuda:11.0.3-base-ubuntu20.04'.

**Download AI/Analytics Framework Container(s)**

Next, you must download needed AI or analytics framework container images from NVIDIA NGC so that they will be available within your guest VM. To download framework containers within the guest VM, follow the instructions outlined in the NVIDIA AI Enterprise Deployment Guide.

**Install and Configure the NetApp DataOps Toolkit**

Next, you must install the NetApp DataOps Toolkit for Traditional Environments within the guest VM. The NetApp DataOps Toolkit can be used to manage scale-out data volumes on your ONTAP system directly from the terminal within the guest VM. To install the NetApp DataOps Toolkit within the guest VM, perform the following tasks.

1. Install pip.
2. Log out of the guest VM terminal and then log back in.

3. Configure the NetApp DataOps Toolkit. In order to complete this step, you will need API access details for your ONTAP system. You may need to obtain these from your storage admin.

```bash
$ netapp_dataops_cli.py config

Enter ONTAP management LIF hostname or IP address (Recommendation: Use SVM management interface): 172.22.10.10
Enter SVM (Storage VM) name: NVAIE-client
Enter SVM NFS data LIF hostname or IP address: 172.22.13.151
Enter default volume type to use when creating new volumes (flexgroup/flexvol) [flexgroup]:
Enter export policy to use by default when creating new volumes [default]:
Enter snapshot policy to use by default when creating new volumes [none]:
Enter unix filesystem user id (uid) to apply by default when creating new volumes (ex. '0' for root user) [0]:
Enter unix filesystem group id (gid) to apply by default when creating new volumes (ex. '0' for root group) [0]:
Enter unix filesystem permissions to apply by default when creating new volumes (ex. '0777' for full read/write permissions for all users and groups) [0777]:
Enter aggregate to use by default when creating new FlexVol volumes: aff_a400_01_NVME_SSD_1
Enter ONTAP API username (Recommendation: Use SVM account): admin
Enter ONTAP API password (Recommendation: Use SVM account):
Verify SSL certificate when calling ONTAP API (true/false): false
Do you intend to use this toolkit to trigger BlueXP Copy and Sync operations? (yes/no): no
Do you intend to use this toolkit to push/pull from S3? (yes/no): no
Created config file: '/home/user/.netapp_dataops/config.json'.
```

Create a Guest VM template

Lastly, you must create a VM template based on your guest VM. You will be able to use this template to quickly create guest VMs for utilizing NVIDIA NGC software.

To create a VM template based on your guest VM, log into VMware vSphere, right-click on the guest VM name, choose 'Clone', choose 'Clone to Template...', and then follow the wizard.
Example Use Case - TensorFlow Training Job

This section describes the tasks that need to be performed in order to execute a TensorFlow training job within an NVIDIA AI Enterprise environment.

Prerequisites

Before you perform the steps that are outlined in this section, we assume that you have already created a guest VM template by following the instructions outlined on the Setup page.

Create Guest VM from Template

First, you must create a new guest VM from the template that you created in the previous section. To create a new guest VM from your template, log into VMware vSphere, right-click on the template name, choose 'New VM from This Template...', and then follow the wizard.
Create and Mount Data Volume

Next, you must create a new data volume on which to store your training dataset. You can quickly create a new data volume using the NetApp DataOps Toolkit. The example command that follows shows the creation of a volume named 'imagenet' with a capacity of 2 TB.

$ netapp_dataops_cli.py create vol -n imagenet -s 2TB

Before you can populate your data volume with data, you must mount it within the guest VM. You can quickly mount a data volume using the NetApp DataOps Toolkit. The example command that follows shows the mounting of the volume that was created in the previous step.

$ sudo -E netapp_dataops_cli.py mount vol -n imagenet -m ~/imagenet

Populate Data Volume

After the new volume has been provisioned and mounted, the training dataset can be retrieved from the source location and placed on the new volume. This typically will involve pulling the data from an S3 or Hadoop data lake and sometimes will involve help from a data engineer.

Execute TensorFlow Training Job

Now, you are ready to execute your TensorFlow training job. To execute your TensorFlow training job, perform the following tasks.

1. Pull the NVIDIA NGC enterprise TensorFlow container image.

   $ sudo docker pull nvcr.io/nvaie/tensorflow-2-1:22.05-tf1-nvaie-2.1-py3

2. Launch an instance of the NVIDIA NGC enterprise TensorFlow container. Use the '-v' option to attach your data volume to the container.

   $ sudo docker run --gpus all -v ~/imagenet:/imagenet -it --rm nvcr.io/nvaie/tensorflow-2-1:22.05-tf1-nvaie-2.1-py3

3. Execute your TensorFlow training program within the container. The example command that follows shows the execution of an example ResNet-50 training program that is included in the container image.

   $ python ./nvidia-examples/cnn/resnet.py --layers 50 -b 64 -i 200 -u batch --precision fp16 --data_dir /imagenet/data

Where to Find Additional Information

To learn more about the information described in this document, refer to the following
documents and/or websites:

• NetApp ONTAP data management software — ONTAP information library

• NetApp DataOps Toolkit
  https://github.com/NetApp/netapp-dataops-toolkit

• NVIDIA AI Enterprise with VMware

Acknowledgments

• Bobby Oommen, Sr. Manager, NetApp
• Ramesh Isaac, Systems Administrator, NetApp
• Roney Daniel, Technical Marketing Engineer, NetApp