NVIDIA AI Enterprise

Product Support Matrix

Table of Contents

Chapter 1. Product Support Matrix

Driver package: NVIDIA AI Enterprise 4.0

Validated partner integrations:

▶ Run: AI: 2.5

▶ Domino Data Lab

> 💬 **Note:** The GPUs listed in the tables are supported only with <u>NVIDIA AI Enterprise compatible</u> servers.

▶ <u>Amazon Web Services (AWS) Nitro Support</u>

▶ <u>Google Cloud Platform (GCP) KVM Support</u>

▶ <u>Linux with KVM Support</u>

▶ <u>Red Hat Enterprise Linux with KVM Support</u>

▶ <u>Red Hat OpenShift Support</u>

▶ <u>SUSE Linux Enterprise Server Support</u>

▶ <u>Ubuntu Support</u>

▶ <u>VMware vSphere Support</u>

Amazon Web Services (AWS) Nitro Support

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA A100 ▶ NVIDIA A10G ▶ NVIDIA T4 ▶ NVIDIA V100	Amazon Web Services (AWS) Nitro		▶ Ubuntu 22.04 LTS ▶ Ubuntu 20.04 LTS	containerd	Amazon Elastic Kubernetes Service (EKS)

Google Cloud Platform (GCP) KVM Support

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA A100 ▶ NVIDIA T4 ▶ NVIDIA V100	Google Cloud Platform (GCP) KVM		▶ Ubuntu 22.04 LTS ▶ Ubuntu 20.04 LTS	containerd	Google Kubernetes Engine (GKE)

Linux with KVM Support

NVIDIA AI Enterprise is supported on Linux with KVM platforms **only** by specific hypervisor software vendors. For information about which NVIDIA AI Enterprise releases and hypervisor software releases are supported, consult the documentation from your hypervisor vendor.

Hypervisor Vendor	Platform	Additional Information
Nutanix	AHV	Obtain the NVIDIA Virtual GPU Manager software directly from Nutanix through the <u>My Nutanix</u> portal (My Nutanix account required). 🗨 **Note:** If the NVIDIA AI Enterprise release that you need is not available from the My Nutanix portal, contact Nutanix. Then follow the instructions on the My Nutanix portal to obtain the correct NVIDIA AI Enterprise graphics drivers from the NVIDIA Licensing Portal.

Red Hat Enterprise Linux with KVM Support

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA A800 PCIe 80GB	▶ Red Hat Enterprise Linux with KVM 9.2, 9.0	▶ NVIDIA vGPU ▶ GPU pass through	▶ Red Hat OpenShift 4.11 and	CRI-O	▶ Red Hat OpenShift 4.11 and later using RHCOS

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA A800 PCIe 80GB liquid cooled ▶ NVIDIA A800 HGX 80GB ▶ NVIDIA A800 40GB PCIe active cooled ▶ NVIDIA AX800[1] ▶ NVIDIA A100X ▶ NVIDIA A100 PCIe 40GB ▶ NVIDIA A100 HGX 40GB ▶ NVIDIA A100 PCIe 80GB ▶ NVIDIA A100 PCIe 80GB liquid cooled ▶ NVIDIA A100 HGX 80GB ▶ NVIDIA A40	▶ Red Hat Enterprise Linux with KVM 8.8, 8.6		later using RHCOS		▶ Upstream Kubernetes 1.21 through 1.25
			▶ Red Hat Enterprise Linux 9.2, 9.0 ▶ Red Hat Enterprise Linux 8.8, 8.6 ▶ Red Hat Enterprise Linux 7.9	Podman	N/A
			▶ Microsoft Windows Server 2022, 2019	N/A	N/A

---

[1] The NVIDIA AX800 GPU is supported only on Linux OSes. Windows is **not** supported.

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA A30X					
▶ NVIDIA A30					
▶ NVIDIA A30 liquid cooled					
▶ NVIDIA A10					
▶ NVIDIA A16					
▶ NVIDIA A2					
▶ NVIDIA H800 PCIe 94GB (H800 NVL)					
▶ NVIDIA H800 PCIe 80GB					
▶ NVIDIA H100 PCIe 94GB (H100 NVL)					
▶ NVIDIA H100 PCIe 80GB					
▶ NVIDIA L40					
▶ NVIDIA L40S					
▶ NVIDIA L4					
▶ NVIDIA RTX 6000 Ada					

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA RTX 5000 Ada ▶ NVIDIA RTX A6000 ▶ NVIDIA RTX A5500 ▶ NVIDIA RTX A5000 ▶ NVIDIA RTX 6000 passive ▶ NVIDIA RTX 8000 passive ▶ NVIDIA T4 ▶ NVIDIA V100					
▶ NVIDIA A800 PCIe 80GB ▶ NVIDIA A800 PCIe 80GB liquid cooled ▶ NVIDIA A800 HGX 80GB	▶ Red Hat Enterprise Linux 9.2, 9.0 ▶ Red Hat Enterprise Linux 8.8, 8.6 ▶ Red Hat Enterprise Linux 7.9	▶ Bare metal	N/A	Podman	N/A
▶ NVIDIA A800 40GB PCIe active cooled	Red Hat Enterprise Linux 8.8, 8.6	▶ Bare metal	N/A	containerd	HPE Ezmeral Runtime Enterprise 5.5

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA AX800[1]					
▶ NVIDIA A100X					
▶ NVIDIA A100 PCIe 40GB					
▶ NVIDIA A100 HGX 40GB					
▶ NVIDIA A100 PCIe 80GB					
▶ NVIDIA A100 PCIe 80GB liquid cooled					
▶ NVIDIA A100 HGX 80GB					
▶ NVIDIA A40					
▶ NVIDIA A30X					
▶ NVIDIA A30					
▶ NVIDIA A30 liquid cooled					
▶ NVIDIA A10					
▶ NVIDIA A16					
▶ NVIDIA A2					
▶ NVIDIA H800 PCIe					

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
94GB (H800 NVL)					
▶ NVIDIA H800 PCIe 80GB					
▶ NVIDIA H100 PCIe 94GB (H100 NVL)					
▶ NVIDIA H100 PCIe 80GB					
▶ NVIDIA H100 SXM5 94GB[2]					
▶ NVIDIA H100 SXM5 80GB[2]					
▶ NVIDIA H100 SXM5 64GB[2]					
▶ NVIDIA L40					
▶ NVIDIA L40S					
▶ NVIDIA L4					
▶ NVIDIA RTX 6000 Ada					

---

[2] NVIDIA H100 SXM5 GPUs are supported only in bare-metal deployments on Red Hat Enterprise Linux and Ubuntu with NVIDIA Datacenter Drivers. NVIDIA vGPU guest drivers are not supported.

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
► NVIDIA RTX 5000 Ada ► NVIDIA RTX A6000 ► NVIDIA RTX A5500 ► NVIDIA RTX A5000 ► NVIDIA RTX 6000 passive ► NVIDIA RTX 8000 passive ► NVIDIA T4 ► NVIDIA V100					

Red Hat OpenShift Support

> 💬 **Note:** NVIDIA AI Enterprise supports every patch release for the listed Red Hat OpenShift release provided that Red Hat also supports it. When a release or patch release is no longer supported by Red Hat, it is no longer supported by NVIDIA AI Enterprise.

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
► NVIDIA A800 PCIe 80GB ► NVIDIA A800 PCIe 80GB	► Red Hat OpenShift 4.11 and later using RHCOS	► Bare metal	N/A	CRI-O	► Red Hat OpenShift 4.11 and later using RHCOS

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
liquid cooled					
▶ NVIDIA A800 HGX 80GB					
▶ NVIDIA A800 40GB PCIe active cooled					
▶ NVIDIA AX800 1					
▶ NVIDIA A100X					
▶ NVIDIA A100 PCIe 40GB					
▶ NVIDIA A100 HGX 40GB					
▶ NVIDIA A100 PCIe 80GB					
▶ NVIDIA A100 PCIe 80GB liquid cooled					
▶ NVIDIA A100 HGX 80GB					
▶ NVIDIA A40					
▶ NVIDIA A30X					
▶ NVIDIA A30					

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA A30 liquid cooled					
▶ NVIDIA A10					
▶ NVIDIA A16					
▶ NVIDIA A2					
▶ NVIDIA H800 PCIe 94GB (H800 NVL)					
▶ NVIDIA H800 PCIe 80GB					
▶ NVIDIA H100 PCIe 94GB (H100 NVL)					
▶ NVIDIA H100 PCIe 80GB					
▶ NVIDIA H100 SXM5 94GB2					
▶ NVIDIA H100 SXM5 80GB2					
▶ NVIDIA H100 SXM5 64GB2					
▶ NVIDIA L40					

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA L40S ▶ NVIDIA L4 ▶ NVIDIA RTX 6000 Ada ▶ NVIDIA RTX 5000 Ada ▶ NVIDIA RTX A6000 ▶ NVIDIA RTX A5500 ▶ NVIDIA RTX A5000 ▶ NVIDIA RTX 6000 passive ▶ NVIDIA RTX 8000 passive ▶ 5000 ▶ NVIDIA T4 ▶ NVIDIA V100					

SUSE Linux Enterprise Server Support

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA A800 PCIe 80GB	▶ SUSE Linux Enterprise	▶ Bare metal	N/A	containerd	▶ Upstream Kubernetes 1.21

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA A800 PCIe 80GB liquid cooled	Server 15 SP2+				through 1.25
▶ NVIDIA A800 HGX 80GB				Docker	N/A
▶ NVIDIA A800 40GB PCIe active cooled					
▶ NVIDIA AX800[1]					
▶ NVIDIA A100X					
▶ NVIDIA A100 PCIe 40GB					
▶ NVIDIA A100 HGX 40GB					
▶ NVIDIA A100 PCIe 80GB					
▶ NVIDIA A100 PCIe 80GB liquid cooled					
▶ NVIDIA A100 HGX 80GB					
▶ NVIDIA A40					
▶ NVIDIA A30X					

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA A30					
▶ NVIDIA A30 liquid cooled					
▶ NVIDIA A10					
▶ NVIDIA A16					
▶ NVIDIA A2					
▶ NVIDIA H800 PCIe 94GB (H800 NVL)					
▶ NVIDIA H800 PCIe 80GB					
▶ NVIDIA H100 PCIe 94GB (H100 NVL)					
▶ NVIDIA H100 PCIe 80GB					
▶ NVIDIA L40					
▶ NVIDIA L40S					
▶ NVIDIA L4					
▶ NVIDIA RTX 6000 Ada					
▶ NVIDIA RTX 5000 Ada					

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA RTX A6000 ▶ NVIDIA RTX A5500 ▶ NVIDIA RTX A5000 ▶ NVIDIA RTX 6000 passive ▶ NVIDIA RTX 8000 passive ▶ NVIDIA T4 ▶ NVIDIA V100					

Ubuntu Support

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA A800 PCIe 80GB ▶ NVIDIA A800 PCIe 80GB liquid cooled ▶ NVIDIA A800 HGX 80GB ▶ NVIDIA A800 40GB PCIe	▶ Ubuntu 22.04 LTS ▶ Ubuntu 20.04 LTS	▶ NVIDIA vGPU ▶ GPU pass through	▶ Ubuntu 22.04 LTS ▶ Ubuntu 20.04 LTS	containerd	▶ Upstream Kubernetes 1.21 through 1.25 ▶ VMware vSphere with Tanzu 7.0 U3c
				Docker	N/A

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
active cooled					
▶ NVIDIA AX800[1]					
▶ NVIDIA A100X					
▶ NVIDIA A100 PCIe 40GB					
▶ NVIDIA A100 HGX 40GB					
▶ NVIDIA A100 PCIe 80GB					
▶ NVIDIA A100 PCIe 80GB liquid cooled					
▶ NVIDIA A100 HGX 80GB					
▶ NVIDIA A40					
▶ NVIDIA A30X					
▶ NVIDIA A30					
▶ NVIDIA A30 liquid cooled					
▶ NVIDIA A10					
▶ NVIDIA A16					
▶ NVIDIA A2					

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA H800 PCIe 94GB (H800 NVL)					
▶ NVIDIA H800 PCIe 80GB					
▶ NVIDIA H100 PCIe 94GB (H100 NVL)					
▶ NVIDIA H100 PCIe 80GB					
▶ NVIDIA L40					
▶ NVIDIA L40S					
▶ NVIDIA L4					
▶ NVIDIA RTX 6000 Ada					
▶ NVIDIA RTX 5000 Ada					
▶ NVIDIA RTX A6000					
▶ NVIDIA RTX A5500					
▶ NVIDIA RTX A5000					

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA RTX 6000 passive ▶ NVIDIA RTX 8000 passive ▶ NVIDIA T4 ▶ NVIDIA V100					
▶ NVIDIA A800 PCIe 80GB ▶ NVIDIA A800 PCIe 80GB liquid cooled ▶ NVIDIA A800 HGX 80GB ▶ NVIDIA A800 40GB PCIe active cooled ▶ NVIDIA AX800 1 ▶ NVIDIA A100X ▶ NVIDIA A100 PCIe 40GB ▶ NVIDIA A100 HGX 40GB	▶ Ubuntu 22.04 LTS ▶ Ubuntu 20.04 LTS	▶ Bare metal	N/A	containerd	▶ Upstream Kubernetes 1.21 through 1.25 ▶ VMware vSphere with Tanzu 7.0 U3c
				Docker	N/A

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA A100 PCIe 80GB					
▶ NVIDIA A100 PCIe 80GB liquid cooled					
▶ NVIDIA A100 HGX 80GB					
▶ NVIDIA A40					
▶ NVIDIA A30X					
▶ NVIDIA A30					
▶ NVIDIA A30 liquid cooled					
▶ NVIDIA A10					
▶ NVIDIA A16					
▶ NVIDIA A2					
▶ NVIDIA H800 PCIe 94GB (H800 NVL)					
▶ NVIDIA H800 PCIe 80GB					
▶ NVIDIA H100 PCIe 94GB (H100 NVL)					

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA H100 PCIe 80GB					
▶ NVIDIA H100 SXM5 94GB2					
▶ NVIDIA H100 SXM5 80GB2					
▶ NVIDIA H100 SXM5 64GB2					
▶ NVIDIA L40					
▶ NVIDIA L40S					
▶ NVIDIA L4					
▶ NVIDIA RTX 6000 Ada					
▶ NVIDIA RTX 5000 Ada					
▶ NVIDIA RTX A6000					
▶ NVIDIA RTX A5500					
▶ NVIDIA RTX A5000					
▶ NVIDIA RTX 6000 passive					

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA RTX 8000 passive ▶ NVIDIA T4 ▶ NVIDIA V100					

VMware vSphere Support

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA A800 PCIe 80GB ▶ NVIDIA A800 PCIe 80GB liquid cooled ▶ NVIDIA A800 HGX 80GB ▶ NVIDIA A800 40GB PCIe active cooled ▶ NVIDIA AX800[1] ▶ NVIDIA A100X ▶ NVIDIA A100 PCIe 40GB ▶ NVIDIA A100 HGX 40GB	▶ VMware vSphere Hypervisor (ESXi) Enterprise Plus Edition 8.0 VMware vCenter Server 8.0 ▶ VMware vSphere Hypervisor (ESXi) Enterprise Plus Edition 7.0 Update 3 VMware vCenter Server 7.0 Update 3 ▶ VMware vSphere Hypervisor (ESXi)	▶ NVIDIA vGPU ▶ GPU pass through	▶ SUSE Linux Enterprise Server 15 SP2+ ▶ Ubuntu 22.04 LTS ▶ Ubuntu 20.04 LTS	containerd	▶ Upstream Kubernetes 1.21 through 1.25 ▶ VMware vSphere with Tanzu 7.0 U3c
				Docker	N/A
			▶ Red Hat OpenShift 4.11 and later using Red Hat Linux CoreOS (RHCOS)	CRI-O	▶ Red Hat OpenShift 4.11 and later using RHCOS
			▶ Red Hat Enterprise Linux 9.2, 9.0 ▶ Red Hat Enterprise	Podman	N/A

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA A100 PCIe 80GB ▶ NVIDIA A100 PCIe 80GB liquid cooled ▶ NVIDIA A100 HGX 80GB ▶ NVIDIA A40 ▶ NVIDIA A30X ▶ NVIDIA A30 ▶ NVIDIA A30 liquid cooled ▶ NVIDIA A10 ▶ NVIDIA A16 ▶ NVIDIA A2 ▶ NVIDIA H800 PCIe 94GB (H800 NVL) ▶ NVIDIA H800 PCIe 80GB ▶ NVIDIA H100 PCIe 94GB	Enterprise Plus Edition 7.0 Update 2 ▶ VMware vCenter Server 7.0 Update 2		Linux 8.8, 8.6 ▶ Red Hat Enterprise Linux 7.9		
			Red Hat Enterprise Linux 8.8, 8.6	containerd	HPE Ezmeral Runtime Enterprise 5.5
			▶ Microsoft Windows Server 2022, 2019 ▶ Microsoft Windows 11, 10	N/A	N/A

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
(H100 NVL)					
▶ NVIDIA H100 PCIe 80GB					
▶ NVIDIA L40					
▶ NVIDIA L40S					
▶ NVIDIA L4					
▶ NVIDIA RTX 6000 Ada					
▶ NVIDIA RTX 5000 Ada					
▶ NVIDIA RTX A6000					
▶ NVIDIA RTX A5500					
▶ NVIDIA RTX A5000					
▶ NVIDIA RTX 6000 passive					
▶ NVIDIA RTX 8000 passive					
▶ NVIDIA T4					
▶ NVIDIA V100					

Chapter 2.   NVIDIA RAPIDS Accelerator for Apache Spark Support

NVIDIA RAPIDS Accelerator for Apache Spark is a software component of NVIDIA AI Enterprise. It uses NVIDIA GPUs to accelerate Spark data frame workloads transparently, that is, without code changes.

NVIDIA AI Enterprise supports RAPIDS Accelerator for Apache Spark on the following platforms:

▶ Google Cloud Dataproc

▶ Databricks on the following cloud services:

    ▶ Amazon Web Services (AWS)

    ▶ Microsoft Azure

▶ Amazon EMR (formerly "Amazon Elastic MapReduce")

Chapter 3. NVIDIA AI Enterprise Supported Cloud Services

NVIDIA AI Enterprise is supported on several cloud services with bring-your-own-license (BYOL) licensing and pay-as-you-go licensing.

▶ [Amazon Web Services Elastic Compute Cloud (AWS EC2)](#)

▶ [Google Cloud Platform (GCP)](#)

▶ [Microsoft Azure](#)

▶ [Oracle Cloud Infrastructure](#)

> **Note:** Red Hat Enterprise Linux guest OS support is limited to running containers by using Docker **without** Kubernetes. NVIDIA AI Enterprise features that depend on Kubernetes, for example, the use of GPU Operator, are not supported on Red Hat Enterprise Linux.

3.1. Amazon Web Services Elastic Compute Cloud (AWS EC2)

> **Note:** Pay-as-you-go licensing is also available for all supported AWS EC2 instances.

GPU	Supported AWS EC2 Instances	Supported Guest Operating Systems
NVIDIA T4	g4dn.xlarge g4dn.2xlarge g4dn.4xlarge g4dn.8xlarge g4dn.12xlarge g4dn.16xlarge	Red Hat Enterprise Linux 8.4 Red Hat Enterprise Linux 7.9 Red Hat OpenShift 4.10 using Red Hat Linux CoreOS (RHCOS) Red Hat OpenShift 4.9 using Red Hat Linux CoreOS (RHCOS) Ubuntu 22.04 Ubuntu 20.04
NVIDIA V100	P3.2xlarge P3.8xlarge	

GPU	Supported AWS EC2 Instances	Supported Guest Operating Systems
	P3.16xlarge	
NVIDIA A10G	g5.xlarge g5.2xlarge g5.4xlarge g5.8xlarge g5.12xlarge g5.16xlarge g5.24xlarge g5.48xlarge	
NVIDIA A100	p4d.24xlarge	

3.2. Google Cloud Platform (GCP)

> 💬 **Note:** Pay-as-you-go licensing is also available for all supported GCP instances.

GPU	Supported GCP Instances	Supported Guest Operating Systems
NVIDIA A100	a2-highgpu-1g a2-highgpu-2g a2-highgpu-4g a2-highgpu-8g a2-megagpu-16g	Red Hat Enterprise Linux 8.4 Red Hat Enterprise Linux 7.9 Red Hat OpenShift 4.10 using Red Hat Linux CoreOS (RHCOS) Red Hat OpenShift 4.9 using Red Hat Linux CoreOS (RHCOS) Ubuntu 22.04 Ubuntu 20.04
NVIDIA L4	g2-standard-4 g2-standard-8 g2-standard-12 g2-standard-16 g2-standard-24 g2-standard-32 g2-standard-48 g2-standard-96	
NVIDIA T4	Any predefined machine type.	
NVIDIA V100	Any custom machine type that can be created in a zone.	

3.3.   Microsoft Azure

> **Note:** Pay-as-you-go licensing is also available for all supported Microsoft Azure instances, **except** NV_A10_v5 instances.

GPU	Supported Azure Instances	Supported Guest Operating Systems
NVIDIA V100	NC6s_v3 NC12s_v3 NC24s_v3 NC24rs_v3 ND40rs_v2	Red Hat Enterprise Linux 8.4 Red Hat Enterprise Linux 7.9 Red Hat OpenShift 4.10 using Red Hat Linux CoreOS (RHCOS) Red Hat OpenShift 4.9 using Red Hat Linux CoreOS (RHCOS) Ubuntu 22.04 Ubuntu 20.04
NVIDIA T4	NC4asT4_v3 NC8asT4_v3 NC16asT4_v3 NC64asT4_v3	
NVIDIA A100	NC24ads_A100_v4 NC48ads_A100_v4 NC96ads_A100_v4 ND96asr_v4 ND96amsr_A100_v4	
NVIDIA A10	NV6ads_A10_v5 NV12ads_A10_v5 NV18ads_A10_v5 NV36ads_A10_v5 NV36adms_A10_v5 NV72ads_A10_v5	

3.4.   Oracle Cloud Infrastructure

> **Note:** Pay-as-you-go licensing is also available for all supported Oracle Cloud Infrastructure shapes.

GPU	Oracle Cloud Infrastructure Shapes	Supported Guest Operating Systems
NVIDIA P100	VM.GPU2.1 BM.GPU2.2	Linux: ▶ Ubuntu 22.04 ▶ Ubuntu 20.04 Windows: ▶ Microsoft Windows Server 2022 ▶ Microsoft Windows Server 2019
NVIDIA V100	VM.GPU3.1 VM.GPU3.2 VM.GPU3.4 BM.GPU3.8	
NVIDIA A100	BM.GPU.GM4.8 BM.GPU4.8	
NVIDIA A10	BM.GPU.GU1.4	

3.5.   NVIDIA GPU Optimized VMI on CSP Marketplace

For ease of use in the cloud, NVIDIA provides compute optimized and validated base Virtual Machine Instances (VMI) through CSP marketplaces. Each VMI includes key technologies and software from NVIDIA for rapid deployment, management, and scaling of AI workloads in the modern hybrid cloud.

Each VMI has the following software pre-installed:

▶ Ubuntu Server 20.04

▶ NVIDIA driver 525 TRD - 525.60.13

▶ Docker-ce 20.10.12

▶ NVIDIA Container Toolkit 1.8.1

▶ NVIDIA Container Runtime 3.8.1

Chapter 4. CPU Only Server Support

NVIDIA AI Enterprise supports deployments on CPU only servers that are part of the NVIDIA Certified Systems list. Customers can deploy both GPU and CPU Only systems with VMware vSphere or Red Hat Enterprise Linux.

NVIDIA AI Enterprise will support the following CPU enabled frameworks:

- ▶ TensorFlow
- ▶ PyTorch
- ▶ Triton Inference Server with FIL backend
- ▶ NVIDIA RAPIDS with XGBoost and Dask

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.nvidia.com