NVIDIA AI Enterprise

Product Support Matrix

Table of Contents

Chapter 1.   Product Support Matrix

Driver package: NVIDIA AI Enterprise 5.1

Validated partner integrations:

- ▶ Run: AI: 2.5
- ▶ Domino Data Lab

Details of NVIDIA AI Enterprise support on various hypervisors and bare-metal operating systems are provided in the following sections:

- ▶ [Amazon Web Services (AWS) Nitro Support](#)
- ▶ [Azure Kubernetes Service (AKS) Support](#)
- ▶ [Google Cloud Platform (GCP) KVM Support](#)
- ▶ [Linux with KVM Support](#)
- ▶ [Red Hat Enterprise Linux Support](#)
- ▶ [Red Hat OpenShift Support](#)
- ▶ [SUSE Linux Enterprise Server Support](#)
- ▶ [Ubuntu Support](#)
- ▶ [VMware vSphere Support](#)

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 H100 ▶ NVIDIA T4	Amazon Web Services (AWS) Nitro		▶ Ubuntu 22.04 LTS ▶ Ubuntu 20.04 LTS	containerd	Amazon Elastic Kubernetes Service (EKS)

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
► NVIDIA V100					

Azure Kubernetes Service (AKS) Support

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
► NVIDIA A100 ► NVIDIA A10 ► NVIDIA H100 ► NVIDIA T4 ► NVIDIA V100	Azure Hypervisor		► Ubuntu 22.04 LTS	containerd	Azure Kubernetes Service (AKS)

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 H100 ► NVIDIA L4 ► 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 My Nutanix 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	OpenStack Platform	Product Documentation for Red Hat OpenStack Platform

Red Hat Enterprise Linux Support

Note: The NVIDIA GPUs listed in the table support NVIDIA AI Enterprise **only** with NVIDIA AI Enterprise compatible servers.

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	▶ Red Hat Enterprise Linux with KVM 9.4, 9.2 ▶ Red Hat Enterprise Linux with KVM 8.10, 8.8	▶ NVIDIA vGPU ▶ GPU pass through	▶ Red Hat OpenShift 4.12 through 4.15 using RHCOS	CRI-O	▶ Red Hat OpenShift 4.12 through 4.15 using RHCOS ▶ Upstream Kubernetes 1.22 through 1.29
			▶ Red Hat Enterprise	Podman	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			Linux 9.4, 9.2 ▶ Red Hat Enterprise Linux 8.10, 8.8 ▶ Red Hat Enterprise Linux 7.9		
			▶ Microsoft Windows Server 2022	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 A16					
► NVIDIA A2					
► NVIDIA H800 PCIe 94GB (H800 NVL)					
► NVIDIA H800 PCIe 80GB					
► NVIDIA H800 SXM5 80GB					
► NVIDIA H100 PCIe 94GB (H100 NVL)					
► NVIDIA H100 PCIe 80GB					
► NVIDIA H100 SXM5 94GB					
► NVIDIA H100 SXM5 80GB					
► NVIDIA H100 SXM5 64GB					
► NVIDIA L40					

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA L40S					
▶ NVIDIA L20					
▶ NVIDIA L20 liquid cooled					
▶ NVIDIA L4					
▶ NVIDIA L2					
▶ NVIDIA RTX 6000 Ada					
▶ NVIDIA RTX 5880 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					

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 ▶ NVIDIA A100 PCIe 80GB ▶ NVIDIA A100 PCIe 80GB liquid cooled ▶ NVIDIA A100 HGX 80GB	▶ Red Hat Enterprise Linux 9.4, 9.2 ▶ Red Hat Enterprise Linux 8.10, 8.8 ▶ Red Hat Enterprise Linux 7.9	▶ Bare metal	N/A	Podman	N/A
	Red Hat Enterprise Linux 8.10, 8.8	▶ Bare metal	N/A	containerd	HPE Ezmeral Runtime Enterprise 5.5
				CRI-O	Upstream Kubernetes 1.22 through 1.29

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA A40					
▶ NVIDIA A30X					
▶ NVIDIA A30					
▶ NVIDIA A30 liquid cooled					
▶ NVIDIA A10					
▶ NVIDIA A16					
▶ NVIDIA A2					
▶ NVIDIA GH200 96GB (CG1) Grace Hopper™ Superchip2					
▶ NVIDIA GH200 Grace Hopper 144GB (CG1) Superchip2					
▶ NVIDIA H800 PCIe 94GB (H800 NVL)					
▶ NVIDIA H800 PCIe 80GB					

---

2 All variants of the NVIDIA GH200 Grace Hopper Superchip are supported only in bare-metal deployments on Red Hat Enterprise Linux, SUSE Linux Enterprise Server, and Ubuntu.

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA H800 SXM5 80GB					
▶ NVIDIA H100 PCIe 94GB (H100 NVL)					
▶ NVIDIA H100 PCIe 80GB					
▶ NVIDIA H100 SXM5 94GB					
▶ NVIDIA H100 SXM5 80GB					
▶ NVIDIA H100 SXM5 64GB					
▶ NVIDIA L40					
▶ NVIDIA L40S					
▶ NVIDIA L20					
▶ NVIDIA L20 liquid cooled					
▶ NVIDIA L4					
▶ NVIDIA L2					
▶ NVIDIA RTX 6000 Ada					

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA RTX 5880 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					

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.

💬 **Note:** The NVIDIA GPUs listed in the table support NVIDIA AI Enterprise **only** with NVIDIA AI Enterprise compatible servers.

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 ▶ NVIDIA A100 PCIe 80GB ▶ NVIDIA A100 PCIe 80GB liquid cooled ▶ NVIDIA A100 HGX 80GB	▶ Red Hat OpenShift 4.12 through 4.15 using RHCOS	▶ Bare metal	N/A	CRI-O	▶ Red Hat OpenShift 4.12 through 4.15 using RHCOS

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ 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 H800 SXM5 80GB					
▶ NVIDIA H100 PCIe 94GB (H100 NVL)					
▶ NVIDIA H100 PCIe 80GB					
▶ NVIDIA H100 SXM5 94GB					

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA H100 SXM5 80GB					
▶ NVIDIA H100 SXM5 64GB					
▶ NVIDIA L40					
▶ NVIDIA L40S					
▶ NVIDIA L20					
▶ NVIDIA L20 liquid cooled					
▶ NVIDIA L4					
▶ NVIDIA L2					
▶ NVIDIA RTX 6000 Ada					
▶ NVIDIA RTX 5880 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 ▶ 5000 ▶ NVIDIA T4 ▶ NVIDIA V100					

SUSE Linux Enterprise Server Support

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

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<u>1</u>	▶ SUSE Linux Enterprise Server 15 SP2+	▶ Bare metal	N/A	Docker	N/A

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ 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 GH200 96GB (CG1)					

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
Grace Hopper Superchip[2]					
▶ NVIDIA GH200 144GB (CG1) Grace Hopper Superchip[2]					
▶ NVIDIA H800 PCIe 94GB (H800 NVL)					
▶ NVIDIA H800 PCIe 80GB					
▶ NVIDIA H800 SXM5 80GB					
▶ NVIDIA H100 PCIe 94GB (H100 NVL)					
▶ NVIDIA H100 PCIe 80GB					
▶ NVIDIA H100 SXM5 94GB					
▶ NVIDIA H100 SXM5 80GB					

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA H100 SXM5 64GB					
▶ NVIDIA L40					
▶ NVIDIA L40S					
▶ NVIDIA L20					
▶ NVIDIA L20 liquid cooled					
▶ NVIDIA L4					
▶ NVIDIA L2					
▶ NVIDIA RTX 6000 Ada					
▶ NVIDIA RTX 5880 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					

Ubuntu Support

> 💬 **Note:** The NVIDIA GPUs listed in the table support NVIDIA AI Enterprise **only** with NVIDIA AI Enterprise compatible servers.

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	▶ Ubuntu 22.04 LTS ▶ Ubuntu 20.04 LTS	▶ NVIDIA vGPU ▶ GPU pass through	▶ Ubuntu 22.04 LTS ▶ Ubuntu 20.04 LTS	containerd	▶ Charmed Kubernetes 1.28 ▶ Upstream Kubernetes 1.22 through 1.29 ▶ VMware vSphere with Tanzu 8.0 ▶ 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 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 94GB (H800 NVL)					

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA H800 PCIe 80GB					
▶ NVIDIA H800 SXM5 80GB					
▶ NVIDIA H100 PCIe 94GB (H100 NVL)					
▶ NVIDIA H100 PCIe 80GB					
▶ NVIDIA H100 SXM5 94GB					
▶ NVIDIA H100 SXM5 80GB					
▶ NVIDIA H100 SXM5 64GB					
▶ NVIDIA L40					
▶ NVIDIA L40S					
▶ NVIDIA L20					
▶ NVIDIA L20 liquid cooled					
▶ NVIDIA L4					
▶ NVIDIA L2					

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA RTX 6000 Ada ▶ NVIDIA RTX 5880 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					
▶ NVIDIA A800 PCIe 80GB ▶ NVIDIA A800 PCIe 80GB liquid cooled	▶ Ubuntu 22.04 LTS ▶ Ubuntu 20.04 LTS	▶ Bare metal	N/A	containerd	▶ Charmed Kubernetes 1.28 ▶ Nutanix DKP 1.27 ▶ Upstream Kubernetes 1.22 through 1.29

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ 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 ▶ NVIDIA A30 liquid cooled					▶ VMware vSphere with Tanzu 8.0 ▶ VMware vSphere with Tanzu 7.0 U3c
				CRI-O	Upstream Kubernetes 1.22 through 1.29
				Docker	N/A

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA A10					
▶ NVIDIA A16					
▶ NVIDIA A2					
▶ NVIDIA GH200 Grace Hopper Superchip[2]					
▶ NVIDIA H800 PCIe 94GB (H800 NVL)					
▶ NVIDIA H800 PCIe 80GB					
▶ NVIDIA H800 SXM5 80GB					
▶ NVIDIA H100 PCIe 94GB (H100 NVL)					
▶ NVIDIA H100 PCIe 80GB					
▶ NVIDIA H100 SXM5 94GB					
▶ NVIDIA H100 SXM5 80GB					

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA H100 SXM5 64GB					
▶ NVIDIA L40					
▶ NVIDIA L40S					
▶ NVIDIA L20					
▶ NVIDIA L20 liquid cooled					
▶ NVIDIA L4					
▶ NVIDIA L2					
▶ NVIDIA RTX 6000 Ada					
▶ NVIDIA RTX 5880 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

💬 **Note:** The NVIDIA GPUs listed in the table support NVIDIA AI Enterprise **only** with [NVIDIA AI Enterprise compatible](#) servers.

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	▶ 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	▶ NVIDIA vGPU ▶ GPU pass through	▶ Ubuntu 22.04 LTS ▶ Ubuntu 20.04 LTS	containerd	▶ Upstream Kubernetes 1.22 through 1.29 ▶ VMware vSphere with Tanzu 8.0 ▶ VMware vSphere with Tanzu 7.0 U3c
			▶ SUSE Linux Enterprise Server 15 SP2+ ▶ Ubuntu 22.04 LTS	Docker	N/A

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ 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 94GB (H800 NVL)	▶ VMware vSphere Hypervisor (ESXi) Enterprise Plus Edition 7.0 Update 2 ▶ VMware vCenter Server 7.0 Update 2		▶ Ubuntu 20.04 LTS		
			▶ Red Hat OpenShift 4.12 through 4.15 using Red Hat Linux CoreOS (RHCOS)	CRI-O	▶ Red Hat OpenShift 4.12 through 4.15 using RHCOS
			▶ Red Hat Enterprise Linux 9.4, 9.2 ▶ Red Hat Enterprise Linux 8.10, 8.8 ▶ Red Hat Enterprise Linux 7.9	Podman	N/A
			Red Hat Enterprise Linux 8.10, 8.8	containerd	HPE Ezmeral Runtime Enterprise 5.5
				▶ containerd ▶ CRI-O	Upstream Kubernetes 1.22 through 1.29
			▶ Microsoft Windows Server 2022	N/A	N/A

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA H800 PCIe 80GB ▶ NVIDIA H800 SXM5 80GB[3] ▶ NVIDIA H100 PCIe 94GB (H100 NVL) ▶ NVIDIA H100 PCIe 80GB ▶ NVIDIA H100 SXM5 94GB ▶ NVIDIA H100 SXM5 80GB[3] ▶ NVIDIA H100 SXM5 64GB ▶ NVIDIA L40 ▶ NVIDIA L40S ▶ NVIDIA L20 ▶ NVIDIA L20 liquid cooled			▶ Microsoft Windows 11, 10		

---

[3] When deployed on an NVIDIA HGX Hopper 8-GPU baseboard, this GPU is supported starting with VMware vSphere 8 update 2. Earlier VMware vSphere releases are **not** supported.

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA L4					
▶ NVIDIA L2					
▶ NVIDIA RTX 6000 Ada					
▶ NVIDIA RTX 5880 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. Pay-as-you-go licensing is also available with some cloud services.

▶ Alibaba

▶ Amazon Web Services Elastic Compute Cloud (AWS EC2)

▶ Google Cloud Platform (GCP)

▶ Microsoft Azure

▶ Oracle Cloud Infrastructure

▶ Tencent Cloud

▶ Volcano Engine

3.1. Alibaba

GPU	Supported Alibaba Instances	Certified Container Orchestration Platforms	Supported Guest Operating Systems
NVIDIA V100	gn6e gn6v	Upstream Kubernetes	▶ Ubuntu 22.04 ▶ Ubuntu 20.04
NVIDIA A10	gn7e gn7i		

3.2. 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	Certified Container Orchestration Platforms	Supported Guest Operating Systems
NVIDIA T4	All G4 series instances	Amazon Elastic Kubernetes Service (EKS) Red Hat OpenShift Upstream Kubernetes	Red Hat Enterprise Linux 8.9, 8.8, 8.6 Red Hat Enterprise Linux 7.9 Red Hat OpenShift 4.12 through 4.15 using Red Hat Linux CoreOS (RHCOS) Ubuntu 22.04 Ubuntu 20.04
NVIDIA V100	All P3 series instances		
NVIDIA A10G	All G5 series instances		
NVIDIA A100	All P4d and P4de series instances		
NVIDIA H100	All P5 series instances		

3.3.   Google Cloud Platform (GCP)

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

GPU	Supported GCP Instances	Certified Container Orchestration Platforms	Supported Guest Operating Systems
NVIDIA A100	All A2 series instances	Google Kubernetes Engine (GKE) Red Hat OpenShift Upstream Kubernetes	Red Hat Enterprise Linux 8.9, 8.8, 8.6 Red Hat Enterprise Linux 7.9 Red Hat OpenShift 4.12 through 15 using Red Hat Linux CoreOS (RHCOS) Ubuntu 22.04 Ubuntu 20.04
NVIDIA H100	All A3 series instances		
NVIDIA L4	All G2 series instances		
NVIDIA T4	Any predefined machine type.		
NVIDIA V100	Any custom machine type that can be created in a zone.		

3.4.   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	Certified Container Orchestration Platforms	Supported Guest Operating Systems
NVIDIA V100	All NC v3 and ND v2 instances	Azure Kubernetes Service (AKS) Red Hat OpenShift Upstream Kubernetes	Red Hat Enterprise Linux 8.4
NVIDIA T4	All NC T4_v3 instances		Red Hat Enterprise Linux 7.9
NVIDIA H100	All ND H100_v5instances		Red Hat OpenShift 4.10 using Red Hat Linux CoreOS (RHCOS)
NVIDIA A100	All NC A100_v4 instances All ND A100_v4 instances		Red Hat OpenShift 4.9 using Red Hat Linux CoreOS (RHCOS)
NVIDIA A10	All NV A10_v5 instances		Ubuntu 22.04 Ubuntu 20.04

3.5.   Oracle Cloud Infrastructure

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

GPU	Oracle Cloud Infrastructure Shapes	Certified Container Orchestration Platforms	Supported Guest Operating Systems
NVIDIA V100	All VM.GPU3 shapes	Upstream Kubernetes	Linux: ▶ Ubuntu 22.04 ▶ Ubuntu 20.04 Windows: ▶ Microsoft Windows Server 2022
NVIDIA H100	BM.GPU.H100.8		
NVIDIA A100	All BM.GPU4 shapes All BM.GPU.A100-v2 shapes		
NVIDIA A10	All VM.GPU.A10 shapes		

3.6.   Tencent Cloud

GPU	Supported Tencent Cloud Instances	Certified Container Orchestration Platforms	Supported Guest Operating Systems
NVIDIA V100	GN10Xp	Upstream Kubernetes	▶ Ubuntu 22.04
NVIDIA A10	PNV4		

GPU	Supported Tencent Cloud Instances	Certified Container Orchestration Platforms	Supported Guest Operating Systems
			▶ Ubuntu 20.04

3.7. Volcano Engine

GPU	Volcano Engine Instances	Certified Container Orchestration Platforms	Supported Guest Operating Systems
NVIDIA A10	ecs.gni2	Upstream Kubernetes	▶ Ubuntu 22.04 ▶ Ubuntu 20.04

3.8. 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