



NVIDIA AI Enterprise

Product Support Matrix

Table of Contents

Chapter 1. Product Support Matrix.....	1
Chapter 2. NVIDIA RAPIDS Accelerator for Apache Spark Support.....	26
Chapter 3. NVIDIA AI Enterprise Supported Cloud Services.....	27
3.1. Alibaba.....	27
3.2. Amazon Web Services Elastic Compute Cloud (AWS EC2).....	28
3.3. Google Cloud Platform (GCP).....	28
3.4. Microsoft Azure.....	29
3.5. Oracle Cloud Infrastructure.....	29
3.6. Tencent Cloud.....	30
3.7. NVIDIA GPU Optimized VMI on CSP Marketplace.....	30
Chapter 4. CPU Only Server Support.....	31

Chapter 1. Product Support Matrix

Driver package: NVIDIA AI Enterprise 4.2

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 with KVM 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 T4 ▶ NVIDIA V100	Amazon Web Services (AWS) Nitro		▶ Ubuntu 22.04 LTS ▶ Ubuntu 20.04 LTS	containerd	Amazon Elastic Kubernetes Service (EKS)

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 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 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.

Hypervisor Vendor	Platform	Additional Information
		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



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 AX800¹ ▶ NVIDIA A100X ▶ NVIDIA A100 PCIe 40GB	▶ Red Hat Enterprise Linux with KVM 9.2, 9.0 ▶ Red Hat Enterprise Linux with KVM 8.9, 8.8, 8.6	▶ NVIDIA vGPU ▶ GPU pass through	▶ Red Hat OpenShift 4.11 and later using RHCOS	CRI-O	▶ Red Hat OpenShift 4.11 and later using RHCOS ▶ Upstream Kubernetes 1.21 through 1.25
			▶ Red Hat Enterprise Linux 9.2, 9.0 ▶ Red Hat Enterprise Linux 8.9, 8.8, 8.6 ▶ Red Hat Enterprise Linux 7.9	Podman	N/A
			▶ Microsoft Windows Server	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 A100 HGX 40GB			2022, 2019		
▶ 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					

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

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ 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 8.9, 8.8, 8.6 ▶ Red Hat Enterprise Linux 7.9	▶ Bare metal	N/A	Podman	N/A
▶ NVIDIA A800 40GB PCIe active cooled ▶ NVIDIA AX800¹ ▶ NVIDIA A100X ▶ NVIDIA A100 PCIe 40GB ▶ NVIDIA A100 HGX 40GB ▶ NVIDIA A100 PCIe 80GB ▶ NVIDIA A100 PCIe 80GB liquid cooled	Red Hat Enterprise Linux 8.9, 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 A100 HGX 80GB					
▶ NVIDIA A40					
▶ NVIDIA A30X					
▶ NVIDIA A30					
▶ NVIDIA A30 liquid cooled					
▶ NVIDIA A10					
▶ NVIDIA A16					
▶ NVIDIA A2					
▶ NVIDIA GH200 Grace Hopper™ Superchip					
▶ NVIDIA H800 PCIe 94GB (H800 NVL)					
▶ NVIDIA H800 PCIe 80GB					
▶ NVIDIA H100 PCIe 94GB (H100 NVL)					

² The NVIDIA GH200 Grace Hopper Superchip is 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 H100 PCIe 80GB					
▶ NVIDIA H100 SXM5 94GB ³					
▶ NVIDIA H100 SXM5 80GB ³					
▶ NVIDIA H100 SXM5 64GB ³					
▶ NVIDIA L40					
▶ NVIDIA L40S					
▶ NVIDIA L20					
▶ NVIDIA L4					
▶ NVIDIA L2					
▶ NVIDIA RTX 6000 Ada					
▶ NVIDIA RTX 5000 Ada					
▶ NVIDIA RTX A6000					
▶ NVIDIA RTX A5500					

³ NVIDIA H100 SXM5 GPUs are supported only in bare-metal deployments on Red Hat Enterprise Linux, SUSE Linux Enterprise Server, 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 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	▶ 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
▶ 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					
▶ NVIDIA A10					

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 H100 PCIe 94GB (H100 NVL)					
▶ NVIDIA H100 PCIe 80GB					
▶ NVIDIA H100 SXM5 94GB ³					
▶ NVIDIA H100 SXM5 80GB ³					
▶ NVIDIA H100 SXM5 64GB ³					
▶ NVIDIA L40					
▶ NVIDIA L40S					
▶ NVIDIA L20					
▶ NVIDIA L4					

GPU	Hypervisor or Bare- Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA L2 ▶ 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



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 AX800¹ ▶ NVIDIA A100X ▶ NVIDIA A100 PCIe 40GB ▶ NVIDIA A100 HGX 40GB ▶ NVIDIA A100 PCIe 80GB ▶ NVIDIA A100 PCIe 80GB liquid cooled ▶ NVIDIA A100 HGX 80GB	▶ SUSE Linux Enterprise Server 15 SP2+	▶ Bare metal	N/A	containerd Docker	▶ Upstream Kubernetes 1.21 through 1.25 N/A

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 Grace Hopper Superchip² ▶ NVIDIA H800 PCIe 94GB (H800 NVL) ▶ NVIDIA H800 PCIe 80GB ▶ NVIDIA H100 PCIe 94GB (H100 NVL) ▶ NVIDIA H100 PCIe 80GB ▶ NVIDIA H100					

GPU	Hypervisor or Bare- Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
SXM5 94GB3					
► NVIDIA H100 SXM5 80GB3					
► NVIDIA H100 SXM5 64GB3					
► NVIDIA L40					
► NVIDIA L40S					
► NVIDIA L20					
► NVIDIA L4					
► NVIDIA L2					
► 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					

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 AX800¹ ▶ NVIDIA A100X	▶ Ubuntu 22.04 LTS ▶ Ubuntu 20.04 LTS	▶ NVIDIA vGPU ▶ GPU pass through	▶ Ubuntu 22.04 LTS ▶ Ubuntu 20.04 LTS	containerd Docker	▶ Upstream Kubernetes 1.21 through 1.25 ▶ VMware vSphere with Tanzu 7.0 U3c 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 H100 PCIe 94GB (H100 NVL)					
▶ NVIDIA H100 PCIe 80GB					
▶ NVIDIA L40					
▶ NVIDIA L40S					
▶ NVIDIA L20					
▶ NVIDIA L4					
▶ NVIDIA L2					
▶ 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					
▶ NVIDIA A800 PCIe 80GB ▶ NVIDIA A800 PCIe 80GB liquid cooled ▶ NVIDIA A800 HGX 80GB ▶ NVIDIA A800 40GB PCIe active cooled ▶ NVIDIA AX800¹ ▶ NVIDIA A100X ▶ NVIDIA A100 PCIe 40GB ▶ NVIDIA A100 HGX 40GB ▶ NVIDIA A100 PCIe 80GB ▶ NVIDIA A100 PCIe	▶ Ubuntu 22.04 LTS ▶ Ubuntu 20.04 LTS	▶ Bare metal	N/A	containerd Docker	▶ Upstream Kubernetes 1.21 through 1.25 ▶ VMware vSphere with Tanzu 7.0 U3c N/A

GPU	Hypervisor or Bare- Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
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 Grace Hopper Superchip ²					
► 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 94GB ³					
▶ NVIDIA H100 SXM5 80GB ³					
▶ NVIDIA H100 SXM5 64GB ³					
▶ NVIDIA L40					
▶ NVIDIA L40S					
▶ NVIDIA L20					
▶ NVIDIA L4					
▶ NVIDIA L2					
▶ NVIDIA RTX 6000 Ada					
▶ NVIDIA RTX 5000 Ada					
▶ NVIDIA RTX A6000					
▶ NVIDIA RTX A5500					

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA RTX A5000 ▶ NVIDIA RTX 6000 passive ▶ 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	▶ VMware vSphere Hypervisor (ESXi) Enterprise Plus Edition 8.0 ▶ VMware vCenter Server 8.0 ▶ VMware vSphere Hypervisor (ESXi) Enterprise Plus	▶ NVIDIA vGPU ▶ GPU pass through	▶ SUSE Linux Enterprise Server 15 SP2+ ▶ Ubuntu 22.04 LTS ▶ Ubuntu 20.04 LTS ▶ Red Hat OpenShift 4.11 and later using Red Hat	containerd Docker CRI-O	▶ Upstream Kubernetes 1.21 through 1.25 ▶ VMware vSphere with Tanzu 7.0 U3c - N/A ▶ 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 AX800 ¹	Edition 7.0 Update 3		Linux CoreOS (RHCOS)		
▶ NVIDIA A100X	VMware vCenter Server 7.0 Update 3		▶ Red Hat Enterprise Linux 9.2, 9.0	Podman	N/A
▶ NVIDIA A100 PCIe 40GB	▶ VMware vSphere Hypervisor (ESXi) Enterprise Plus Edition 7.0 Update 2		▶ Red Hat Enterprise Linux 8.9, 8.8, 8.6		
▶ NVIDIA A100 HGX 40GB			▶ Red Hat Enterprise Linux 7.9		
▶ NVIDIA A100 PCIe 80GB	▶ VMware vCenter Server 7.0 Update 2		Red Hat Enterprise Linux 8.9, 8.8, 8.6	containerd	HPE Ezmeral Runtime Enterprise 5.5
▶ NVIDIA A100 PCIe 80GB liquid cooled			▶ Microsoft Windows Server 2022, 2019	N/A	N/A
▶ NVIDIA A100 HGX 80GB			▶ Microsoft Windows 11, 10		
▶ 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 L40					
► NVIDIA L40S					
► NVIDIA L20					
► NVIDIA L4					
► NVIDIA L2					
► 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					

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](#)



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. 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 8.4
NVIDIA V100	All P3 series instances		Red Hat Enterprise Linux 7.9
NVIDIA A10G	All G5 series instances		Red Hat OpenShift 4.10 using Red Hat Linux CoreOS (RHCOS)
NVIDIA A100	All P4d and P4de series instances		Red Hat OpenShift 4.9 using Red Hat Linux CoreOS (RHCOS)
NVIDIA H100	All P5 series instances		Ubuntu 22.04 Ubuntu 20.04

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 8.4
NVIDIA H100	All A3 series instances		Red Hat Enterprise Linux 7.9
NVIDIA L4	All G2 series instances		Red Hat OpenShift 4.10 using Red Hat Linux CoreOS (RHCOS)
NVIDIA T4	Any predefined machine type .		Red Hat OpenShift 4.9 using Red Hat Linux CoreOS (RHCOS)
NVIDIA V100	Any custom machine type that can be created in a zone.		Ubuntu 22.04

GPU	Supported GCP Instances	Certified Container Orchestration Platforms	Supported Guest Operating Systems
			Ubuntu 20.04

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_v5 instances		Red Hat OpenShift 4.10 using Red Hat Linux CoreOS (RHCOS)
NVIDIA A100	All NC A100_v4 instances		Red Hat OpenShift 4.9 using Red Hat Linux CoreOS (RHCOS)
	All ND A100_v4 instances		Ubuntu 22.04
NVIDIA A10	All NV A10_v5 instances		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:
NVIDIA H100	BM.GPU.H100.8		
NVIDIA A100	All BM.GPU4 shapes All BM.GPU.A100-v2 shapes		

GPU	Oracle Cloud Infrastructure Shapes	Certified Container Orchestration Platforms	Supported Guest Operating Systems
NVIDIA A10	All VM.GPU.A10 shapes		▶ Microsoft Windows Server 2022 ▶ Microsoft Windows Server 2019

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		▶ Ubuntu 20.04

3.7. 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

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation ("NVIDIA") makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer ("Terms of Sale"). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer's own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer's product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA's aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

VESA DisplayPort

DisplayPort and DisplayPort Compliance Logo, DisplayPort Compliance Logo for Dual-mode Sources, and DisplayPort Compliance Logo for Active Cables are trademarks owned by the Video Electronics Standards Association in the United States and other countries.

HDMI

HDMI, the HDMI logo, and High-Definition Multimedia Interface are trademarks or registered trademarks of HDMI Licensing LLC.

OpenCL

OpenCL is a trademark of Apple Inc. used under license to the Khronos Group Inc.

Trademarks

NVIDIA, the NVIDIA logo, NVIDIA Maxwell, NVIDIA Pascal, NVIDIA Turing, NVIDIA Volta, Quadro, and Tesla are trademarks or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

© 2024 NVIDIA Corporation & affiliates. All rights reserved.

