NVIDIA AI Enterprise

Product Support Matrix

PSM-10616-001 _v3.3 | November 2023

Table of Contents

Chapter 1. Product Support Matrix

Driver package: NVIDIA AI Enterprise 3.3

Validated partner integrations:

- Run: AI: 2.5
- Domino Data Lab

Note:

- The GPUs listed in the table are supported only with NVIDIA AI Enterprise compatible servers.
- 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 liquid cooled ▸ NVIDIA A800 HGX 80GB ▸ 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	▸ 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 4.9 and later using Red Hat Linux CoreOS (RHCOS)	CRI-O	▸ Red Hat OpenShift 4.9 and later using RHCOS

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 A10 ▶ NVIDIA A16 ▶ NVIDIA A2 ▶ NVIDIA H800 PCIe 80GB ▶ NVIDIA H100 PCIe 80GB ▶ NVIDIA L40 ▶ NVIDIA L4 ▶ NVIDIA RTX 6000 Ada ▶ NVIDIA RTX A6000 ▶ NVIDIA RTX A5500 ▶ NVIDIA RTX A5000	Server 7.0 Update 3 ▶ VMware vSphere Hypervisor (ESXi) Enterprise Plus Edition 7.0 Update 2 ▶ VMware vCenter Server 7.0 Update 2		▶ 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 ▶ Microsoft Windows 11, 10	N/A	N/A

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 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	▶ Red Hat Enterprise Linux with KVM 9.2, 9.0 ▶ Red Hat Enterprise Linux with KVM 8.8, 8.6	▶ NVIDIA vGPU ▶ GPU pass through	▶ Red Hat OpenShift 4.9 and later using RHCOS	CRI-O	▶ Red Hat OpenShift 4.9 and later using RHCOS
			▶ 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

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA A10 ▶ NVIDIA A16 ▶ NVIDIA A2 ▶ NVIDIA H800 PCIe 80GB ▶ NVIDIA H100 PCIe 80GB ▶ NVIDIA L40 ▶ NVIDIA L4 ▶ NVIDIA RTX 6000 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	▶ Upstream Kubernetes 1.21 through 1.25 ▶ VMware vSphere with Tanzu 7.0 U3c

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA A800 HGX 80GB ▶ 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 A10 ▶ NVIDIA A16 ▶ NVIDIA A2 ▶ NVIDIA H800 PCIe 80GB ▶ NVIDIA H100 PCIe 80GB ▶ NVIDIA L40 ▶ NVIDIA L4 ▶ NVIDIA RTX 6000 Ada				Docker	N/A

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					
▶ NVIDIA A800 PCIe 80GB ▶ NVIDIA A800 PCIe 80GB liquid cooled ▶ NVIDIA A800 HGX 80GB ▶ NVIDIA A100X ▶ NVIDIA A100 PCIe 40GB ▶ NVIDIA A100 HGX 40GB ▶ NVIDIA A100 PCIe 80GB ▶ NVIDIA A100 PCIe 80GB liquid cooled	▶ SUSE Linux Enterprise Server 15 SP2+	▶ Bare metal	N/A	containerd	▶ Upstream Kubernetes 1.21 through 1.25
				Docker	N/A

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 A10 ▶ NVIDIA A16 ▶ NVIDIA A2 ▶ NVIDIA H800 PCIe 80GB ▶ NVIDIA H100 PCIe 80GB ▶ NVIDIA L40 ▶ NVIDIA L4 ▶ NVIDIA RTX 6000 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 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 A10 ▶ NVIDIA A16 ▶ NVIDIA A2 ▶ NVIDIA H800 PCIe 80GB	▶ Red Hat OpenShift 4.9 and later using RHCOS	▶ Bare metal	N/A	CRI-O	▶ Red Hat OpenShift 4.9 and later using RHCOS

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ NVIDIA H100 PCIe 80GB ▶ NVIDIA L40 ▶ NVIDIA L4 ▶ NVIDIA RTX 6000 Ada ▶ NVIDIA RTX A6000 ▶ NVIDIA RTX A5500 ▶ NVIDIA RTX A5000 ▶ NVIDIA RTX 6000 passive ▶ NVIDIA RTX 8000 passive ▶ 5000 ▶ NVIDIA T4 ▶ NVIDIA V100					
▶ NVIDIA A800 PCIe 80GB ▶ NVIDIA A800 PCIe 80GB liquid cooled ▶ NVIDIA A800 HGX 80GB ▶ NVIDIA A100X ▶ NVIDIA A100 PCIe 40GB	▶ 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

GPU	Hypervisor or Bare-Metal OS	GPU Deployment	Guest OS Support	Container Engine	Container Orchestration Platform
▶ 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 A10 ▶ NVIDIA A16 ▶ NVIDIA A2 ▶ NVIDIA H800 PCIe 80GB ▶ NVIDIA H100 PCIe 80GB ▶ NVIDIA L40 ▶ NVIDIA L4 ▶ NVIDIA RTX 6000 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					
▶ 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)
▶ NVIDIA A100 ▶ NVIDIA T4 ▶ NVIDIA V100	Google Cloud Platform (GCP) KVM		▶ Ubuntu 22.04 LTS ▶ Ubuntu 20.04 LTS	containerd	Google Kubernetes Engine (GKE)

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.

- 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)

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 P3.16xlarge	
NVIDIA A10G	g5.xlarge	

GPU	Supported AWS EC2 Instances	Supported Guest Operating Systems
	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

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

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:
NVIDIA V100	VM.GPU3.1 VM.GPU3.2 VM.GPU3.4	

GPU	Oracle Cloud Infrastructure Shapes	Supported Guest Operating Systems
	BM.GPU3.8	▶ Microsoft Windows Server 2022 ▶ Microsoft Windows Server 2019
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

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

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