Deploying NVIDIA AI Enterprise on Oracle Compute Cloud@Customer

Step-by-Step to Deploying and Running Models with NVIDIA AI Enterprise on Oracle Compute Cloud@Customer with NVIDIA L40S

Version [1.0]
Copyright © 2025, Oracle and/or its affiliates
Public

Purpose statement

This document provides insights into deploying NVIDIA AI Enterprise on Oracle Compute Cloud@Customer for running AI/ML workloads and serves as a technical resource for understanding system pre-requisites, installation, and configuration. This document is a step-by-step guide for deploying NVIDIA AI Enterprise on Oracle Compute Cloud@Customer for both Ubuntu and Oracle Linux 8 operating systems.

Disclaimer

This document in any form, software or printed matter, contains proprietary information that is the exclusive property of Oracle. Your access to and use of this confidential material is subject to the terms and conditions of your Oracle software license and service agreement, which has been executed and with which you agree to comply. This document and information contained herein may not be disclosed, copied, reproduced or distributed to anyone outside Oracle without prior written consent of Oracle. This document is not part of your license agreement, nor can it be incorporated into any contractual agreement with Oracle or its subsidiaries or affiliates.

This document is for informational purposes only and is intended solely to assist you in planning for the implementation and upgrade of the product features described. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, timing, and pricing of any features or functionality described in this document remains at the sole discretion of Oracle. Due to the nature of the product architecture, it may not be possible to safely include all features described in this document without risking significant destabilization of the code.

Copyright © 2025, Oracle and/or its affiliates / Public

Table of contents

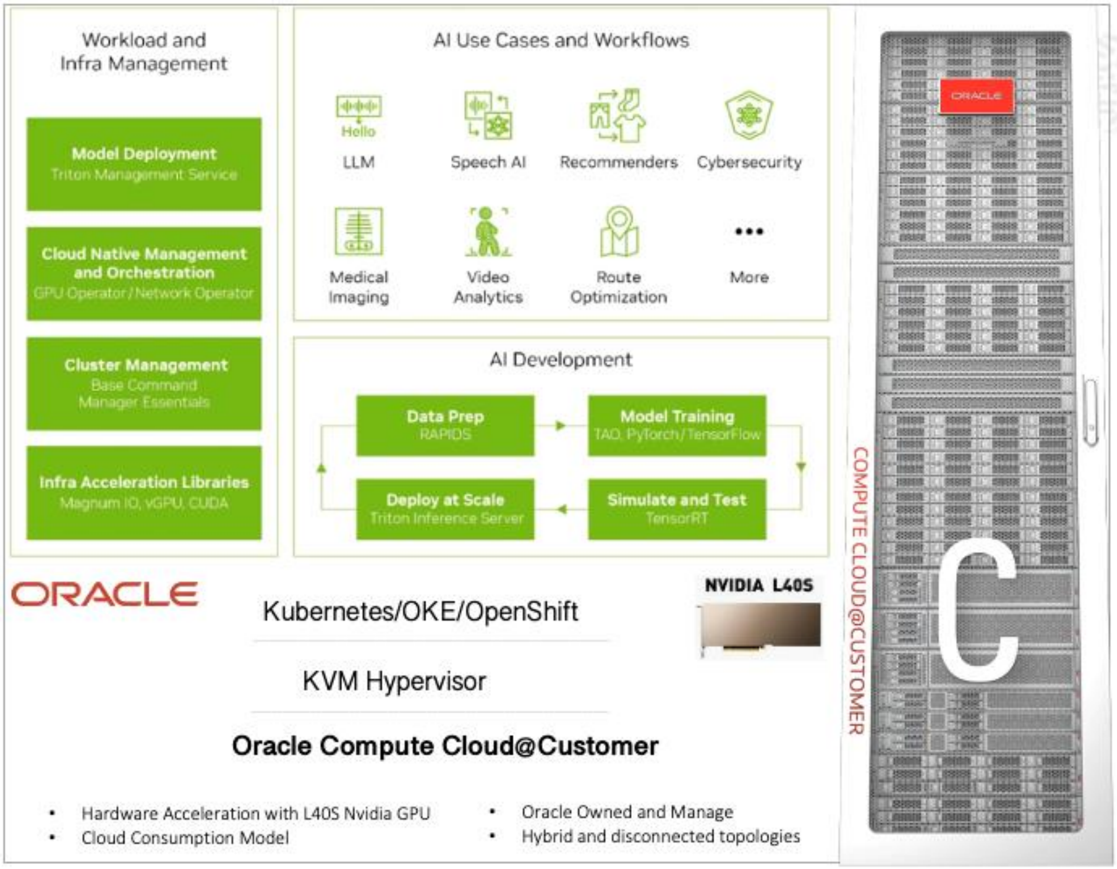
List of figures

Copyright © 2025, Oracle and/or its affiliates / Public

Introduction

Oracle Compute Cloud@Customer enables organizations to run cloud infrastructure on-premises with full control and compliance. This document explains how to deploy NVIDIA AI Enterprise on Compute Cloud@Customer, enabling businesses to run AI/ML workloads using NVIDIA L40S GPUs. NVIDIA AI Enterprise provides a production-ready AI environment optimized for generative AI, co-pilots, and machine learning models. By fully utilizing L40S GPUs, NVIDIA AI Enterprise on Compute Cloud@Customer accelerates AI workloads. This guide outlines the deployment process.

Figure 1. Oracle Compute Cloud@Customer


Workload and Infra Management
Model Deployment
Triton Management Service
Cloud Native Management and Orchestration
GPU Operator / Network Operator
Cluster Management
Base Command Manager Essentials
Infra Acceleration Libraries
Magnum IO, vGPU, CUDA
AI Use Cases and Workflows
Hello
LLM
Speech AI
Recommenders
Cybersecurity
Medical Imaging
Video Analytics
Route Optimization
•••
More
AI Development
Data Prep
RAPIDS
Model Training
TAO, PyTorch / TensorFlow
Simulate and Test
TensorRT
Deploy at Scale
Triton Inference Server
ORACLE
NVIDIA L40S
Kubernetes/OKE/OpenShift
KVM Hypervisor
Oracle Compute Cloud@Customer
• Hardware Acceleration with L40S Nvidia GPU
• Cloud Consumption Model
• Oracle Owned and Manage
• Hybrid and disconnected topologies
COMPUTE CLOUD@CUSTOMER
ORACLE
C

This content is provided for informational purposes and self-supported guidance only. Consultancy or other assistance related to the content is not covered under the Oracle Support contract or associated service requests. If you have questions or additional needs, then please do reach out to your Oracle Sales contact directly.

Deployment Considerations and Prerequisites

For this installation we have used NVIDIA AI Enterprise platform release 6.0, within the Compute Cloud@Customer, M3.10.3.

Prerequisites

Prior to commencing the installation, it is essential to ensure that all necessary prerequisites are satisfied.

Copyright © 2025, Oracle and/or its affiliates / Public

- **Oracle Compute Cloud@Customer Environment:** Access to a GPU-enabled Oracle Compute Cloud@Customer deployment with administrative privileges. Select shapes that support NVIDIA L40S GPUs.
- **Compatible OS Image:** An NVIDIA AI Enterprise image (based on Ubuntu) or a custom-built Oracle Linux 8 image should be imported and ready as a custom image on Compute Cloud@Customer.
- **System Resource Configuration:** Use a VM shape with the L40S GPU, assign adequate memory/OCPUs, and set the boot volume to VPU 20 for optimal GPU performance.
- **Operating System and Driver Compatibility:** Ensure the Ubuntu OS version is compatible with the required NVIDIA drivers, CUDA Toolkit, and NVIDIA Container Toolkit
- **NVIDIA Driver and CUDA Toolkit Compatibility:** Ensure the correct version of the NVIDIA driver for the L40S GPU is selected and verify its compatibility with the CUDA Toolkit version being installed.
- **NVIDIA Software Components (available locally or via NGC):**
  - NVIDIA GPU driver
  - CUDA Toolkit
  - NVIDIA Container Toolkit
  - NGC CLI
- **NGC CLI Compatibility**: Ensure the NGC CLI version is compatible with the operating system version used in the NVAIE deployment.
- **NGC Account and API Key:** Required for accessing models and containers from NVIDIA's catalog.
- **Network and Proxy Configuration:** Outbound HTTPS access to ngc.nvidia.com, nvcr.io, etc. If using a proxy, configure it system-wide for CLI and package tools.
- **Docker Runtime with NVIDIA Support:** Docker must be installed and configured to use the NVIDIA runtime. NGC registry login requires the API key.
- **Admin Access and Security:** Sudo privileges are required for installation steps. Ensure firewall allows necessary outbound traffic.

NVIDIA AI Enterprise Deployment on Oracle Compute Cloud@Customer

These steps describe how to deploy NVIDIA AI Enterprise on Oracle Compute Cloud@Customer for running AI/ML solutions using NGC images and pre-trained models.

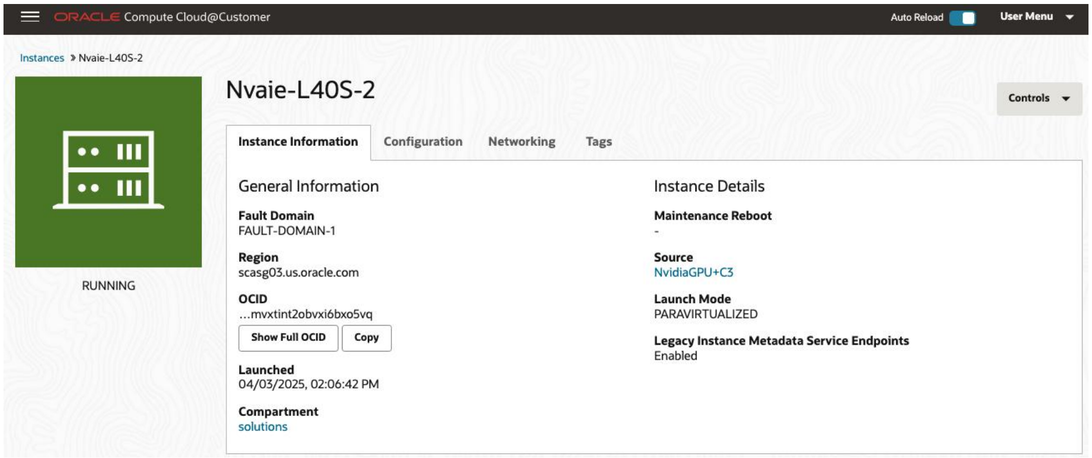
Step1: Launching NVIDIA AI Enterprise Instance

Launch the NVIDIA AI Enterprise instance by creating an instance on Oracle Compute Cloud@Customer.

Note: Select a shape that includes the NVIDIA L40S GPU to leverage hardware acceleration for AI workloads. Set the VPU count to 20 for proper configuration.

Copyright © 2025, Oracle and/or its affiliates / Public

Figure 2. NVIDIA AI Enterprise on Oracle Compute Cloud@Customer



Nvaie-L40S-2

RUNNING

General Information

Fault Domain
FAULT-DOMAIN-1

Region
scasg03.us.oracle.com

OCID
...mvxtint2obvxi6bxo5vq

Launched
04/03/2025, 02:06:42 PM

Compartment
solutions

Instance Details

Maintenance Reboot
-

Source
NvidiaGPU+C3

Launch Mode
PARAVIRTUALIZED

Legacy Instance Metadata Service Endpoints
Enabled

Step 2: OS-Specific Configuration and Installation

Follow the instructions specific to your chosen operating system.

2.1 Configuration for Ubuntu

1. Install Docker

Log into the instance, then update packages and install Docker.

```
# 1. Remove any old Docker-related packages

for pkg in docker.io docker-doc docker-compose docker-compose-v2 podman-docker
containerd runc; do sudo apt-get remove -y $pkg; done


# 2. Update and install prerequisites

sudo apt-get update

sudo apt-get install -y ca-certificates curl gnupg build-essential


# 3. Add Docker's official GPG key

sudo install -m 0755 -d /etc/apt/keyrings

sudo curl -fsSL https://download.docker.com/linux/ubuntu/gpg -o
/etc/apt/keyrings/docker.asc

sudo chmod a+r /etc/apt/keyrings/docker.asc


# 4. Add Docker repository to Apt sources

echo \

  "deb [arch=$(dpkg --print-architecture) signed-by=/etc/apt/keyrings/docker.asc] \

  https://download.docker.com/linux/ubuntu \

  $(. /etc/os-release && echo ${UBUNTU_CODENAME:-$VERSION_CODENAME}) stable" | \

  sudo tee /etc/apt/sources.list.d/docker.list > /dev/null
```

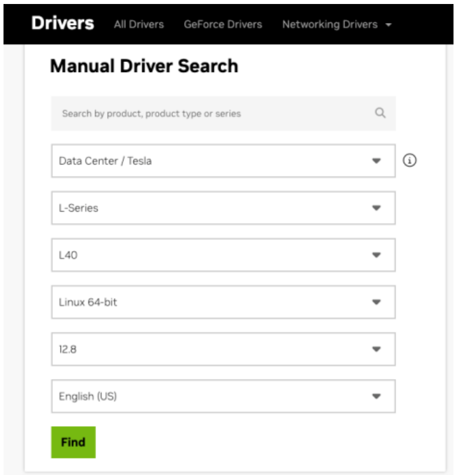
Copyright © 2025, Oracle and/or its affiliates / Public

```
# 5. Update apt and install Docker components

sudo apt-get update

sudo apt-get install -y docker-ce docker-ce-cli containerd.io docker-buildx-plugin
docker-compose-plugin


# 6. Test Docker installation

sudo docker run hello-world
```

2. Install NVIDIA Driver

Download the driver from NVIDIA Driver Downloads. Select the product type and series manually.

Figure 3. NVIDIA Driver Search



Data Center / Tesla

L-Series

L40

Linux 64-bit

12.8

English (US)

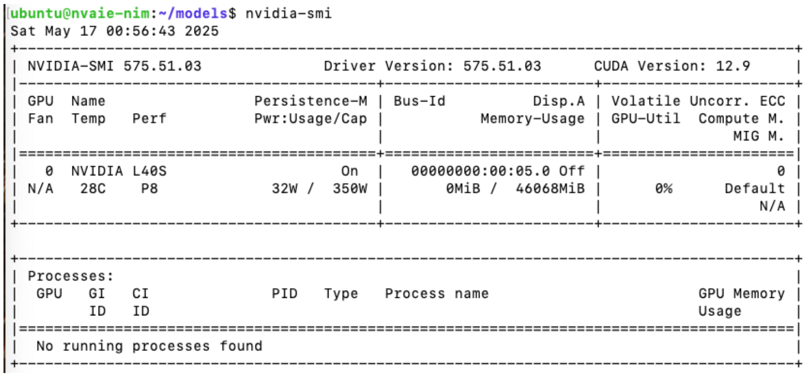
Copy the downloaded .run file to the instance, make it executable, and run the installer.

```
sudo chmod +x NVIDIA-Linux-x86_64-*.run

sudo sh ./NVIDIA-Linux-x86_64-*.run

sudo reboot
```

After rebooting, verify the installation:

```
nvidia-smi
```

Figure 4. NVIDIA Driver Verification

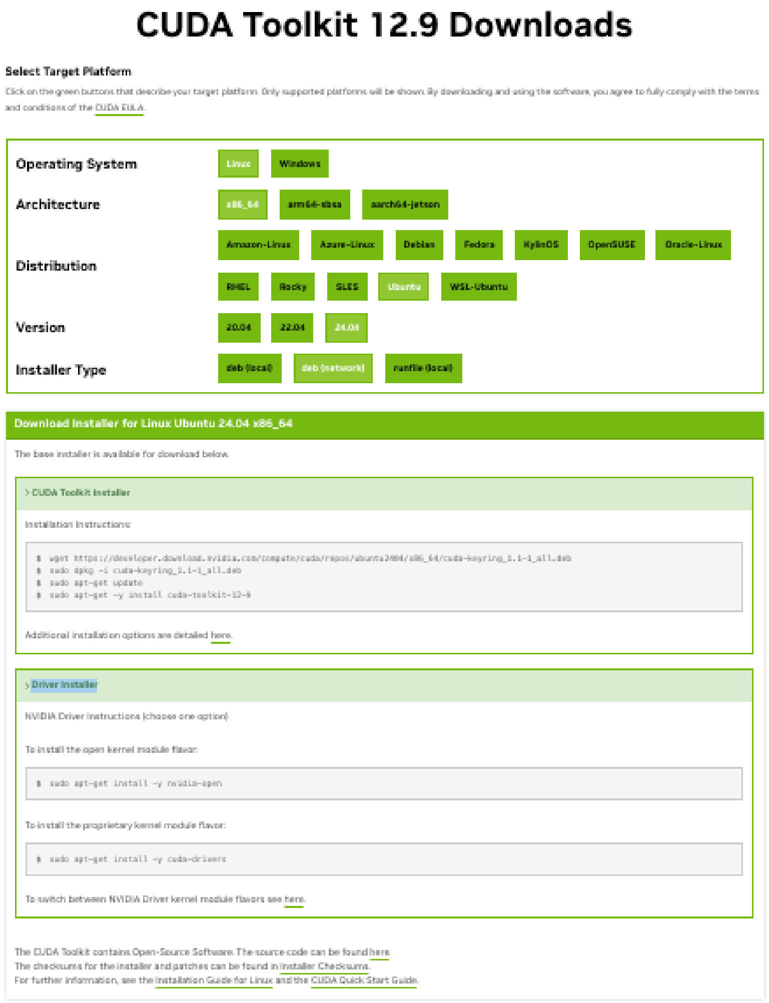


```
ubuntu@nvaie-nim:~/models$ nvidia-smi
Sat May 17 00:56:43 2025
+-----------------------------------------------------------------------------------------+
| NVIDIA-SMI 575.51.03              Driver Version: 575.51.03      CUDA Version: 12.9     |
|-----------------------------------------+------------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M. |
|                                         |                        |               MIG M. |
|=========================================+========================+======================|
|   0  NVIDIA L40S                    On  |   00000000:00:05.0 Off |                    0 |
| N/A   28C    P8             32W /  350W |       0MiB /  46068MiB |      0%      Default |
|                                         |                        |                  N/A |
+-----------------------------------------+------------------------+----------------------+

+-----------------------------------------------------------------------------------------+
| Processes:                                                                              |
|  GPU   GI   CI              PID   Type   Process name                        GPU Memory |
|        ID   ID                                                               Usage      |
|=========================================================================================|
|  No running processes found                                                             |
+-----------------------------------------------------------------------------------------+
```

3. Install CUDA Toolkit

Download the installer from the CUDA Toolkit Download Page.

Figure 5. CUDA Toolkit Download Page



CUDA Toolkit 12.9 Downloads

Select Target Platform

Click on the green buttons that describe your target platform. Only supported platforms will be shown. By downloading and using the software, you agree to fully comply with the terms and conditions of the CUDA EULA.

Download Installer for Linux Ubuntu 24.04 x86_64

The base installer is available for download below.

> CUDA Toolkit Installer

Installation Instructions:

```
$ wget https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2404/x86_64/cuda-keyring_1.1-1_all.deb
$ sudo dpkg -i cuda-keyring_1.1-1_all.deb
$ sudo apt-get update
$ sudo apt-get -y install cuda-toolkit-12-9
```

Additional installation options are detailed here.

> Driver Installer

NVIDIA Driver Instructions (choose one option)

To install the open kernel module flavor:

```
$ sudo apt-get install -y nvidia-open
```

To install the proprietary kernel module flavor:

```
$ sudo apt-get install -y cuda-drivers
```

To switch between NVIDIA Driver kernel module flavors see here.

The CUDA Toolkit contains Open-Source Software. The source code can be found here.
The checksums for the installer and patches can be found in Installer Checksums.
For further information, see the Installation Guide for Linux and the CUDA Quick Start Guide.

```
wget https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64/cuda-keyring_1.1-1_all.deb

sudo dpkg -i cuda-keyring_1.1-1_all.deb

sudo apt-get update

sudo apt-get -y install cuda-toolkit-12-9
```

Verify the Installation:

```
nvcc --version
```

4. Install NVIDIA Container Toolkit

Follow the official guide at Installing NVIDIA Container Toolkit and configure the Docker runtime as instructed.

2.2 Configuration for Oracle Linux 8

1. Base Requirements

Create an Oracle Linux 8 VM instance using the "Minimal Install" image group. Refer to the Oracle Linux 8 Installation Guide for details. All commands should be run as the `opc` user unless specified.

Copyright © 2025, Oracle and/or its affiliates / Public

2. Install CUDA Toolkit

```
sudo dnf config-manager --add-repo https://developer.download.nvidia.com/compute/cuda/repos/rhel8/x86_64/cuda-rhel8.repo

sudo dnf clean all

sudo dnf -y install cuda-toolkit-12-9
```

3. Install CUDA Driver

```
sudo dnf install -y oracle-epel-release-el8

sudo dnf install -y kernel-uek-devel-$(uname -r)

sudo dnf install -y dkms

sudo dnf install -y gcc-toolset-14
```

Download the latest NVIDIA driver `.run` file as shown in **Figure 3**. Then, run the installer as root.

```
sudo su -

chmod +x NVIDIA-Linux-x86_64-*.run

scl enable gcc-toolset-14 bash

./NVIDIA-Linux-x86_64-*.run
```

4. Install Docker

```
sudo dnf config-manager --add-repo https://download.docker.com/linux/rhel/docker-ce.repo

sudo dnf install -y docker-ce docker-ce-cli containerd.io docker-buildx-plugin docker-compose-plugin

sudo systemctl enable --now docker

sudo docker run hello-world
```

5. Enable `opc` User for docker Access

```
sudo usermod -aG docker opc
```

Log out and log back in. Verify access with `docker ps`.

6. Install NVIDIA Container Toolkit

```
curl -s -L https://nvidia.github.io/libnvidia-container/stable/rpm/nvidia-container-toolkit.repo | sudo tee /etc/yum.repos.d/nvidia-container-toolkit.repo
sudo dnf install -y nvidia-container-toolkit
```

Step 3: Accessing Enterprise Catalog and Running AI/ML

The NGC Catalog is a curated set of GPU-optimized software for AI, HPC, and Visualization.

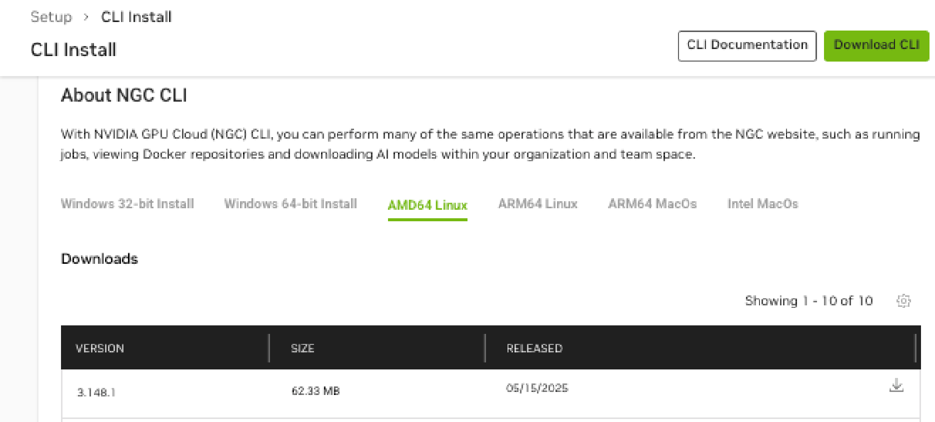
1. Set Up NGC Environment

First, Sign Up for an NVIDIA Cloud Account and Generate a Personal API Key.

Download the NGC CLI "AMD64 Linux" version from the NGC website and transfer it to your server.

Copyright © 2025, Oracle and/or its affiliates / Public

Figure 6. NGC CLI



About NGC CLI

With NVIDIA GPU Cloud (NGC) CLI, you can perform many of the same operations that are available from the NGC website, such as running jobs, viewing Docker repositories and downloading AI models within your organization and team space.

Downloads

Showing 1 - 10 of 10

VERSION	SIZE	RELEASED
3.148.1	62.33 MB	05/15/2025

Install the CLI:

```
sudo apt-get update && sudo apt-get install unzip # For Ubuntu
```

```
sudo dnf install unzip # For Oracle Linux
```

```
unzip ngccli_linux.zip
```

```
cd ngc-cli
```

```
./ngc --version
```

```
export PATH=$PATH:/path/to/ngc-cli
```

A successful installation will display the NGC CLI version number (e.g., NGC CLI 3.148.1).

Add the `ngc-cli` directory to your path to make the command available system-wide permanently. You will need to reload your shell configuration for the change to take effect by running `source ~/.bashrc` or by logging out and back in.

```
**echo 'export PATH=$PATH:/path/to/ngc-cli' >> ~/.bashrc**
```

2. Authenticate with NGC:

```
./ngc config set
```

Enter your API key and other details when prompted. Then, upgrade the CLI.

```
./ngc version upgrade
```

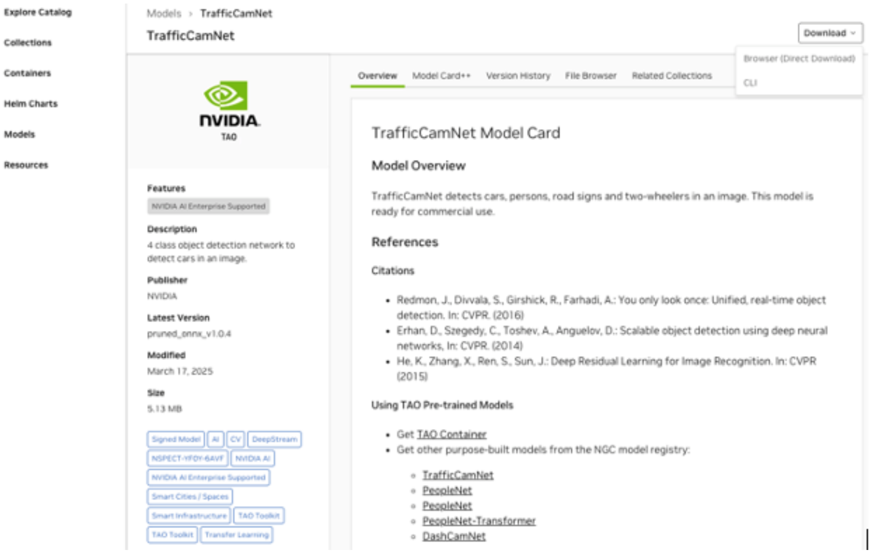
```
./ngc --version
```

3. Pulling NGC Containers and Pre-trained Models

Go to the NGC catalog, find a model, and copy the CLI download command.

Copyright © 2025, Oracle and/or its affiliates / Public

Figure 7. NGC Model



TrafficCamNet

Features

NVIDIA AI Enterprise Supported

Description

4 class object detection network to detect cars in an image.

Publisher

NVIDIA

Latest Version

pruned_onnx_v1.0.4

Modified

March 17, 2025

Size

5.13 MB

TrafficCamNet Model Card

Model Overview

TrafficCamNet detects cars, persons, road signs and two-wheelers in an image. This model is ready for commercial use.

References

Citations

- Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR. (2016)
- Erhan, D., Szegedy, C., Toshev, A., Anguelov, D.: Scalable object detection using deep neural networks, In: CVPR. (2014)
- He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: CVPR (2015)

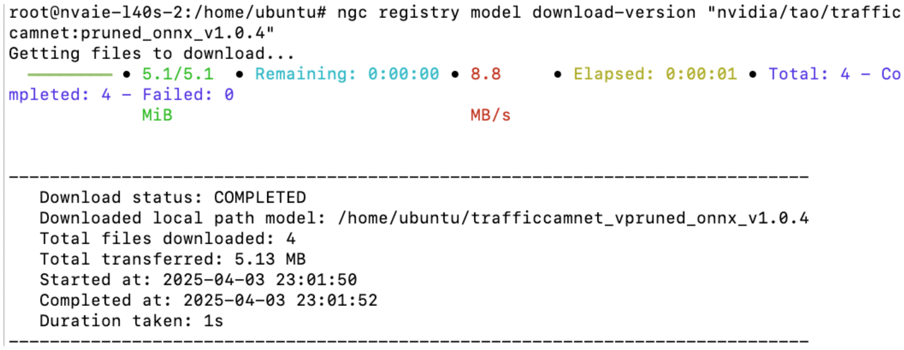
Using TAO Pre-trained Models

- Get TAO Container
- Get other purpose-built models from the NGC model registry:
  - TrafficCamNet
  - PeopleNet
  - PeopleNet
  - PeopleNet-Transformer
  - DashCamNet

Use the `ngc` CLI to pull the model:

```
ngc registry model download-version "nvidia/tao/trafficcamnet:pruned_onnx_v1.0.4"
```

Figure 8. NGC Model Download



```
root@nvaie-l40s-2:/home/ubuntu# ngc registry model download-version "nvidia/tao/traffic
camnet:pruned_v1.0.4"
Getting files to download...
  ━━━━━━━━━━ • 5.1/5.1  • Remaining: 0:00:00 • 8.8     • Elapsed: 0:00:01 • Total: 4 - Co
mpleted: 4 - Failed: 0
             MiB                               MB/s


-----------------------------------------------------------------------------
   Download status: COMPLETED
   Downloaded local path model: /home/ubuntu/trafficcamnet_vpruned_onnx_v1.0.4
   Total files downloaded: 4
   Total transferred: 5.13 MB
   Started at: 2025-04-03 23:01:50
   Completed at: 2025-04-03 23:01:52
   Duration taken: 1s
-----------------------------------------------------------------------------
```

4. Running AI/ML workloads on NVIDIA AI Enterprise

First, log in to the NVIDIA container registry.

```
docker login nvcr.io
```

When prompted, use `$oauthtoken` as the username and your NGC API key as the password.

Pull the desired container and run it, ensuring all GPUs are available to the container.

```
docker pull nvcr.io/nvidia/tao/trafficcamnet:pruned_onnx_v1.0.4
```

```
docker run --rm --runtime=nvidia --gpus all
nvcr.io/nvidia/tao/trafficcamnet:pruned_onnx_v1.0.4
```

This command runs the TrafficCamNet container, utilizing the available GPUs on your Oracle Compute Cloud@Customer instance. The `--rm` flag ensures the container is removed after execution, and the `--gpus all` option ensures that all available GPUs are used for the task.

Additional Resources

- Running Inferencing using NVIDIA NIM on Oracle Compute Cloud@Customer

Copyright © 2025, Oracle and/or its affiliates / Public

- Stream fraud detection with NVIDIA Morpheus on Oracle Compute Cloud@Customer

Acknowledgements

- Sheetal Sabharwal, Principal Product Manager, Oracle Edge Cloud
- Salman Ashfaq, Master Principal Sales Consultant, Oracle Solutions Center
- Anderson Souza, Senior Director, Product Management & Engineering Solutions for Oracle Edge Cloud

Copyright © 2025, Oracle and/or its affiliates / Public

Connect with us

Call **+1.800.ORACLE1** or visit **oracle.com**. Outside North America, find your local office at: **oracle.com/contact**.

blogs.oracle.com   facebook.com/oracle   twitter.com/oracle

Copyright © 2025, Oracle and/or its affiliates. This document is provided for information purposes only, and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document, and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.

Oracle, Java, MySQL, and NetSuite are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.