

Release Notes

RN-11516-001_v26.11 | November 2026

Table of Contents

Chapter 1. CUDA DL Overview 1
Chapter 2. Pulling A Container 2
Chapter 3. Running CUDA DL 3
Chapter 5. CUDA DL Release 26.11 4
Chapter 6. CUDA DL Release 26.09 6
Chapter 7. CUDA DL Release 26.08 8
Chapter 8. CUDA DL Release 26.07 10
Chapter 9. CUDA DL Release 26.06 12
Chapter 10. CUDA DL Release 26.05 14
Chapter 11. CUDA DL Release 26.04 16
Chapter 12. CUDA DL Release 26.03 18
Chapter 13. CUDA DL Release 26.02 20
Chapter 14. CUDA DL Release 26.01 22
Chapter 15. CUDA DL Release 25.12 24
Chapter 16. CUDA DL Release 25.11 26
Chapter 17. CUDA DL Release 25.10 28
Chapter 18. CUDA DL Release 25.09 30
Chapter 19. CUDA DL Release 25.08 32
Chapter 20. CUDA DL Release 25.06 34
Chapter 21. CUDA DL Release 25.05 36
Chapter 22. CUDA DL Release 25.04 38
Chapter 23. CUDA DL Release 25.03 40
Chapter 24. CUDA DL Release 25.02 42

Chapter 1. CUDA DL Overview

CUDA, developed by NVIDIA, is a parallel computing platform and programming model for GPU. With CUDA, developers can dramatically speed up computing applications by harnessing the power of GPUs. The CUDA Toolkit includes libraries, a compiler, development tools, and the CUDA runtime needed for GPU-accelerated development.

Chapter 2. Pulling A Container

About this task

Before you can pull a container from the NGC container registry:

- Install Docker.
  - For NVIDIA DGX™ users, see Preparing to use NVIDIA Containers Getting Started Guide.
  - For non-DGX users, see NVIDIA® GPU Cloud™ (NGC) container registry installation documentation based on your platform.
- Ensure that you have an NGC API Key to log in to the NGC container registry.

  See the NGC Getting Started Guide for more information.

The deep learning frameworks, the NGC Docker containers, and the deep learning framework containers are stored in the

`nvcr.io/nvidia` repository.

Chapter 3. Running CUDA DL

Before you can run an NGC deep learning framework container, your Docker environment must support NVIDIA GPUs. To run a container, issue the appropriate command as explained in Running A Container and specify the registry, repository, and tags.

On a system with GPU support for NGC containers, when you run a container, the following occurs when running a container:

- The container runtime loads the image into a container which runs the software.
- You define the container's runtime resources by including the additional flags and settings that are used with the command.

  These flags and settings are described in Running A Container.
- The GPUs are explicitly defined for the Docker® container, which defaults to all GPUs, but can be specified by using the `NVIDIA_VISIBLE_DEVICES` environment variable.

To run a container, issue the appropriate command as explained in the Running A Container chapter in the NVIDIA Containers For Deep Learning Frameworks User's Guide and specify the registry, repository, and tags. For more information about using NGC, refer to the NGC Container User Guide.

A typical command to launch the runtime/devel container is:

```
docker run --gpus all -it --rm
          nvcri.io/nvidia/cuda-dl-base:YY.MM-cuda<xx.y>-<devel|runtime>-
ubuntu<YY.MM>
```

Where:

YY.MM-cuda<xx.y>-<devel|runtime>-ubuntu<YY.MM> is the container version with "YY.MM" as the release number, "cuda<xx.y>" as CUDA version used in this container and "ubuntu<YY.MM>" as OS this container is built for.

For example:

```
docker run --gpus all -it --rm
          nvcr.io/nvidia/cuda-dl-base:25.03-cuda12.9-<devel|runtime>-ubuntu24.04
```

Chapter 5. CUDA DL Release 26.11

The NVIDIA container image for CUDA Deep Learning is available on NGC.

.

Contents of the CUDA DL container

The container includes the following (not a complete list):

- Ubuntu 24.04 including Python 3.12
- NVIDIA CUDA 13.5.0
- NVIDIA cuBLAS 13.10.0.20
- NVIDIA cuDNN 9.28.0.42
- NVIDIA NCCL 2.32.6
- rdma-core 65.0
- NVIDIA HPC-X 2.51
- NVIDIA GDRCopy 2.5.1
- NVIDIA DOCA 3.5.0
- AWS OFI NCCL 1.12.1
- Nsight Compute 2026.4.0.10
- Nsight Systems 2026.5.2.323
- NVIDIA TensorRT™ 11.5.0.61

Driver Requirements

Release 26.11 is based on CUDA 13.5.0. For comprehensive and up-to-date driver compatibility information, please refer to the following documentation:

- NVIDIA CUDA Compatibility Guide - Compati.bility information between CUDA versions and driver releases.
- CUDA Toolkit Release Notes - Driver version requirements and compatibility matrices.
- NVIDIA Drivers Download - Latest NVIDIA drivers.

Key Features and Enhancements

This CUDA DL release includes the following key features and enhancements.

- CUDA DL container image is based on Ubuntu 24.04.
- NIXL, NVIDIA's high-performance network data transfer library, is now included in inference-level containers.
- The GCP NCCL plugin package is now installed in the CUDA base image. On Google Cloud Platform hosts (A3 Ultra, A4, A4X and A4X Max machine types), the container automatically detects and activates the GCP NCCL network, tuner, profiler, and environment plugins for optimized multi-node communication.
  - On GCP hosts with the GCP NCCL plugins installed, NCCL telemetry is enabled by default starting in this release.
  - A runtime advisory message is displayed when running on GCP hosts with the GCP NCCL plugins active, directing users to the GCP NCCL telemetry documentation.

Announcements

- Starting in 26.11, the DLFW containers introduce general availability support for the NVIDIA Rubin GPU architecture.
- Starting in 26.07, the CUDA DL images contain the GCP NCCL (gIB) plugin.
- Starting in 26.03, the CUDA DL Inference Devel container includes NIXL, for optimized cross-node data transfers, along with the nixlbench benchmarking tool.
- Starting in with the 26.01 release, the CUDA DL Base container is available in four variants optimized for both training and inference use cases
  - Introducing inference-runtime and inference-devel container variants, providing size reduction compared to full development images through stub library optimization for CUSPARSE, CUSOLVER, CUFFT, and NVJITLINK.
- Starting with 25.08 release, CUDA DL Image will be on CUDA 13.0
- Starting with the 25.03 release, NVIDIA provides two CUDA DL images: Developer and Runtime

Known Issues

- Certain distributed jobs may encounter UCX errors involving MLX5, in which case we recommend the use of the environment variable as a workaround: UCX_DC_MLX5_AR_ENABLE=n.

Chapter 6. CUDA DL Release 26.09

The NVIDIA container image for CUDA Deep Learning is available on NGC.

.

Contents of the CUDA DL container

The container includes the following (not a complete list):

- Ubuntu 24.04 including Python 3.12
- NVIDIA CUDA 13.4.1
- NVIDIA cuBLAS 13.84
- NVIDIA cuDNN 9.26.0
- NVIDIA NCCL 2.31
- rdma-core 64.0
- NVIDIA HPC-X 2.51
- NVIDIA GDRCopy 2.5.1
- NVIDIA DOCA 3.5.0
- AWS OFI NCCL 1.17.3
- Nsight Compute 2026.3.0
- Nsight Systems 2026.5.1
- NVIDIA TensorRT™ 11.3.0

Driver Requirements

Release 26.09 is based on CUDA 13.4.1. For comprehensive and up-to-date driver compatibility information, please refer to the following documentation:

- NVIDIA CUDA Compatibility Guide - Compati.bility information between CUDA versions and driver releases.
- CUDA Toolkit Release Notes - Driver version requirements and compatibility matrices.
- NVIDIA Drivers Download - Latest NVIDIA drivers.

Key Features and Enhancements

This CUDA DL release includes the following key features and enhancements.

- CUDA DL container image is based on Ubuntu 24.04.
- NIXL, NVIDIA's high-performance network data transfer library, is now included in inference-level containers.
- The GCP NCCL plugin package is now installed in the CUDA base image. On Google Cloud Platform hosts (A3 Ultra, A4, A4X and A4X Max machine types), the container automatically detects and activates the GCP NCCL network, tuner, profiler, and environment plugins for optimized multi-node communication.
  - On GCP hosts with the GCP NCCL plugins installed, NCCL telemetry is enabled by default starting in this release.
  - A runtime advisory message is displayed when running on GCP hosts with the GCP NCCL plugins active, directing users to the GCP NCCL telemetry documentation.

Announcements

- The 26.09 DLFW containers introduce initial support for the NVIDIA Rubin GPU architecture.
- Starting in 26.07, the CUDA DL images contain the GCP NCCL (gIB) plugin.
- Starting in 26.03, the CUDA DL Inference Devel container includes NIXL, for optimized cross-node data transfers, along with the nixlbench benchmarking tool.
- Starting in with the 26.01 release, the CUDA DL Base container is available in four variants optimized for both training and inference use cases
  - Introducing inference-runtime and inference-devel container variants, providing size reduction compared to full development images through stub library optimization for CUSPARSE, CUSOLVER, CUFFT, and NVJITLINK.
- Starting with 25.08 release, CUDA DL Image will be on CUDA 13.0
- Starting with the 25.03 release, NVIDIA provides two CUDA DL images: Developer and Runtime

Known Issues

- Certain distributed jobs may encounter UCX errors involving MLX5, in which case we recommend the use of the environment variable as a workaround: UCX_DC_MLX5_AR_ENABLE=n.

Chapter 7. CUDA DL Release 26.08

The NVIDIA container image for CUDA Deep Learning is available on NGC.

.

Contents of the CUDA DL container

The container includes the following (not a complete list):

- Ubuntu 24.04 including Python 3.12
- NVIDIA CUDA 13.4.1
- NVIDIA cuBLAS 13.7
- NVIDIA cuDNN 9.25.0
- NVIDIA NCCL 2.30.7
- NVIDIA HPC-X 2.50
- NVIDIA GDRCopy 2.5.1
- NVIDIA DOCA 3.4.0
- AWS OFI NCCL 1.17.3
- TensorBoard 2.16.2
- Nsight Compute 2026.3.0
- Nsight Systems 2026.5.1
- NVIDIA TensorRT™ 11.2.1.2
- Elastic Fabric Adapter 1.48.0
- rdma-core 63.0
- UCX 1.2.1
- GCP NCCL (gIB) 1.1.2

Driver Requirements

Release 26.08 is based on CUDA 13.4.1. For comprehensive and up-to-date driver compatibility information, please refer to the following documentation:

- NVIDIA CUDA Compatibility Guide - Compati.bility information between CUDA versions and driver releases.
- CUDA Toolkit Release Notes - Driver version requirements and compatibility matrices.
- NVIDIA Drivers Download - Latest NVIDIA drivers.

Key Features and Enhancements

This CUDA DL release includes the following key features and enhancements.

- CUDA DL container image is based on Ubuntu 24.04.
- NIXL, NVIDIA's high-performance network data transfer library, is now included in inference-level containers.
- The GCP NCCL plugin package is now installed in the CUDA base image. On Google Cloud Platform hosts (A3 Ultra, A4, A4X and A4X Max machine types), the container automatically detects and activates the GCP NCCL network, tuner, profiler, and environment plugins for optimized multi-node communication.
  - On GCP hosts with the GCP NCCL plugins installed, NCCL telemetry is enabled by default starting in this release.
  - A runtime advisory message is displayed when running on GCP hosts with the GCP NCCL plugins active, directing users to the GCP NCCL telemetry documentation.

Announcements

- Starting in 26.07, the CUDA DL images contain the GCP NCCL (gIB) plugin.
- Starting in 26.03, the CUDA DL Inference Devel container includes NIXL, for optimized cross-node data transfers, along with the nixlbench benchmarking tool.
- Starting in with the 26.01 release, the CUDA DL Base container is available in four variants optimized for both training and inference use cases
  - Introducing inference-runtime and inference-devel container variants, providing size reduction compared to full development images through stub library optimization for CUSPARSE, CUSOLVER, CUFFT, and NVJITLINK.
- Starting with 25.08 release, CUDA DL Image will be on CUDA 13.0
- Starting with the 25.03 release, NVIDIA provides two CUDA DL images: Developer and Runtime

Known Issues

- Certain distributed jobs may encounter UCX errors involving MLX5, in which case we recommend the use of the environment variable as a workaround: UCX_DC_MLX5_AR_ENABLE=n.

Chapter 8. CUDA DL Release 26.07

The NVIDIA container image for CUDA Deep Learning is available on NGC.

.

Contents of the CUDA DL container

The container includes the following (not a complete list):

- Ubuntu 24.04 including Python 3.12
- NVIDIA CUDA 13.3.1
- NVIDIA cuBLAS 13.6.0.2
- NVIDIA cuDNN 9.24.0.43
- NVIDIA NCCL 2.30.7
- NVIDIA HPC-X 2.50
- NVIDIA GDRCopy 2.5.1
- NVIDIA DOCA 3.4.0
- AWS OFI NCCL 1.19.0
- TensorBoard 2.16.2
- Nsight Compute 2026.2.1.5
- Nsight Systems 2026.3.1.117
- NVIDIA TensorRT™ 11.1.0.106
- Elastic Fabric Adapter 1.48.0
- rdma-core 63.0
- UCX 1.2.1
- GCP NCCL (gIB) 1.1.2

Driver Requirements

Release 26.07 is based on CUDA 13.3.1. For comprehensive and up-to-date driver compatibility information, please refer to the following documentation:

- NVIDIA CUDA Compatibility Guide - Compati.bility information between CUDA versions and driver releases.
- CUDA Toolkit Release Notes - Driver version requirements and compatibility matrices.
- NVIDIA Drivers Download - Latest NVIDIA drivers.

Key Features and Enhancements

This CUDA DL release includes the following key features and enhancements.

- CUDA DL container image is based on Ubuntu 24.04.
- NIXL, NVIDIA's high-performance network data transfer library, is now included in inference-level containers.
- The GCP NCCL plugin package is now installed in the CUDA base image. On Google Cloud Platform hosts (A3 Ultra, A4, A4X and A4X Max machine types), the container automatically detects and activates the GCP NCCL network, tuner, profiler, and environment plugins for optimized multi-node communication.
  - On GCP hosts with the GCP NCCL plugins installed, NCCL telemetry is enabled by default starting in this release.
  - A runtime advisory message is displayed when running on GCP hosts with the GCP NCCL plugins active, directing users to the GCP NCCL telemetry documentation.

Announcements

- Starting in 26.07, the CUDA DL images contain the GCP NCCL (gIB) plugin.
- Starting in 26.03, the CUDA DL Inference Devel container includes NIXL, for optimized cross-node data transfers, along with the nixlbench benchmarking tool.
- Starting in with the 26.01 release, the CUDA DL Base container is available in four variants optimized for both training and inference use cases
  - Introducing inference-runtime and inference-devel container variants, providing size reduction compared to full development images through stub library optimization for CUSPARSE, CUSOLVER, CUFFT, and NVJITLINK.
- Starting with 25.08 release, CUDA DL Image will be on CUDA 13.0
- Starting with the 25.03 release, NVIDIA provides two CUDA DL images: Developer and Runtime

Known Issues

- Certain distributed jobs may encounter UCX errors involving MLX5, in which case we recommend the use of the environment variable as a workaround: UCX_DC_MLX5_AR_ENABLE=n.

Chapter 9. CUDA DL Release 26.06

The NVIDIA container image for CUDA Deep Learning is available on NGC.

.

Contents of the CUDA DL container

The container includes the following (not a complete list):

- Ubuntu 24.04 including Python 3.12
- NVIDIA CUDA 13.3.0
- NVIDIA cuBLAS 13.5.1.27
- NVIDIA cuDNN 9.23.0.39
- NVIDIA NCCL 2.30.4
- rdma-core 63.0
- NVIDIA HPC-X 2.50
- OpenUCX 1.20.0
- NVIDIA GDRCopy 2.5.1
- NVIDIA DOCA 3.4.0
- AWS OFI NCCL 1.17.3
- TensorBoard 2.16.2
- Nsight Compute 2026.2.0.7
- Nsight Systems 2026.3.1.117
- NVIDIA TensorRT™ 11.0.0114

Driver Requirements

Release 26.06 is based on CUDA 13.3.0. For comprehensive and up-to-date driver compatibility information, please refer to the following documentation:

- NVIDIA CUDA Compatibility Guide - Compati.bility information between CUDA versions and driver releases.

- CUDA Toolkit Release Notes - Driver version requirements and compatibility matrices.
- NVIDIA Drivers Download - Latest NVIDIA drivers.

Key Features and Enhancements

This CUDA DL release includes the following key features and enhancements.

- CUDA DL container image is based on Ubuntu 24.04.
- NIXL, NVIDIA's high-performance network data transfer library, is now included in inference-level containers.

Announcements

- Starting in 26.03, the CUDA DL Inference Devel container includes NIXL, for optimized cross-node data transfers, along with the nixlbench benchmarking tool.
- Starting in with the 26.01 release, the CUDA DL Base container is available in four variants optimized for both training and inference use cases
  - Introducing inference-runtime and inference-devel container variants, providing size reduction compared to full development images through stub library optimization for CUSPARSE, CUSOLVER, CUFFT, and NVJITLINK.
- Starting with 25.08 release, CUDA DL Image will be on CUDA 13.0
- Starting with the 25.03 release, NVIDIA provides two CUDA DL images: Developer and Runtime

Known Issues

- Certain distributed jobs may encounter UCX errors involving MLX5, in which case we recommend the use of the environment variable as a workaround: UCX_DC_MLX5_AR_ENABLE=n.

Chapter 10. CUDA DL Release 26.05

The NVIDIA container image for CUDA Deep Learning is available on NGC.

.

Contents of the CUDA DL container

The container includes the following (not a complete list):

- Ubuntu 24.04 including Python 3.12
- NVIDIA CUDA 13.2.1
- NVIDIA cuBLAS 13.4.1.1
- NVIDIA cuDNN 9.22.0.52
- NVIDIA NCCL 2.30.4
- rdma-core 59.1
- NVIDIA HPC-X 2.26
- OpenUCX 1.20.0
- NVIDIA GDRCopy 2.5.1
- NVIDIA DOCA 3.2.0
- AWS OFI NCCL 1.17.3
- TensorBoard 2.16.2
- Nsight Compute 2026.1.1.2
- Nsight Systems 2026.2.1.210
- NVIDIA TensorRT™ 10.16.1.11

Driver Requirements

Release 26.05 is based on CUDA 13.2.1. For comprehensive and up-to-date driver compatibility information, please refer to the following documentation:

- NVIDIA CUDA Compatibility Guide - Compati.bility information between CUDA versions and driver releases.

- CUDA Toolkit Release Notes - Driver version requirements and compatibility matrices.
- NVIDIA Drivers Download - Latest NVIDIA drivers.

Key Features and Enhancements

This CUDA DL release includes the following key features and enhancements.

- CUDA DL container image is based on Ubuntu 24.04.
- NIXL, NVIDIA's high-performance network data transfer library, is now included in inference-level containers.

Announcements

- Starting in 26.03, the CUDA DL Inference Devel container includes NIXL, for optimized cross-node data transfers, along with the nixlbench benchmarking tool.
- Starting in with the 26.01 release, the CUDA DL Base container is available in four variants optimized for both training and inference use cases
  - Introducing inference-runtime and inference-devel container variants, providing size reduction compared to full development images through stub library optimization for CUSPARSE, CUSOLVER, CUFFT, and NVJITLINK.
- Starting with 25.08 release, CUDA DL Image will be on CUDA 13.0
- Starting with the 25.03 release, NVIDIA provides two CUDA DL images: Developer and Runtime

Known Issues

- Certain distributed jobs may encounter UCX errors involving MLX5, in which case we recommend the use of the environment variable as a workaround: UCX_DC_MLX5_AR_ENABLE=n.

Chapter 11. CUDA DL Release 26.04

The NVIDIA container image for CUDA Deep Learning is available on NGC.

.

Contents of the CUDA DL container

The container includes the following (not a complete list):

- Ubuntu 24.04 including Python 3.12
- NVIDIA CUDA 13.2.1
- NVIDIA cuBLAS 13.4.0.1
- NVIDIA cuDNN 9.21.0.82
- NVIDIA NCCL 2.29.7
- rdma-core 59.1
- NVIDIA HPC-X 2.26
- OpenUCX 1.20.0
- NVIDIA GDRCopy 2.5.1
- NVIDIA DOCA 3.2.0
- AWS OFI NCCL 1.17.3
- TensorBoard 2.16.2
- Nsight Compute 2026.1.1.2
- Nsight Systems 2026.2.1.210
- NVIDIA TensorRT™ 10.16.1.11

Driver Requirements

Release 26.04 is based on CUDA 13.2.1. For comprehensive and up-to-date driver compatibility information, please refer to the following documentation:

- NVIDIA CUDA Compatibility Guide - Compati.bility information between CUDA versions and driver releases.

- CUDA Toolkit Release Notes - Driver version requirements and compatibility matrices.
- NVIDIA Drivers Download - Latest NVIDIA drivers.

Key Features and Enhancements

This CUDA DL release includes the following key features and enhancements.

- CUDA DL container image is based on Ubuntu 24.04.
- NIXL, NVIDIA's high-performance network data transfer library, is now included in inference-level containers.

Announcements

- Starting in 26.03, the CUDA DL Inference Devel container includes NIXL, for optimized cross-node data transfers, along with the nixlbench benchmarking tool.
- Starting in with the 26.01 release, the CUDA DL Base container is available in four variants optimized for both training and inference use cases
  - Introducing inference-runtime and inference-devel container variants, providing size reduction compared to full development images through stub library optimization for CUSPARSE, CUSOLVER, CUFFT, and NVJITLINK.
- Starting with 25.08 release, CUDA DL Image will be on CUDA 13.0
- Starting with the 25.03 release, NVIDIA provides two CUDA DL images: Developer and Runtime

Known Issues

- Certain distributed jobs may encounter UCX errors involving MLX5, in which case we recommend the use of the environment variable as a workaround: UCX_DC_MLX5_AR_ENABLE=n.

Chapter 12. CUDA DL Release 26.03

The NVIDIA container image for CUDA Deep Learning is available on NGC.

.

Contents of the CUDA DL container

The container includes the following (not a complete list):

- Ubuntu 24.04 including Python 3.12
- NVIDIA CUDA 13.2.0.046
- NVIDIA cuBLAS 13.3.0.5
- NVIDIA cuDNN 9.20.0.48
- NVIDIA NCCL 2.29.7
- rdma-core 56.0
- NVIDIA HPC-X 2.26
- OpenUCX 1.20.0
- NVIDIA GDRCopy 2.5.1
- NVIDIA DOCA 3.2.0
- AWS OFI NCCL 1.17.0
- TensorBoard 2.16.2
- Nsight Compute 2026.1.0.9
- Nsight Systems 2026.1.2.63
- NVIDIA TensorRT™ 10.16.0.72

Driver Requirements

Release 26.03 is based on CUDA 13.2.0. For comprehensive and up-to-date driver compatibility information, please refer to the following documentation:

- NVIDIA CUDA Compatibility Guide - Compati.bility information between CUDA versions and driver releases.

- CUDA Toolkit Release Notes - Driver version requirements and compatibility matrices.
- NVIDIA Drivers Download - Latest NVIDIA drivers.

Key Features and Enhancements

This CUDA DL release includes the following key features and enhancements.

- CUDA DL container image is based on Ubuntu 24.04.
- NIXL, NVIDIA's high-performance network data transfer library, is now included in inference-level containers.

Announcements

- Starting in 26.03, the CUDA DL Inference Devel container includes NIXL, for optimized cross-node data transfers, along with the nixlbench benchmarking tool.
- Starting in with the 26.01 release, the CUDA DL Base container is available in four variants optimized for both training and inference use cases
  - Introducing inference-runtime and inference-devel container variants, providing size reduction compared to full development images through stub library optimization for CUSPARSE, CUSOLVER, CUFFT, and NVJITLINK.
- Starting with 25.08 release, CUDA DL Image will be on CUDA 13.0
- Starting with the 25.03 release, NVIDIA provides two CUDA DL images: Developer and Runtime

Known Issues

- Certain distributed jobs may encounter UCX errors involving MLX5, in which case we recommend the use of the environment variable as a workaround: UCX_DC_MLX5_AR_ENABLE=n.

Chapter 13. CUDA DL Release 26.02

The NVIDIA container image for CUDA Deep Learning is available on NGC.

.

Contents of the CUDA DL container

The container includes the following (not a complete list):

- Ubuntu 24.04 including Python 3.12
- NVIDIA CUDA 13.1.1.006
- NVIDIA cuBLAS 13.2.1.1
- NVIDIA cuDNN 9.19.0.56
- NVIDIA NCCL 2.29.2
- rdma-core 56.0
- NVIDIA HPC-X 2.25.1
- OpenUCX 1.20.0
- NVIDIA GDRCopy 2.5.1
- NVIDIA DOCA 3.1.0
- AWS OFI NCCL 1.17.0
- TensorBoard 2.16.2
- Nsight Compute 2025.4.12
- Nsight Systems 2026.1.1.204
- NVIDIA TensorRT™ 10.15.1.26

Driver Requirements

Release 26.02 is based on CUDA 13.1.1. For comprehensive and up-to-date driver compatibility information, please refer to the following documentation:

- NVIDIA CUDA Compatibility Guide - Compati.bility information between CUDA versions and driver releases.

- CUDA Toolkit Release Notes - Driver version requirements and compatibility matrices.
- NVIDIA Drivers Download - Latest NVIDIA drivers.

Key Features and Enhancements

This CUDA DL release includes the following key features and enhancements.

- CUDA DL container image is based on Ubuntu 24.04.

Announcements

- Starting in with the 26.01 release, the CUDA DL Base container is available in four variants optimized for both training and inference use cases
  - Introducing inference-runtime and inference-devel container variants, providing size reduction compared to full development images through stub library optimization for CUSPARSE, CUSOLVER, CUFFT, and NVJITLINK.
- Starting with 25.08 release, CUDA DL Image will be on CUDA 13.0
- Starting with the 25.03 release, NVIDIA provides two CUDA DL images: Developer and Runtime

Known Issues

- Certain distributed jobs may encounter UCX errors involving MLX5, in which case we recommend the use of the environment variable as a workaround: UCX_DC_MLX5_AR_ENABLE=n.

Chapter 14. CUDA DL Release 26.01

The NVIDIA container image for CUDA Deep Learning is available on NGC.

.

Contents of the CUDA DL container

The container includes the following (not a complete list):

- Ubuntu 24.04 including Python 3.12
- NVIDIA CUDA 13.1.1.006
- NVIDIA cuBLAS 13.2.1.1
- NVIDIA cuDNN 9.17.1.4
- NVIDIA NCCL 2.29.2
- rdma-core 56.0
- NVIDIA HPC-X 2.25.1
- OpenUCX 1.20.0
- NVIDIA GDRCopy 2.5.1
- NVIDIA DOCA 3.1.0
- AWS OFI NCCL 1.17.0
- TensorBoard 2.16.2
- Nsight Compute 2025.4.12
- Nsight Systems 2025.6.1.190
- NVIDIA TensorRT™ 10.14.1.48

Driver Requirements

Release 26.01 is based on CUDA 13.1.1. For comprehensive and up-to-date driver compatibility information, please refer to the following documentation:

- NVIDIA CUDA Compatibility Guide - Compati.bility information between CUDA versions and driver releases.

- CUDA Toolkit Release Notes - Driver version requirements and compatibility matrices.
- NVIDIA Drivers Download - Latest NVIDIA drivers.

Key Features and Enhancements

This CUDA DL release includes the following key features and enhancements.

- CUDA DL container image is based on Ubuntu 24.04.

Announcements

- Starting in with the 26.01 release, the CUDA DL Base container is available in four variants optimized for both training and inference use cases
  - Introducing inference-runtime and inference-devel container variants, providing size reduction compared to full development images through stub library optimization for CUSPARSE, CUSOLVER, CUFFT, and NVJITLINK.
- Starting with 25.08 release, CUDA DL Image will be on CUDA 13.0
- Starting with the 25.03 release, NVIDIA provides two CUDA DL images: Developer and Runtime

Known Issues

- Certain distributed jobs may encounter UCX errors involving MLX5, in which case we recommend the use of the environment variable as a workaround: UCX_DC_MLX5_AR_ENABLE=n.

Chapter 15. CUDA DL Release 25.12

The NVIDIA container image for CUDA Deep Learning is available on NGC.

.

Contents of the CUDA DL container

The container includes the following (not a complete list):

- Ubuntu 24.04 including Python 3.12
- NVIDIA CUDA 13.1.0
- NVIDIA cuBLAS 13.2.0.9
- NVIDIA cuDNN 9.17.0.29
- NVIDIA NCCL 2.28.9
- rdma-core 56.0
- NVIDIA HPC-X 2.25.1
- OpenUCX 1.20.0
- GDRCopy 2.5.1
- TensorBoard 2.16.2
- Nsight Compute 2025.4.0.12
- Nsight Systems 2025.5.2.266
- NVIDIA TensorRT™ 10.14.1

Driver Requirements

Release 25.12 is based on CUDA 13.1.0. For comprehensive and up-to-date driver compatibility information, please refer to the following documentation:

- NVIDIA CUDA Compatibility Guide - Compati.bility information between CUDA versions and driver releases.
- CUDA Toolkit Release Notes - Driver version requirements and compatibility matrices.
- NVIDIA Drivers Download - Latest NVIDIA drivers.

Key Features and Enhancements

This CUDA DL release includes the following key features and enhancements.

- CUDA DL container image is based on Ubuntu 24.04.

Announcements

- Starting with 25.08 release, CUDA DL Image will be on CUDA 13.0
- Starting with the 25.03 release, NVIDIA provides two CUDA DL images: Developer and Runtime

Known Issues

- Certain distributed jobs may encounter UCX errors involving MLX5, in which case we recommend the use of the environment variable as a workaround: UCX_DC_MLX5_AR_ENABLE=n.

Chapter 16. CUDA DL Release 25.11

The NVIDIA container image for CUDA Deep Learning is available on NGC.

.

Contents of the CUDA DL container

The container includes the following (not a complete list):

- Ubuntu 24.04 including Python 3.12
- NVIDIA CUDA 13.0.2.0.0.6
- NVIDIA cuBLAS 13.1.0.3
- NVIDIA cuDNN 9.15.0.58
- NVIDIA NCCL 2.27.7
- rdma-core 56.0
- NVIDIA HPC-X 2.24.1
- GDRCopy 2.5.1
- TensorBoard 2.16.2
- Nsight Compute 2025.3.1.4
- Nsight Systems 2025.5.1.121
- NVIDIA TensorRT™ 10.14.1.48

Driver Requirements

Release 25.11 is based on CUDA 13.0.2. For comprehensive and up-to-date driver compatibility information, please refer to the following documentation:

- NVIDIA CUDA Compatibility Guide - Compati.bility information between CUDA versions and driver releases.
- CUDA Toolkit Release Notes - Driver version requirements and compatibility matrices.
- NVIDIA Drivers Download - Latest NVIDIA drivers.

Key Features and Enhancements

This CUDA DL release includes the following key features and enhancements.

- CUDA DL container image is based on Ubuntu 24.04.

Announcements

- Starting with 25.08 release, CUDA DL Image will be on CUDA 13.0
- Starting with the 25.03 release, NVIDIA provides two CUDA DL images: Developer and Runtime

Known Issues

- Certain distributed jobs may encounter UCX errors involving MLX5, in which case we recommend the use of the environment variable as a workaround: UCX_DC_MLX5_AR_ENABLE=n.

Chapter 17. CUDA DL Release 25.10

The NVIDIA container image for CUDA Deep Learning is available on NGC.

.

Contents of the CUDA DL container

The container includes the following (not a complete list):

- Ubuntu 24.04 including Python 3.12
- NVIDIA CUDA 13.0.2.0.0.6
- NVIDIA cuBLAS 13.1.0.3
- NVIDIA cuDNN 9.14.0.64
- NVIDIA NCCL 2.27.7
- rdma-core 50.0
- NVIDIA HPC-X 2.24.1
- GDRCopy 2.4.1
- TensorBoard 2.16.2
- Nsight Compute 2025.3.1.4
- Nsight Systems 2025.5.1.121
- NVIDIA TensorRT™ 10.13.3.9

Driver Requirements

Release 25.10 is based on CUDA 13.0.2. For comprehensive and up-to-date driver compatibility information, please refer to the following documentation:

- NVIDIA CUDA Compatibility Guide - Compati.bility information between CUDA versions and driver releases.
- CUDA Toolkit Release Notes - Driver version requirements and compatibility matrices.
- NVIDIA Drivers Download - Latest NVIDIA drivers.

Key Features and Enhancements

This CUDA DL release includes the following key features and enhancements.

- CUDA DL container image is based on Ubuntu 24.04.

Announcements

- Starting with 25.08 release, CUDA DL Image will be on CUDA 13.0
- Starting with the 25.03 release, NVIDIA provides two CUDA DL images: Developer and Runtime

Known Issues

- Certain distributed jobs may encounter UCX errors involving MLX5, in which case we recommend the use of the environment variable as a workaround: UCX_DC_MLX5_AR_ENABLE=n.

Chapter 18. CUDA DL Release 25.09

The NVIDIA container image for CUDA Deep Learning is available on NGC.

.

Contents of the CUDA DL container

The container includes the following (not a complete list):

- Ubuntu 24.04 including Python 3.12
- NVIDIA CUDA 13.0.1.012
- NVIDIA cuBLAS 13.0.2.14
- NVIDIA cuDNN 9.13.0.48
- NVIDIA NCCL 2.27.7
- rdma-core 50.0
- NVIDIA HPC-X 2.24.1
- GDRCopy 2.4.1
- TensorBoard 2.16.2
- Nsight Compute 2025.3.1.4
- Nsight Systems 2025.5.1.121
- NVIDIA TensorRT™ 10.13.3.9

Driver Requirements

Release 25.09 is based on CUDA 13.0.1. For comprehensive and up-to-date driver compatibility information, please refer to the following documentation:

- NVIDIA CUDA Compatibility Guide - Compati.bility information between CUDA versions and driver releases.
- CUDA Toolkit Release Notes - Driver version requirements and compatibility matrices.
- NVIDIA Drivers Download - Latest NVIDIA drivers.

Key Features and Enhancements

This CUDA DL release includes the following key features and enhancements.

- CUDA DL container image is based on Ubuntu 24.04.

Announcements

- Starting with 25.08 release, CUDA DL Image will be on CUDA 13.0
- Starting with the 25.03 release, NVIDIA provides two CUDA DL images: Developer and Runtime

Known Issues

- Certain distributed jobs may encounter UCX errors involving MLX5, in which case we recommend the use of the environment variable as a workaround: UCX_DC_MLX5_AR_ENABLE=n.

Chapter 19. CUDA DL Release 25.08

The NVIDIA container image for CUDA Deep Learning, release 25.08, is available on NGC.

.

Contents of the CUDA DL container

The container includes the following (not a complete list):

- Ubuntu 24.04 including Python 3.12
- NVIDIA CUDA 13.0.0.044
- NVIDIA cuBLAS 13.0.0.19
- NVIDIA cuDNN 9.12.0.46
- NVIDIA NCCL 2.27.7
- rdma-core 50.0
- NVIDIA HPC-X 2.24
- GDRCopy 2.4.1
- TensorBoard 2.16.2
- Nsight Compute 2025.3.0.19
- Nsight Systems 2025.4.1.136
- NVIDIA TensorRT™ 10.13.2.2

Driver Requirements

Release 25.08 is based on CUDA 13.0.0. For comprehensive and up-to-date driver compatibility information, please refer to the following documentation:

- NVIDIA CUDA Compatibility Guide - Compati.bility information between CUDA versions and driver releases.
- CUDA Toolkit Release Notes - Driver version requirements and compatibility matrices.
- NVIDIA Drivers Download - Latest NVIDIA drivers.

Key Features and Enhancements

This CUDA DL release includes the following key features and enhancements.

- CUDA DL container image is based on Ubuntu 24.04.

Announcements

- Starting 25.08 release, CUDA DL Image will be on CUDA 13.0
- Starting with the 25.03 release, NVIDIA provides two CUDA DL images: Developer and Runtime

Known Issues

- Certain distributed jobs may encounter UCX errors involving MLX5, in which case we recommend the use of the environment variable as a workaround: UCX_DC_MLX5_AR_ENABLE=n.

Chapter 20. CUDA DL Release 25.06

The NVIDIA container image for CUDA Deep Learning, release 25.06, is available on NGC.

.

Contents of the CUDA DL container

The container includes the following (not a complete list):

- Ubuntu 24.04 including Python 3.12
- NVIDIA CUDA 12.9.1
- NVIDIA cuBLAS 12.9.1.4
- NVIDIA cuDNN 9.10.2.21
- NVIDIA NCCL 2.27.3
- rdma-core 50.0
- NVIDIA HPC-X 2.23
- GDRCopy 2.4.1
- TensorBoard 2.16.2
- Nsight Compute 2025.2.1.3
- Nsight Systems 2025.3.1.90
- NVIDIA TensorRT™ 10.11.0.33

Driver Requirements

Release 25.06 is based on CUDA 12.9.1. For comprehensive and up-to-date driver compatibility information, please refer to the following documentation:

- NVIDIA CUDA Compatibility Guide - Compati.bility information between CUDA versions and driver releases.
- CUDA Toolkit Release Notes - Driver version requirements and compatibility matrices.
- NVIDIA Drivers Download - Latest NVIDIA drivers.

Key Features and Enhancements

This CUDA DL release includes the following key features and enhancements.

- CUDA DL container image is based on Ubuntu 24.04.

Announcements

- Starting with the 25.03 release, NVIDIA provides two CUDA DL images: Developer and Runtime

Known Issues

- None.

Chapter 21. CUDA DL Release 25.05

The NVIDIA container image for CUDA Deep Learning, release 25.05, is available on NGC.

.

Contents of the CUDA DL container

This container image contains the complete source of the version of PyTorch in `/opt/pytorch`. It is prebuilt and installed in the default Python environment (`/usr/local/lib/python3.10/dist-packages/torch`) in the container image.

The container also includes the following:

- Ubuntu 24.04 including Python 3.12
- NVIDIA CUDA 12.9.0
- NVIDIA cuBLAS 12.9.0
- NVIDIA cuDNN 9.10.1.3
- NVIDIA NCCL 2.26.5
- rdma-core 50.0
- NVIDIA HPC-X 2.23
- GDRCopy 2.4.1
- TensorBoard 2.16.2
- Nsight Compute 2025.2.0.11
- Nsight Systems 2025.3.1.90
- NVIDIA TensorRT™ 10.10.0.31

Driver Requirements

Release 25.05 is based on CUDA 12.9.0. For comprehensive and up-to-date driver compatibility information, please refer to the following documentation:

- NVIDIA CUDA Compatibility Guide - Compati.bility information between CUDA versions and driver releases.

- CUDA Toolkit Release Notes - Driver version requirements and compatibility matrices.
- NVIDIA Drivers Download - Latest NVIDIA drivers.

Key Features and Enhancements

This CUDA DL release includes the following key features and enhancements.

- CUDA DL container image is based on Ubuntu 24.04.

Announcements

- Starting with the 25.03 release, NVIDIA provides two CUDA DL images: Developer and Runtime

Known Issues

- None.

Chapter 22. CUDA DL Release 25.04

The NVIDIA container image for CUDA Deep Learning, release 25.04, is available on NGC.

.

Contents of the CUDA DL container

This container image contains the complete source of the version of PyTorch in `/opt/pytorch`. It is prebuilt and installed in the default Python environment (`/usr/local/lib/python3.10/dist-packages/torch`) in the container image.

The container also includes the following:

- Ubuntu 24.04 including Python 3.12
- NVIDIA CUDA 12.9.0
- NVIDIA cuBLAS 12.9.0
- NVIDIA cuDNN 9.9.0.52
- NVIDIA NCCL 2.26.3
- rdma-core 50.0
- NVIDIA HPC-X 2.22.1
- GDRCopy 2.4.1
- TensorBoard 2.16.2
- Nsight Compute 2025.2.0.11
- Nsight Systems 2025.2.1.130
- NVIDIA TensorRT™ 10.9.0.34

Driver Requirements

Release 25.06 is based on CUDA 12.9.1. For comprehensive and up-to-date driver compatibility information, please refer to the following documentation:

- NVIDIA CUDA Compatibility Guide - Compati.bility information between CUDA versions and driver releases.

- CUDA Toolkit Release Notes - Driver version requirements and compatibility matrices.
- NVIDIA Drivers Download - Latest NVIDIA drivers.

Key Features and Enhancements

This CUDA DL release includes the following key features and enhancements.

- CUDA DL container image is based on Ubuntu 24.04.

Announcements

- Starting with the 25.03 release, NVIDIA provides two CUDA DL images: Developer and Runtime

Known Issues

- None.

Chapter 23. CUDA DL Release 25.03

The NVIDIA container image for PyTorch, release 25.03, is available on NGC.

.

Contents of the PyTorch container

This container image contains the complete source of the version of PyTorch in `/opt/pytorch`. It is prebuilt and installed in the default Python environment (`/usr/local/lib/python3.10/dist-packages/torch`) in the container image.

The container also includes the following:

- Ubuntu 24.04 including Python 3.12
- NVIDIA CUDA 12.8.1.012
- NVIDIA cuBLAS 12.8.4.1
- NVIDIA cuDNN 9.8.0.87
- NVIDIA NCCL 2.25.1
- rdma-core 50.0
- NVIDIA HPC-X 2.21
- SHARP 3.9.0
- UCC 1.4.0
- UCX 1.18.0
- OpenMPI 4.1.7
- GDRCopy 2.4.1
- TensorBoard 2.16.2
- Nsight Compute 2025.1.1.2
- Nsight Systems 2025.1.1.110
- NVIDIA TensorRT™ 10.9.0.34

Driver Requirements

Release 25.03 is based on CUDA 12.8.1 which requires NVIDIA Driver release 570 or later. However, if you are running on a data center GPU (for example, T4 or any other data center GPU), you can use NVIDIA driver release 470.57 (or later R470), 525.85 (or later R525), 535.86 (or later R535), or 545.23 (or later R545).

The CUDA driver's compatibility package only supports particular drivers. Thus, users should upgrade from all R418, R440, R450, R460, R510, R520, R530, R545, R555 and R560 drivers, which are not forward-compatible with CUDA 12.8. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

.

Key Features and Enhancements

This CUDA DL release includes the following key features and enhancements.

- CUDA DL container image version 25.03 is based on Ubuntu 24.04.

Announcements

- Starting with the 25.03 release, NVIDIA provides two CUDA DL images: Developer and Runtime

Known Issues

- 25.03 CUDA DL Runtime image is currently not ready for the networking stack.

Chapter 24. CUDA DL Release 25.02

NVIDIA's CUDA Deep Learning images extend the standard CUDA images by integrating enhanced networking support and a suite of optimized libraries —including cuDNN, cuTensor, NCCL, and HPC-x—to accelerate deep learning workloads. These images serve as a foundation for building your own GPU-accelerated containers.

.

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.

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.

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

Trademarks

NVIDIA, the NVIDIA logo, and cuBLAS, CUDA, DALI, DGX, DGX-1, DGX-2, DGX Station, DLProf, Jetson, Kepler, Maxwell, NCCL, Nsight Compute, Nsight Systems, NvCaffe, PerfWorks, Pascal, SDK Manager, Tegra, TensorRT, Triton Inference Server, Tesla, TF-TRT, and Volta are trademarks and/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

© 2021-2026 NVIDIA Corporation & Affiliates. All rights reserved.

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