TensorRT

Container Release Notes

RN-08823-001 _v24.07 | October 2024

Table of Contents

Chapter 1. TensorRT Overview

The core of NVIDIA® TensorRT™ is a C++ library that facilitates high-performance inference on NVIDIA graphics processing units (GPUs). TensorRT takes a trained network, which consists of a network definition and a set of trained parameters, and produces a highly optimized runtime engine that performs inference for that network.

You can describe a TensorRT network by using a C++ or Python API, or you can import an existing Caffe, ONNX, or TensorFlow model by using one of the provided parsers.

TensorRT provides APIs through C++ and Python that help express deep learning models by using the Network Definition API or load a predefined model by using parsers that allows TensorRT to optimize and run them on an NVIDIA GPU. TensorRT applies graph optimizations, layer fusion, and other optimizations, while also finding the fastest implementation of that model by leveraging a diverse collection of highly optimized kernels. TensorRT also supplies a runtime that you can use to execute this network on all NVIDIA's GPUs from the NVIDIA Pascal™ generation onwards.

TensorRT also includes optional high-speed, mixed precision capabilities that were introduced in Tegra X1 and were extended with the NVIDIA Pascal, NVIDIA Volta™, and NVIDIA Turing™ architectures.

The TensorRT container allows TensorRT samples to be built, modified, and executed. For more information about the TensorRT samples, see the TensorRT Sample Support Guide.

For a complete list of installation options and instructions, refer to Installing TensorRT.

Chapter 2. Pulling A Container

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 access and can log in to the NGC container registry.

  Refer to 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 TensorRT

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:

- The Docker engine loads the image into a container which runs the software.
- You define the runtime resources of the container 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.

  For more information, refer to the nvidia-docker documentation.

> Note: Starting in Docker 19.03, complete the steps below.

The method implemented in your system depends on the DGX OS version that you installed (for DGX systems), the NGC Cloud Image that was provided by a Cloud Service Provider, or the software that you installed to prepare to run NGC containers on TITAN PCs, Quadro PCs, or NVIDIA Virtual GPUs (vGPUs).

1. Issue the command for the applicable release of the container that you want.

   The following command assumes that you want to pull the latest container.

   ```
   docker pull nvcr.io/nvidia/tensorrt:24.02-py3
   ```
2. Open a command prompt and paste the `pull` command.

   Ensure that the pull process successfully completes before you proceed to step 3.
3. Run the container image.
   - If you have Docker 19.03 or later, a typical command to launch the container is:

     ```
     docker run --gpus all -it --rm -v local_dir:container_dir nvcr.io/nvidia/
     tensorrt:<xx.xx>-py<x>
     ```
   - If you have Docker 19.02 or earlier, a typical command to launch the container is:

```
nvidia-docker run -it --rm -v local_dir:container_dir nvcr.io/nvidia/
tensorrt:<xx.xx>-py<x>
```

4. To extend the TensorRT container, select one of the following options:
   - Add to or modify the source code in this container and run your customized version.
   - To add additional packages, use `docker build` to add your customizations on top of this container.

Note: NVIDIA recommends using the `docker build` option for ease of migration to later versions of the TensorRT container.

Chapter 4. TensorRT Release 24.07

The NVIDIA container image for TensorRT, release 24.07, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.6.3.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 22.04

  Note: Container image `24.07-py3` contains Python 3.10.
- NVIDIA CUDA 12.5.0.23
- NVIDIA cuBLAS 12.5.2.13

- NVIDIA cuDNN 9.1.0.70
- NVIDIA NCCL 2.21.5

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 39.0
- OpenMPI 4.1.4+
- OpenUCX 1.15.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.19
- Nsight Compute 2024.2.0.16
- Nsight Systems 2024.2.3.38

Driver Requirements

Release 24.07 is based on NVIDIA CUDA 12.5.0.23, which requires NVIDIA Driver release 545 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, and R520 drivers, which are not forward-compatible with CUDA 12.3. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 24.07 supports CUDA compute capability 6.0 and later. This corresponds to GPUs in the NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, NVIDIA Hopper™, and NVIDIA Ada Lovelace architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 24.07 is based on TensorRT 10.1.

  For a list of the features and enhancements that were introduced in this version of TensorRT, refer to the TensorRT release notes.
- All dependencies on cuDNN have been removed from the TensorRT starting with the 8.6.3 release to reduce the overall container size. Any TensorRT features which

depend on cuDNN, which are primarily some plugins and samples, will not work with this release.

- Latest version of Ubuntu 22.04 with October 2023 updates.

Announcements

- Starting with the 23.11 release, TensorRT containers supporting iGPU architectures are published, and run on Jetson devices. Please refer to the Frameworks Support Matrix for information regarding which iGPU hardware/software is supported by which container.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is no longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 24.07 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.

- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
24.07		NVIDIA CUDA 12.5.1	TensorRT 10.2.0.17
24.06	22.04	NVIDIA CUDA 12.5.0.23	TensorRT 10.1.0.27
24.05		NVIDIA CUDA 12.4.1	TensorRT 10.0.1.6
24.04			TensorRT 8.6.3
24.03		NVIDIA CUDA 12.4.0.41	
24.02		NVIDIA CUDA 12.3.2	
24.01			TensorRT 8.6.1.6
23.12			
23.11		NVIDIA CUDA 12.3.0	
23.10		NVIDIA CUDA 12.2.1	
23.09		NVIDIA CUDA 12.2.1	
23.08			
23.07		NVIDIA CUDA 12.1.1	
23.06			
23.05			TensorRT 8.6.1.2
23.04	20.04	NVIDIA CUDA 12.1.0	TensorRT 8.6.1
23.03			TensorRT 8.5.3
23.02		NVIDIA CUDA 12.0.1	
23.01			TensorRT 8.5.2.2
22.12		NVIDIA CUDA 11.8.0	TensorRT 8.5.1
22.11			
22.10			TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4

Container Version	Ubuntu	CUDA Toolkit	TensorRT
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03 20.02 20.01		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12 19.11			TensorRT 6.0.1

Container Version	Ubuntu	CUDA Toolkit	TensorRT
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

- The `onnx_graphsurgeon` Python module on ARM Server systems is not compatible with ONNX version 1.11.0, which is normally recommended for the included TensorRT release. You will instead need to use ONNX version 1.15.0 to resolve a possible segmentation fault.
- With r545 or r550 drivers, some models may run into "*Unspecified Launch Failure*" during engine building. This can be worked around by downgrading the driver version to r535.
- TensorRT's version compatibility feature has not been extensively tested and is therefore not supported with TensorRT 8.6.3. This TensorRT release is a special release that removes cuDNN as a dependency. Version compatibility between TensorRT 8.6.1 and future versions as documented will still be supported.
- Due to removing TensorRT's dependency on cuDNN the following networks may show performance regressions:
  - BasicUnet
  - DynUnet
  - HighResNet
  - StableDiffusion VAE-encoder
  - StableDiffusion VAE-decoder

Chapter 5. TensorRT Release 24.06

The NVIDIA container image for TensorRT, release 24.06, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.6.3.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 22.04

  Note: Container image `24.06-py3` contains Python 3.10.
- NVIDIA CUDA 12.5.0.23
- NVIDIA cuBLAS 12.5.2.13

- NVIDIA cuDNN 9.1.0.70
- NVIDIA NCCL 2.21.5

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 39.0
- OpenMPI 4.1.4+
- OpenUCX 1.15.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.19
- Nsight Compute 2024.2.0.16
- Nsight Systems 2024.2.3.38

Driver Requirements

Release 24.06 is based on NVIDIA CUDA 12.5.0.23, which requires NVIDIA Driver release 545 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, and R520 drivers, which are not forward-compatible with CUDA 12.3. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 24.06 supports CUDA compute capability 6.0 and later. This corresponds to GPUs in the NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, NVIDIA Hopper™, and NVIDIA Ada Lovelace architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 24.06 is based on TensorRT 10.1.

  For a list of the features and enhancements that were introduced in this version of TensorRT, refer to the TensorRT release notes.
- All dependencies on cuDNN have been removed from the TensorRT starting with the 8.6.3 release to reduce the overall container size. Any TensorRT features which

depend on cuDNN, which are primarily some plugins and samples, will not work with this release.

- Latest version of Ubuntu 22.04 with October 2023 updates.

Announcements

- Starting with the 23.11 release, TensorRT containers supporting iGPU architectures are published, and run on Jetson devices. Please refer to the Frameworks Support Matrix for information regarding which iGPU hardware/software is supported by which container.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is no longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 24.06 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.

- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
24.06	22.04	NVIDIA CUDA 12.5.0.23	TensorRT 10.1.0.27
24.05		NVIDIA CUDA 12.4.1	TensorRT 10.0.1.6
24.04			TensorRT 8.6.3
24.03		NVIDIA CUDA 12.4.0.41	
24.02		NVIDIA CUDA 12.3.2	
24.01			TensorRT 8.6.1.6
23.12			
23.11		NVIDIA CUDA 12.3.0	
23.10		NVIDIA CUDA 12.2.1	
23.09		NVIDIA CUDA 12.2.1	
23.08			
23.07		NVIDIA CUDA 12.1.1	
23.06			
23.05			TensorRT 8.6.1.2
23.04	20.04	NVIDIA CUDA 12.1.0	TensorRT 8.6.1
23.03			TensorRT 8.5.3
23.02		NVIDIA CUDA 12.0.1	
23.01			TensorRT 8.5.2.2
22.12		NVIDIA CUDA 11.8.0	TensorRT 8.5.1
22.11			
22.10			TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1

Container Version	Ubuntu	CUDA Toolkit	TensorRT
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	

Container Version	Ubuntu	CUDA Toolkit	TensorRT
19.09			
19.08			TensorRT 5.1.5

Known Issues

- The `onnx_graphsurgeon` Python module on ARM Server systems is not compatible with ONNX version 1.11.0, which is normally recommended for the included TensorRT release. You will instead need to use ONNX version 1.15.0 to resolve a possible segmentation fault.
- With r545 or r550 drivers, some models may run into "*Unspecified Launch Failure*" during engine building. This can be worked around by downgrading the driver version to r535.
- TensorRT's version compatibility feature has not been extensively tested and is therefore not supported with TensorRT 8.6.3. This TensorRT release is a special release that removes cuDNN as a dependency. Version compatibility between TensorRT 8.6.1 and future versions as documented will still be supported.
- Due to removing TensorRT's dependency on cuDNN the following networks may show performance regressions:
  - BasicUnet
  - DynUnet
  - HighResNet
  - StableDiffusion VAE-encoder
  - StableDiffusion VAE-decoder

Chapter 6. TensorRT Release 24.05

The NVIDIA container image for TensorRT, release 24.05, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.6.3.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 22.04

  Note: Container image `24.05-py3` contains Python 3.10.
- NVIDIA CUDA 12.4.1
- NVIDIA cuBLAS 12.4.5.8

- NVIDIA cuDNN 9.1.0.70
- NVIDIA NCCL 2.21.5

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 39.0
- OpenMPI 4.1.4+
- OpenUCX 1.15.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.19
- Nsight Compute 2024.1.1.4
- Nsight Systems 2024.2.1.106

Driver Requirements

Release 24.05 is based on NVIDIA CUDA 12.4.0.41, which requires NVIDIA Driver release 545 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, and R520 drivers, which are not forward-compatible with CUDA 12.3. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 24.05 supports CUDA compute capability 6.0 and later. This corresponds to GPUs in the NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, NVIDIA Hopper™, and NVIDIA Ada Lovelace architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 24.05 is based on TensorRT 8.6.3.1.

  For a list of the features and enhancements that were introduced in TensorRT 8.6, refer to the TensorRT release notes.

- All dependencies on cuDNN have been removed from the TensorRT 8.6.3 release to reduce the overall container size. Any TensorRT features which depend on cuDNN, which are primarily some plugins and samples, will not work with this release.
- Latest version of Ubuntu 22.04 with October 2023 updates.

Announcements

- Starting with the 23.11 release, TensorRT containers supporting iGPU architectures are published, and run on Jetson devices. Please refer to the Frameworks Support Matrix for information regarding which iGPU hardware/software is supported by which container.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 24.05 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

Use the TensorFlow Container to accelerate through TF-TRT instead.

- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
24.05	22.04	NVIDIA CUDA 12.4.1	TensorRT 10.0.1.6
24.04			TensorRT 8.6.3
24.03		NVIDIA CUDA 12.4.0.41	
24.02		NVIDIA CUDA 12.3.2	
24.01			TensorRT 8.6.1.6
23.12			
23.11		NVIDIA CUDA 12.3.0	
23.10		NVIDIA CUDA 12.2.1	
23.09		NVIDIA CUDA 12.2.1	
23.08			
23.07		NVIDIA CUDA 12.1.1	
23.06			
23.05			TensorRT 8.6.1.2
23.04	20.04	NVIDIA CUDA 12.1.0	TensorRT 8.6.1
23.03			TensorRT 8.5.3
23.02		NVIDIA CUDA 12.0.1	
23.01			TensorRT 8.5.2.2
22.12		NVIDIA CUDA 11.8.0	TensorRT 8.5.1
22.11			
22.10			TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5

Container Version	Ubuntu	CUDA Toolkit	TensorRT
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			

Container Version	Ubuntu	CUDA Toolkit	TensorRT
19.08			TensorRT 5.1.5

Known Issues

- The `onnx_graphsurgeon` Python module on ARM Server systems is not compatible with ONNX version 1.11.0, which is normally recommended for the included TensorRT release. You will instead need to use ONNX version 1.15.0 to resolve a possible segmentation fault.
- With r545 or r550 drivers, some models may run into "*Unspecified Launch Failure*" during engine building. This can be worked around by downgrading the driver version to r535.
- TensorRT's version compatibility feature has not been extensively tested and is therefore not supported with TensorRT 8.6.3. This TensorRT release is a special release that removes cuDNN as a dependency. Version compatibility between TensorRT 8.6.1 and future versions as documented will still be supported.
- Due to removing TensorRT's dependency on cuDNN the following networks may show performance regressions:
  - BasicUnet
  - DynUnet
  - HighResNet
  - StableDiffusion VAE-encoder
  - StableDiffusion VAE-decoder

Chapter 7. TensorRT Release 24.04

The NVIDIA container image for TensorRT, release 24.04, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.6.3.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 22.04

  Note: Container image `24.04-py3` contains Python 3.10.
- NVIDIA CUDA 12.4
- NVIDIA cuBLAS 12.4.5.8

- NVIDIA cuDNN 9.1.0.70
- NVIDIA NCCL 2.21.5

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 39.0
- OpenMPI 4.1.4+
- OpenUCX 1.15.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.18
- Nsight Compute 2024.1.1.4
- Nsight Systems 2024.2.1.106

Driver Requirements

Release 24.04 is based on NVIDIA CUDA 12.4.0.41, which requires NVIDIA Driver release 545 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, and R520 drivers, which are not forward-compatible with CUDA 12.3. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 24.04 supports CUDA compute capability 6.0 and later. This corresponds to GPUs in the NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, NVIDIA Hopper™, and NVIDIA Ada Lovelace architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 24.04 is based on TensorRT 8.6.3.1.

  For a list of the features and enhancements that were introduced in TensorRT 8.6, refer to the TensorRT release notes.

- All dependencies on cuDNN have been removed from the TensorRT 8.6.3 release to reduce the overall container size. Any TensorRT features which depend on cuDNN, which are primarily some plugins and samples, will not work with this release.
- Latest version of Ubuntu 22.04 with October 2023 updates.

Announcements

- Starting with the 23.11 release, TensorRT containers supporting iGPU architectures are published, and run on Jetson devices. Please refer to the Frameworks Support Matrix for information regarding which iGPU hardware/software is supported by which container.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 24.04 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

Use the TensorFlow Container to accelerate through TF-TRT instead.

- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
24.04	22.04	NVIDIA CUDA 12.4	TensorRT 8.6.3
24.03		NVIDIA CUDA 12.4.0.41	
24.02		NVIDIA CUDA 12.3.2	
24.01			TensorRT 8.6.1.6
23.12			
23.11		NVIDIA CUDA 12.3.0	
23.10		NVIDIA CUDA 12.2.1	
23.09		NVIDIA CUDA 12.2.1	
23.08			
23.07		NVIDIA CUDA 12.1.1	
23.06			
23.05			TensorRT 8.6.1.2
23.04	20.04	NVIDIA CUDA 12.1.0	TensorRT 8.6.1
23.03			TensorRT 8.5.3
23.02		NVIDIA CUDA 12.0.1	
23.01			TensorRT 8.5.2.2
22.12		NVIDIA CUDA 11.8.0	TensorRT 8.5.1
22.11			
22.10			TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	

Container Version	Ubuntu	CUDA Toolkit	TensorRT
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

- The `onnx_graphsurgeon` Python module on ARM Server systems is not compatible with ONNX version 1.11.0, which is normally recommended for the included TensorRT release. You will instead need to use ONNX version 1.15.0 to resolve a possible segmentation fault.
- With r545 or r550 drivers, some models may run into "*Unspecified Launch Failure*" during engine building. This can be worked around by downgrading the driver version to r535.
- TensorRT's version compatibility feature has not been extensively tested and is therefore not supported with TensorRT 8.6.3. This TensorRT release is a special release that removes cuDNN as a dependency. Version compatibility between TensorRT 8.6.1 and future versions as documented will still be supported.
- Due to removing TensorRT's dependency on cuDNN the following networks may show performance regressions:
  - BasicUnet
  - DynUnet
  - HighResNet
  - StableDiffusion VAE-encoder
  - StableDiffusion VAE-decoder

Chapter 8. TensorRT Release 24.03

The NVIDIA container image for TensorRT, release 24.03, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.6.3.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 22.04

  Note: Container image `24.03-py3` contains Python 3.10.
- NVIDIA CUDA 12.4.0.41
- NVIDIA cuBLAS 12.4.2.65

- NVIDIA cuDNN 9.0.0.306
- NVIDIA NCCL 2.20

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 39.0
- OpenMPI 4.1.4+
- OpenUCX 1.15.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.18
- Nsight Compute 2024.1.0.13
- Nsight Systems 2024.2.1.38

Driver Requirements

Release 24.03 is based on NVIDIA CUDA 12.4.0.41, which requires NVIDIA Driver release 545 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, and R520 drivers, which are not forward-compatible with CUDA 12.3. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 24.03 supports CUDA compute capability 6.0 and later. This corresponds to GPUs in the NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, NVIDIA Hopper™, and NVIDIA Ada Lovelace architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 24.03 is based on TensorRT 8.6.3.1.

  For a list of the features and enhancements that were introduced in TensorRT 8.6, refer to the TensorRT 8.6 release notes.

- All dependencies on cuDNN have been removed from the TensorRT 8.6.3 release to reduce the overall container size. Any TensorRT features which depend on cuDNN, which are primarily some plugins and samples, will not work with this release.
- Latest version of Ubuntu 22.04 with October 2023 updates.

Announcements

- Starting with the 23.11 release, TensorRT containers supporting iGPU architectures are published, and run on Jetson devices. Please refer to the Frameworks Support Matrix for information regarding which iGPU hardware/software is supported by which container.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 24.03 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

Use the TensorFlow Container to accelerate through TF-TRT instead.

- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
24.03	22.04	NVIDIA CUDA 12.4.0.41	TensorRT 8.6.3
24.02		NVIDIA CUDA 12.3.2	
24.01			TensorRT 8.6.1.6
23.12			
23.11		NVIDIA CUDA 12.3.0	
23.10		NVIDIA CUDA 12.2.1	
23.09		NVIDIA CUDA 12.2.1	
23.08			
23.07		NVIDIA CUDA 12.1.1	
23.06			
23.05			TensorRT 8.6.1.2
23.04	20.04	NVIDIA CUDA 12.1.0	TensorRT 8.6.1
23.03			TensorRT 8.5.3
23.02		NVIDIA CUDA 12.0.1	
23.01			TensorRT 8.5.2.2
22.12		NVIDIA CUDA 11.8.0	TensorRT 8.5.1
22.11			
22.10			TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2

Container Version	Ubuntu	CUDA Toolkit	TensorRT
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03 20.02 20.01		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

- The `onnx_graphsurgeon` Python module on ARM Server systems is not compatible with ONNX version 1.11.0, which is normally recommended for the included TensorRT release. You will instead need to use ONNX version 1.15.0 to resolve a possible segmentation fault.
- With r545 or r550 drivers, some models may run into "*Unspecified Launch Failure*" during engine building. This can be worked around by downgrading the driver version to r535.
- TensorRT's version compatibility feature has not been extensively tested and is therefore not supported with TensorRT 8.6.3. This TensorRT release is a special release that removes cuDNN as a dependency. Version compatibility between TensorRT 8.6.1 and future versions as documented will still be supported.
- Due to removing TensorRT's dependency on cuDNN the following networks may show performance regressions:
  - BasicUnet
  - DynUnet
  - HighResNet
  - StableDiffusion VAE-encoder
  - StableDiffusion VAE-decoder

Chapter 9. TensorRT Release 24.02

The NVIDIA container image for TensorRT, release 24.02, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.6.3.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 22.04

  Note: Container image `24.01-py3` contains Python 3.10.
- NVIDIA CUDA® 12.3.2
- NVIDIA cuBLAS 12.3.4.1

- NVIDIA cuDNN 9.0.0.306
- NVIDIA NCCL 2.19.4

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 39.0
- OpenMPI 4.1.4+
- OpenUCX 1.15.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.16rc4
- Nsight Compute 2023.3.1.1
- Nsight Systems 2023.4.1.97

Driver Requirements

Release 24.02 is based on CUDA 12.3.2, which requires NVIDIA Driver release 545 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, and R520 drivers, which are not forward-compatible with CUDA 12.3. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 24.02 supports CUDA compute capability 6.0 and later. This corresponds to GPUs in the NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, NVIDIA Hopper™, and NVIDIA Ada Lovelace architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 24.02 is based on TensorRT 8.6.3.1.

  For a list of the features and enhancements that were introduced in TensorRT 8.6, refer to the TensorRT 8.6 release notes.

- All dependencies on cuDNN have been removed from the TensorRT 8.6.3 release to reduce the overall container size. Any TensorRT features which depend on cuDNN, which are primarily some plugins and samples, will not work with this release.
- Latest version of Ubuntu 22.04 with October 2023 updates.

Announcements

- Starting with the 23.11 release, TensorRT containers supporting iGPU architectures are published, and run on Jetson devices. Please refer to the Frameworks Support Matrix for information regarding which iGPU hardware/software is supported by which container.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 24.01 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

Use the TensorFlow Container to accelerate through TF-TRT instead.

- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
24.02	22.04	NVIDIA CUDA 12.3.2	TensorRT 8.6.3
24.01		NVIDIA CUDA 12.3.2	TensorRT 8.6.1.6
23.12			
23.11		NVIDIA CUDA 12.3.0	
23.10		NVIDIA CUDA 12.2.1	
23.09		NVIDIA CUDA 12.2.1	
23.08			
23.07		NVIDIA CUDA 12.1.1	
23.06			
23.05			TensorRT 8.6.1.2
23.04	20.04	NVIDIA CUDA 12.1.0	TensorRT 8.6.1
23.03			TensorRT 8.5.3
23.02		NVIDIA CUDA 12.0.1	
23.01			TensorRT 8.5.2.2
22.12		NVIDIA CUDA 11.8.0	TensorRT 8.5.1
22.11			
22.10			TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3

Container Version	Ubuntu	CUDA Toolkit	TensorRT
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03 20.02 20.01		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

- The `onnx_graphsurgeon` Python module on ARM Server systems is not compatible with ONNX version 1.11.0, which is normally recommended for the included TensorRT release. You will instead need to use ONNX version 1.15.0 to resolve a possible segmentation fault.
- With r545 or r550 drivers, some models may run into "*Unspecified Launch Failure*" during engine building. This can be worked around by downgrading the driver version to r535.
- TensorRT's version compatibility feature has not been extensively tested and is therefore not supported with TensorRT 8.6.3. This TensorRT release is a special release that removes cuDNN as a dependency. Version compatibility between TensorRT 8.6.1 and future versions as documented will still be supported.
- Due to removing TensorRT's dependency on cuDNN the following networks may show performance regressions:
  - BasicUnet
  - DynUnet
  - HighResNet
  - StableDiffusion VAE-encoder
  - StableDiffusion VAE-decoder

Chapter 10. TensorRT Release 24.01

The NVIDIA container image for TensorRT, release 24.01, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.6.1.6.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 22.04

  Note: Container image `24.01-py3` contains Python 3.10.
- NVIDIA CUDA® 12.3.2
- NVIDIA cuBLAS 12.3.4.1

- NVIDIA cuDNN 8.9.7.29
- NVIDIA NCCL 2.19.4

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 39.0
- OpenMPI 4.1.4+
- OpenUCX 1.15.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.16rc4
- Nsight Compute 2023.3.1.1
- Nsight Systems 2023.3.4.1.97

Driver Requirements

Release 24.01 is based on CUDA 12.3.2, which requires NVIDIA Driver release 545 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, and R520 drivers, which are not forward-compatible with CUDA 12.3. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 24.01 supports CUDA compute capability 6.0 and later. This corresponds to GPUs in the NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, NVIDIA Hopper™, and NVIDIA Ada Lovelace architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 24.01 is based on TensorRT 8.6.1.6.

  For a list of the features and enhancements that were introduced in TensorRT 8.6.1, refer to the TensorRT 8.6 release notes.
- Ubuntu 22.04 with October 2023 updates.

Announcements

- Starting with the 23.11 release, TensorRT containers supporting iGPU architectures are published, and run on Jetson devices. Please refer to the Frameworks Support Matrix for information regarding which iGPU hardware/software is supported by which container.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 24.01 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
24.01	22.04	NVIDIA CUDA 12.3.2	TensorRT 8.6.1.6
23.12			
23.11		NVIDIA CUDA 12.3.0	
23.10		NVIDIA CUDA 12.2.1	
23.09		NVIDIA CUDA 12.2.1	
23.08			
23.07		NVIDIA CUDA 12.1.1	
23.06			
23.05			TensorRT 8.6.1.2
23.04	20.04	NVIDIA CUDA 12.1.0	TensorRT 8.6.1
23.03			TensorRT 8.5.3
23.02		NVIDIA CUDA 12.0.1	
23.01			TensorRT 8.5.2.2
22.12		NVIDIA CUDA 11.8.0	TensorRT 8.5.1
22.11			
22.10			TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	TensorRT 8.0.2.2 for Arm SBSA Linux
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 11. TensorRT Release 23.12

The NVIDIA container image for TensorRT, release 23.12, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.6.1.6.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 22.04

  Note: Container image `23.12-py3` contains Python 3.10.
- NVIDIA CUDA® 12.3.2
- NVIDIA cuBLAS 12.3.4.1

- NVIDIA cuDNN 8.9.7.29
- NVIDIA NCCL 2.19.3

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 39.0
- OpenMPI 4.1.4+
- OpenUCX 1.15.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.16
- Nsight Compute 2023.3.1.1
- Nsight Systems 2023.3.4.1

Driver Requirements

Release 23.12 is based on CUDA 12.3.2, which requires NVIDIA Driver release 545 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 450.51 (or later R450), 470.57 (or later R470), 510.47 (or later R510), or 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, R460, and R520 drivers, which are not forward-compatible with CUDA 12.3. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 23.12 supports CUDA compute capability 6.0 and later. This corresponds to GPUs in the NVIDIA Kepler, Maxwell, NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, NVIDIA Hopper™, and NVIDIA Ada Lovelace architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 23.12 is based on TensorRT 8.6.1.6.

  For a list of the features and enhancements that were introduced in TensorRT 8.6.1, refer to the TensorRT 8.6 release notes.
- Ubuntu 22.04 with October 2023 updates.

Announcements

- Starting with the 23.11 release, TensorRT containers supporting iGPU architectures are published, and run on Jetson devices. Please refer to the Frameworks Support Matrix for information regarding which iGPU hardware/software is supported by which container.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 23.12 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
23.12	22.04	NVIDIA CUDA 12.3.2	TensorRT 8.6.1.6
23.11		NVIDIA CUDA 12.3.0	
23.10		NVIDIA CUDA 12.2.1	
23.09		NVIDIA CUDA 12.2.1	
23.08			
23.07		NVIDIA CUDA 12.1.1	
23.06			
23.05			TensorRT 8.6.1.2
23.04	20.04	NVIDIA CUDA 12.1.0	TensorRT 8.6.1
23.03			TensorRT 8.5.3
23.02		NVIDIA CUDA 12.0.1	
23.01			TensorRT 8.5.2.2
22.12		NVIDIA CUDA 11.8.0	TensorRT 8.5.1
22.11			
22.10			TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	

Container Version	Ubuntu	CUDA Toolkit	TensorRT
			TensorRT 8.0.2.2 for Arm SBSA Linux
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 12. TensorRT Release 23.11

The NVIDIA container image for TensorRT, release 23.11, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.6.1.6.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 22.04

  Note: Container image `23.11-py3` contains Python 3.10.
- NVIDIA CUDA® 12.3.0
- NVIDIA cuBLAS 12.3.2.1

- NVIDIA cuDNN 8.9.6
- NVIDIA NCCL 2.19.3

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 39.0
- OpenMPI 4.1.4+
- OpenUCX 1.15.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.16
- Nsight Compute 2023.3.0.12
- Nsight Systems 2023.3.1.92

Driver Requirements

Release 23.11 is based on CUDA 12.3.0, which requires NVIDIA Driver release 545 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 450.51 (or later R450), 470.57 (or later R470), 510.47 (or later R510), or 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, R460, and R520 drivers, which are not forward-compatible with CUDA 12.3. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 23.11 supports CUDA compute capability 6.0 and later. This corresponds to GPUs in the NVIDIA Kepler, Maxwell, NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, NVIDIA Hopper™, and NVIDIA Ada Lovelace architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 23.11 is based on TensorRT 8.6.1.6.

  For a list of the features and enhancements that were introduced in TensorRT 8.6.1, refer to the TensorRT 8.6 release notes.
- Ubuntu 22.04 with October 2023 updates.

Announcements

- Starting with the 23.11 release, TensorRT containers supporting iGPU architectures are published, and run on Jetson devices. Please refer to the Frameworks Support Matrix for information regarding which iGPU hardware/software is supported by which container.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 23.11 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
23.11	22.04	NVIDIA CUDA 12.3.0	TensorRT 8.6.1.6
23.10		NVIDIA CUDA 12.2.1	
23.09		NVIDIA CUDA 12.2.1	
23.08			
23.07		NVIDIA CUDA 12.1.1	
23.06			
23.05			TensorRT 8.6.1.2
23.04	20.04	NVIDIA CUDA 12.1.0	TensorRT 8.6.1
23.03			TensorRT 8.5.3
23.02		NVIDIA CUDA 12.0.1	
23.01			TensorRT 8.5.2.2
22.12		NVIDIA CUDA 11.8.0	TensorRT 8.5.1
22.11			
22.10			TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 13. TensorRT Release 23.10

The NVIDIA container image for TensorRT, release 23.10, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.6.1.6.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 22.04

  Note: Container image `23.10-py3` contains Python 3.10.
- NVIDIA CUDA® 12.2.2
- NVIDIA cuBLAS 12.2.5.6

- NVIDIA cuDNN 8.9.5
- NVIDIA NCCL 2.19.3

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 39.0
- OpenMPI 4.1.4+
- OpenUCX 1.15.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.16
- Nsight Compute 2023.2.1.3
- Nsight Systems 2023.3.1.92

Driver Requirements

Release 23.10 is based on CUDA 12.2.2, which requires NVIDIA Driver release 535 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 450.51 (or later R450), 470.57 (or later R470), 510.47 (or later R510), or 525.85 (or later R525), or 535.86 (or later R535).

The CUDA driver's compatibility package only supports particular drivers. Thus, users should upgrade from all R418, R440, R460, and R520 drivers, which are not forward-compatible with CUDA 12.2. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 23.10 supports CUDA compute capability 6.0 and later. This corresponds to GPUs in the NVIDIA Kepler, Maxwell, NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, NVIDIA Hopper™, and NVIDIA Ada Lovelace architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 23.10 is based on TensorRT 8.6.1.6.

  For a list of the features and enhancements that were introduced in TensorRT 8.6.1, refer to the TensorRT 8.6 release notes.
- Ubuntu 22.04 with September 2023 updates.

Announcements

- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 22.05 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
23.09	22.04	NVIDIA CUDA 12.2.1	TensorRT 8.6.1.6
23.08			
23.07		NVIDIA CUDA 12.1.1	
23.06			
23.05			TensorRT 8.6.1.2
23.04	20.04	NVIDIA CUDA 12.1.0	TensorRT 8.6.1
23.03			TensorRT 8.5.3
23.02		NVIDIA CUDA 12.0.1	
23.01			TensorRT 8.5.2.2
22.12		NVIDIA CUDA 11.8.0	TensorRT 8.5.1
22.11			
22.10			TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 14. TensorRT Release 23.09

The NVIDIA container image for TensorRT, release 23.09, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.6.1.6.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 22.04

  Note: Container image `23.09-py3` contains Python 3.10.
- NVIDIA CUDA® 12.2.1
- NVIDIA cuBLAS 12.2.5.6

- NVIDIA cuDNN 8.9.5
- NVIDIA NCCL 2.18.5

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 39.0
- OpenMPI 4.1.4+
- OpenUCX 1.15.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.16
- Nsight Compute 2023.2.1.3
- Nsight Systems 2023.3.1.92

Driver Requirements

Release 23.09 is based on CUDA 12.2.1, which requires NVIDIA Driver release 535 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 450.51 (or later R450), 470.57 (or later R470), 510.47 (or later R510), or 525.85 (or later R525), or 535.86 (or later R535).

The CUDA driver's compatibility package only supports particular drivers. Thus, users should upgrade from all R418, R440, R460, and R520 drivers, which are not forward-compatible with CUDA 12.2. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 23.09 supports CUDA compute capability 6.0 and later. This corresponds to GPUs in the NVIDIA Kepler, Maxwell, NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, NVIDIA Hopper™, and NVIDIA Ada Lovelace architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 23.09 is based on TensorRT 8.6.1.6.

  For a list of the features and enhancements that were introduced in TensorRT 8.6.1, refer to the TensorRT 8.6 release notes.
- Ubuntu 22.04 with August 2023 updates.

Announcements

- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 22.05 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
23.09	22.04	NVIDIA CUDA 12.2.1	TensorRT 8.6.1.6
23.08			
23.07		NVIDIA CUDA 12.1.1	
23.06			
23.05			TensorRT 8.6.1.2
23.04	20.04	NVIDIA CUDA 12.1.0	TensorRT 8.6.1
23.03			TensorRT 8.5.3
23.02		NVIDIA CUDA 12.0.1	
23.01			TensorRT 8.5.2.2
22.12		NVIDIA CUDA 11.8.0	TensorRT 8.5.1
22.11			
22.10			TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 15. TensorRT Release 23.08

The NVIDIA container image for TensorRT, release 23.08, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.6.1.6.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 22.04

  Note: Container image `23.08-py3` contains Python 3.10.
- NVIDIA CUDA® 12.2.1
- NVIDIA cuBLAS 12.2.5.1

- NVIDIA cuDNN 8.9.4
- NVIDIA NCCL 2.18.3

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 39.0
- OpenMPI 4.1.4+
- OpenUCX 1.15.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.15
- Nsight Compute 2023.2.1.3
- Nsight Systems 2023.2.3.1001

Driver Requirements

Release 23.08 is based on CUDA 12.2.1, which requires NVIDIA Driver release 535 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 450.51 (or later R450), 470.57 (or later R470), 510.47 (or later R510), or 525.85 (or later R525), or 535.86 (or later R535).

The CUDA driver's compatibility package only supports particular drivers. Thus, users should upgrade from all R418, R440, R460, and R520 drivers, which are not forward-compatible with CUDA 12.2. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 23.08 supports CUDA compute capability 6.0 and later. This corresponds to GPUs in the NVIDIA Kepler, Maxwell, NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, NVIDIA Hopper™, and NVIDIA Ada Lovelace architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 23.08 is based on TensorRT 8.6.1.6.

  For a list of the features and enhancements that were introduced in TensorRT 8.6.1, refer to the TensorRT 8.6 release notes.
- Ubuntu 22.04 with July 2023 updates.

Announcements

- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 22.05 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
23.08	22.04	NVIDIA CUDA 12.2.1	TensorRT 8.6.1.6
23.07		NVIDIA CUDA 12.1.1	
23.06			
23.05			TensorRT 8.6.1.2
23.04	20.04	NVIDIA CUDA 12.1.0	TensorRT 8.6.1
23.03			TensorRT 8.5.3
23.02		NVIDIA CUDA 12.0.1	
23.01			TensorRT 8.5.2.2
22.12		NVIDIA CUDA 11.8.0	TensorRT 8.5.1
22.11			
22.10			TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 16. TensorRT Release 23.07

The NVIDIA container image for TensorRT, release 23.07, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.6.1.6.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 22.04

  Note: Container image `23.07-py3` contains Python 3.10.
- NVIDIA CUDA® 12.1.1
- NVIDIA cuBLAS 12.1.3.1

- NVIDIA cuDNN 8.9.3
- NVIDIA NCCL 2.18.3

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 39.0
- OpenMPI 4.1.4+
- OpenUCX 1.15.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.15
- Nsight Compute 2023.1.1.4
- Nsight Systems 2023.2.3.1001

Driver Requirements

Release 23.07 is based on CUDA 12.1.1, which requires NVIDIA Driver release 530 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 450.51 (or later R450), 470.57 (or later R470), 510.47 (or later R510), 515.65 (or later R515), 525.85 (or later R525), or 530.30 (or later R530).

The CUDA driver's compatibility package only supports particular drivers. Thus, users should upgrade from all R418, R440, R460, and R520 drivers, which are not forward-compatible with CUDA 12.1. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 23.07 supports CUDA compute capability 3.5 and later. This corresponds to GPUs in the NVIDIA Kepler, Maxwell, NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, NVIDIA Hopper™, and NVIDIA Ada Lovelace architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 23.07 is based on TensorRT 8.6.1.6.

  For a list of the features and enhancements that were introduced in TensorRT 8.6.1, refer to the TensorRT 8.6 release notes.
- Ubuntu 22.04 with June 2023 updates.

Announcements

- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 22.05 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
23.07	22.04	NVIDIA CUDA 12.1.1	TensorRT 8.6.1.6
23.06			
23.05			TensorRT 8.6.1.2
23.04	20.04	NVIDIA CUDA 12.1.0	TensorRT 8.6.1
23.03			TensorRT 8.5.3
23.02		NVIDIA CUDA 12.0.1	
23.01			TensorRT 8.5.2.2
22.12		NVIDIA CUDA 11.8.0	TensorRT 8.5.1
22.11			
22.10			TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03 20.02 20.01		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 17. TensorRT Release 23.06

The NVIDIA container image for TensorRT, release 23.06, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.6.1.6.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 22.04

  Note: Container image `23.06-py3` contains Python 3.10.
- NVIDIA CUDA® 12.1.1
- NVIDIA cuBLAS 12.1.3.1

- NVIDIA cuDNN 8.9.2
- NVIDIA NCCL 2.18.1

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 39.0
- OpenMPI 4.1.4+
- OpenUCX 1.15.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.15
- Nsight Compute 2023.1.1.4
- Nsight Systems 2023.2.3.1001

Driver Requirements

Release 23.06 is based on CUDA 12.1.1, which requires NVIDIA Driver release 530 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 450.51 (or later R450), 470.57 (or later R470), 510.47 (or later R510), 515.65 (or later R515), 525.85 (or later R525), or 530.30 (or later R530).

The CUDA driver's compatibility package only supports particular drivers. Thus, users should upgrade from all R418, R440, R460, and R520 drivers, which are not forward-compatible with CUDA 12.1. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 23.06 supports CUDA compute capability 3.5 and later. This corresponds to GPUs in the NVIDIA Kepler, Maxwell, NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, NVIDIA Hopper™, and NVIDIA Ada Lovelace architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 23.06 is based on TensorRT 8.6.1.6.

  For a list of the features and enhancements that were introduced in TensorRT 8.6.1, refer to the TensorRT 8.6 release notes.
- Ubuntu 22.04 with May 2023 updates.

Announcements

- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 22.05 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
23.06	22.04	NVIDIA CUDA 12.1.1	TensorRT 8.6.1.6
23.05			TensorRT 8.6.1.2
23.04	20.04	NVIDIA CUDA 12.1.0	TensorRT 8.6.1
23.03			TensorRT 8.5.3
23.02		NVIDIA CUDA 12.0.1	
23.01			TensorRT 8.5.2.2
22.12		NVIDIA CUDA 11.8.0	TensorRT 8.5.1
22.11			
22.10			TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03 20.02 20.01		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 18. TensorRT Release 23.05

The NVIDIA container image for TensorRT, release 23.05, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.6.1.2.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 22.04

  Note: Container image `23.05-py3` contains Python 3.10.
- NVIDIA CUDA® 12.1.1
- NVIDIA cuBLAS 12.1.3.1

- NVIDIA cuDNN 8.9.1.23
- NVIDIA NCCL 2.18.1

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- OpenMPI 4.1.4+
- OpenUCX 1.14.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.14
- Nsight Compute 2023.1.1.4
- Nsight Systems 2023.2

Driver Requirements

Release 23.05 is based on CUDA 12.1.1, which requires NVIDIA Driver release 530 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 450.51 (or later R450), 470.57 (or later R470), 510.47 (or later R510), 515.65 (or later R515), 525.85 (or later R525), or 530.30 (or later R530).

The CUDA driver's compatibility package only supports particular drivers. Thus, users should upgrade from all R418, R440, R460, and R520 drivers, which are not forward-compatible with CUDA 12.1. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 23.05 supports CUDA compute capability 3.5 and later. This corresponds to GPUs in the NVIDIA Kepler, Maxwell, NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, NVIDIA Hopper™, and NVIDIA Ada Lovelace architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 23.05 is based on TensorRT 8.6.1.2.

  For a list of the features and enhancements that were introduced in TensorRT 8.6.1, refer to the TensorRT 8.6 release notes.
- Ubuntu 22.04 with April 2023 updates.

Announcements

- Starting with the 22.05 release, the TensorRT container is available for the Arm SBSA platform.

  For example, when you pull the `nvcr.io/nvidia/tensorrt:22.05-py3` Docker image on an Arm SBSA machine, the Arm-specific image is automatically fetched.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 22.05 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
23.05	22.04	NVIDIA CUDA 12.1.1	TensorRT 8.6.1.2
23.04	20.04	NVIDIA CUDA 12.1.0	TensorRT 8.6.1
23.03			TensorRT 8.5.3
23.02		NVIDIA CUDA 12.0.1	
23.01			TensorRT 8.5.2.2
22.12		NVIDIA CUDA 11.8.0	TensorRT 8.5.1
22.11			
22.10			TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 19. TensorRT Release 23.04

The NVIDIA container image for TensorRT, release 23.04, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.6.1.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 20.04

  Note: Container image `23.04-py3` contains Python 3.8.
- NVIDIA CUDA® 12.1.0
- NVIDIA cuBLAS 12.1.3

- NVIDIA cuDNN 8.9.0
- NVIDIA NCCL 2.17.1

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- OpenMPI 4.1.4+
- OpenUCX 1.14.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.13
- Nsight Compute 2023.1.0.15
- Nsight Systems 2023.1.1.127

Driver Requirements

Release 23.04 is based on CUDA 12.1.0, which requires NVIDIA Driver release 530 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 450.51 (or later R450), 470.57 (or later R470), 510.47 (or later R510), 515.65 (or later R515), 525.85 (or later R525), or 530.30 (or later R530).

The CUDA driver's compatibility package only supports particular drivers. Thus, users should upgrade from all R418, R440, R460, and R520 drivers, which are not forward-compatible with CUDA 12.1. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 23.04 supports CUDA compute capability 3.5 and later. This corresponds to GPUs in the NVIDIA Kepler, Maxwell, NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, NVIDIA Hopper™, and NVIDIA Ada Lovelace architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 23.04 is based on TensorRT 8.6.1.

  For a list of the features and enhancements that were introduced in TensorRT 8.6.1, refer to the TensorRT 8.6 release notes.
- Ubuntu 20.04 with March 2023 updates.

Announcements

- Starting with the 22.05 release, the TensorRT container is available for the Arm SBSA platform.

  For example, when you pull the `nvcr.io/nvidia/tensorrt:22.05-py3` Docker image on an Arm SBSA machine, the Arm-specific image is automatically fetched.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 22.05 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
23.04	20.04	NVIDIA CUDA 12.1.0	TensorRT 8.6.1
23.03			TensorRT 8.5.3
23.02		NVIDIA CUDA 12.0.1	
23.01			TensorRT 8.5.2.2
22.12		NVIDIA CUDA 11.8.0	TensorRT 8.5.1
22.11			
22.10			TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.02	18.04	NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11		NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 20. TensorRT Release 23.03

The NVIDIA container image for TensorRT, release 23.03, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.5.3.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 20.04

  Note: Container image `23.03-py3` contains Python 3.8.
- NVIDIA CUDA® 12.1.0
- NVIDIA cuBLAS from CUDA 12.1.0

- NVIDIA cuDNN 8.8.1.3
- NVIDIA NCCL 2.17.1

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- OpenMPI 4.1.4+
- OpenUCX 1.14.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.13
- Nsight Compute 2023.1.0.15
- Nsight Systems 2023.1.1.127

Driver Requirements

Release 23.03 is based on CUDA 12.1.0, which requires NVIDIA Driver release 530 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 450.51 (or later R450), 470.57 (or later R470), 510.47 (or later R510), 515.65 (or later R515), 525.85 (or later R525), or 530.30 (or later R530).

The CUDA driver's compatibility package only supports particular drivers. Thus, users should upgrade from all R418, R440, R460, and R520 drivers, which are not forward-compatible with CUDA 12.1. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 23.03 supports CUDA compute capability 3.5 and later. This corresponds to GPUs in the NVIDIA Kepler, Maxwell, NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, and NVIDIA Hopper™ architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 23.03 is based on TensorRT 8.5.3.

  For a list of the features and enhancements that were introduced in TensorRT 8.5.3, refer to the TensorRT 8.5 release notes.
- Ubuntu 20.04 with February 2023 updates.

Announcements

- Starting with the 22.05 release, the TensorRT container is available for the Arm SBSA platform.

  For example, when you pull the `nvcr.io/nvidia/tensorrt:22.05-py3` Docker image on an Arm SBSA machine, the Arm-specific image is automatically fetched.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 22.05 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
23.03	20.04	NVIDIA CUDA 12.1.0	TensorRT 8.5.3
23.02		NVIDIA CUDA 12.0.1	
23.01			TensorRT 8.5.2.2
22.12		NVIDIA CUDA 11.8.0	TensorRT 8.5.1
22.11			
22.10			TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03 20.02 20.01		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 21. TensorRT Release 23.02

The NVIDIA container image for TensorRT, release 23.02, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.5.3.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 20.04

  Note: Container image `23.02-py3` contains Python 3.8.
- NVIDIA CUDA® 12.0.1
- NVIDIA cuBLAS from CUDA 12.0.1

- NVIDIA cuDNN 8.7.0
- NVIDIA NCCL 2.16.5

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- OpenMPI 4.1.4+
- OpenUCX 1.14.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.13
- Nsight Compute 2022.4.1.6
- Nsight Systems 2022.5.1.93

Driver Requirements

Release 23.02 is based on CUDA 12.0.1, which requires NVIDIA Driver release 525 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 450.51 (or later R450), 470.57 (or later R470), 510.47 (or later R510), 515.65 (or later R515), or 525.85 (or later R525).

The CUDA driver's compatibility package only supports particular drivers. Thus, users should upgrade from all R418, R440, R460, and R520 drivers, which are not forward-compatible with CUDA 12.0. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 23.02 supports CUDA compute capability 3.5 and later. This corresponds to GPUs in the NVIDIA Kepler, Maxwell, NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, and NVIDIA Hopper™ architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 23.02 is based on TensorRT 8.5.3.

  For a list of the features and enhancements that were introduced in TensorRT 8.5.3, refer to the TensorRT 8.5 release notes.
- Ubuntu 20.04 with January 2023 updates.

Announcements

- Starting with the 22.05 release, the TensorRT container is available for the Arm SBSA platform.

  For example, when you pull the `nvcr.io/nvidia/tensorrt:22.05-py3` Docker image on an Arm SBSA machine, the Arm-specific image is automatically fetched.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 22.05 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

Note: Since the 22.09 release is based on an early access version of TensorRT 8.5, which is not accompanied by the publication of a corresponding TensorRT Open Source Software (OSS) release to GitHub, please specify building from the *main* branch in `install_opensource.sh` until the TensorRT OSS 8.5.1 release is posted.

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
23.02	20.04	NVIDIA CUDA 12.0.1	TensorRT 8.5.3
23.01			TensorRT 8.5.2.2
22.12		NVIDIA CUDA 11.8.0	TensorRT 8.5.1
22.11			
22.10			TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 22. TensorRT Release 23.01

The NVIDIA container image for TensorRT, release 23.01, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.5.2.2.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 20.04

  Note: Container image `23.01-py3` contains Python 3.8.
- NVIDIA CUDA® 12.0.1
- NVIDIA cuBLAS from CUDA 12.0.1

- NVIDIA cuDNN 8.7.0
- NVIDIA NCCL 2.16.5

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- OpenMPI 4.1.4+
- OpenUCX 1.14.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.13
- Nsight Compute 2022.4.1.6
- Nsight Systems 2022.5.1

Driver Requirements

Release 23.01 is based on CUDA 12.0.1, which requires NVIDIA Driver release 525 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 450.51 (or later R450), 470.57 (or later R470), 510.47 (or later R510), 515.65 (or later R515), or 525.85 (or later R525).

The CUDA driver's compatibility package only supports particular drivers. Thus, users should upgrade from all R418, R440, and R460 drivers, which are not forward-compatible with CUDA 12.0. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 23.01 supports CUDA compute capability 3.5 and later. This corresponds to GPUs in the NVIDIA Kepler, Maxwell, NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, and NVIDIA Hopper™ architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 23.01 is based on TensorRT 8.5.2.2.

  For a list of the features and enhancements that were introduced in TensorRT 8.5.2, refer to the TensorRT 8.5 release notes.
- Ubuntu 20.04 with December 2022 updates.

Announcements

- Starting with the 22.05 release, the TensorRT container is available for the Arm SBSA platform.

  For example, when you pull the `nvcr.io/nvidia/tensorrt:22.05-py3` Docker image on an Arm SBSA machine, the Arm-specific image is automatically fetched.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 22.05 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

Note: Since the 22.09 release is based on an early access version of TensorRT 8.5, which is not accompanied by the publication of a corresponding TensorRT Open Source Software (OSS) release to GitHub, please specify building from the *main* branch in `install_opensource.sh` until the TensorRT OSS 8.5.1 release is posted.

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
23.01	20.04	NVIDIA CUDA 12.0.1	TensorRT 8.5.2.2
22.12		NVIDIA CUDA 11.8.0	TensorRT 8.5.1
22.11			
22.10			TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 23. TensorRT Release 22.12

The NVIDIA container image for TensorRT, release 22.12, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.5.1.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 20.04

  Note: Container image `22.12-py3` contains Python 3.8.
- NVIDIA CUDA® 11.8.0
- NVIDIA cuBLAS 11.11.3.6

- NVIDIA cuDNN 8.7.0
- NVIDIA NCCL 2.15.5

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- OpenMPI 4.1.4+
- OpenUCX 1.14.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.13
- Nsight Compute 2022.3.0.22
- Nsight Systems 2022.4.2.1

Driver Requirements

Release 22.12 is based on CUDA 11.8.0, which requires NVIDIA Driver release 520 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 450.51 (or later R450), 470.57 (or later R470), 510.47 (or later R510), or 515.65 (or later R515).

The CUDA driver's compatibility package only supports particular drivers. Thus, users should upgrade from all R418, R440, and R460 drivers, which are not forward-compatible with CUDA 11.8. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 22.12 supports CUDA compute capability 3.5 and later. This corresponds to GPUs in the NVIDIA Kepler, Maxwell, NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, and NVIDIA Hopper™ architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 22.12 is based on TensorRT 8.5.1.

  For a list of the features and enhancements that were introduced in TensorRT 8.5.1, refer to the TensorRT 8.5 release notes.
- Ubuntu 20.04 with November 2022 updates.

Announcements

- Starting with the 22.05 release, the TensorRT container is available for the Arm SBSA platform.

  For example, when you pull the `nvcr.io/nvidia/tensorrt:22.05-py3` Docker image on an Arm SBSA machine, the Arm-specific image is automatically fetched.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 22.05 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

Note: Since the 22.09 release is based on an early access version of TensorRT 8.5, which is not accompanied by the publication of a corresponding TensorRT Open Source Software (OSS) release to GitHub, please specify building from the *main* branch in `install_opensource.sh` until the TensorRT OSS 8.5.1 release is posted.

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
22.12	20.04	NVIDIA CUDA 11.8.0	TensorRT 8.5.1
22.11			
22.10			TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 24. TensorRT Release 22.11

The NVIDIA container image for TensorRT, release 22.10, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.5.1.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 20.04

  Note: Container image `22.11-py3` contains Python 3.8.
- NVIDIA CUDA® 11.8.0
- NVIDIA cuBLAS 11.11.3.6

- NVIDIA cuDNN 8.7.0
- NVIDIA NCCL 2.15.5

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- OpenMPI 4.1.4+
- OpenUCX 1.14.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.12.2tp1
- Nsight Compute 2022.3.0.22
- Nsight Systems 2022.4.2.1

Driver Requirements

Release 22.11 is based on CUDA 11.8.0, which requires NVIDIA Driver release 520 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 450.51 (or later R450), 470.57 (or later R470), 510.47 (or later R510), or 515.65 (or later R515).

The CUDA driver's compatibility package only supports particular drivers. Thus, users should upgrade from all R418, R440, and R460 drivers, which are not forward-compatible with CUDA 11.8. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 22.11 supports CUDA compute capability 3.5 and later. This corresponds to GPUs in the NVIDIA Kepler, Maxwell, NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, and NVIDIA Hopper™ architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 22.11 is based on TensorRT 8.5.1.

  For a list of the features and enhancements that were introduced in TensorRT 8.5.1, refer to the TensorRT 8.5 release notes.
- Ubuntu 20.04 with October 2022 updates.

Announcements

- Starting with the 22.05 release, the TensorRT container is available for the Arm SBSA platform.

  For example, when you pull the `nvcr.io/nvidia/tensorrt:22.05-py3` Docker image on an Arm SBSA machine, the Arm-specific image is automatically fetched.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 22.05 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

Note: Since the 22.09 release is based on an early access version of TensorRT 8.5, which is not accompanied by the publication of a corresponding TensorRT Open Source Software (OSS) release to GitHub, please specify building from the *main* branch in `install_opensource.sh` until the TensorRT OSS 8.5.1 release is posted.

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
22.11	20.04	NVIDIA CUDA 11.8.0	TensorRT 8.5.1
22.10			TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 25. TensorRT Release 22.10

The NVIDIA container image for TensorRT, release 22.10, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.5 EA.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 20.04

  Note: Container image `22.10-py3` contains Python 3.8.
- NVIDIA CUDA® 11.8.0
- NVIDIA cuBLAS 11.11.3.6

- NVIDIA cuDNN 8.6.0.163
- NVIDIA NCCL 2.15.5

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- OpenMPI 4.1.5a1
- OpenUCX 1.12.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.12.2tp1
- Nsight Compute 2022.3.0.22
- Nsight Systems 2022.4.2.1

Driver Requirements

Release 22.10 is based on CUDA 11.8.0, which requires NVIDIA Driver release 520 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 450.51 (or later R450), 470.57 (or later R470), 510.47 (or later R510), or 515.65 (or later R515).

The CUDA driver's compatibility package only supports particular drivers. Thus, users should upgrade from all R418, R440, and R460 drivers, which are not forward-compatible with CUDA 11.8. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 22.10 supports CUDA compute capability 3.5 and later. This corresponds to GPUs in the NVIDIA Kepler, Maxwell, NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, and NVIDIA Hopper™ architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 22.10 is based on TensorRT 8.5 EA.

  For a list of the features and enhancements that were introduced in TensorRT 8.5.0.12, refer to the TensorRT 8.5 release notes.
- Ubuntu 20.04 with September 2022 updates.

Announcements

- Starting with the 22.05 release, the TensorRT container is available for the Arm SBSA platform.

  For example, when you pull the `nvcr.io/nvidia/tensorrt:22.05-py3` Docker image on an Arm SBSA machine, the Arm-specific image is automatically fetched.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 22.05 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

Note: Since the 22.09 release is based on an early access version of TensorRT 8.5, which is not accompanied by the publication of a corresponding TensorRT Open Source Software (OSS) release to GitHub, please specify building from the *main* branch in `install_opensource.sh` until the TensorRT OSS 8.5.1 release is posted.

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
22.10	20.04	NVIDIA CUDA 11.8.0	TensorRT 8.5 EA
22.09			
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 26. TensorRT Release 22.09

The NVIDIA container image for TensorRT, release 22.09, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.5 EA.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software.

The container also includes the following:

- Ubuntu 20.04

  Note: Container image `22.09-py3` contains Python 3.8.
- NVIDIA CUDA® 11.8.0
- NVIDIA cuBLAS 11.11.3.6

- NVIDIA cuDNN 8.6.0.158
- NVIDIA NCCL 2.15.1

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- OpenMPI 4.1.2rc4+
- OpenUCX 1.12.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.12.1a0
- Nsight Compute 2022.3.0.22
- Nsight Systems 2022.3.1.43

Driver Requirements

Release 22.09 is based on CUDA 11.8.0, which requires NVIDIA Driver release 520 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 450.51 (or later R450), 470.57 (or later R470), 510.47 (or later R510), or 515.65 (or later R515).

The CUDA driver's compatibility package only supports particular drivers. Thus, users should upgrade from all R418, R440, and R460 drivers, which are not forward-compatible with CUDA 11.8. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 22.09 supports CUDA compute capability 3.5 and later. This corresponds to GPUs in the NVIDIA Kepler, Maxwell, NVIDIA Pascal, NVIDIA Volta™, NVIDIA Turing™, NVIDIA Ampere architecture, and NVIDIA Hopper™ architecture families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 22.09 is based on TensorRT 8.5 EA.

  For a list of the features and enhancements that were introduced in TensorRT 8.5.0.12, refer to the TensorRT 8.5 release notes.
- Ubuntu 20.04 with August 2022 updates.

Announcements

- Starting with the 22.05 release, the TensorRT container is available for the Arm SBSA platform.

  For example, when you pull the `nvcr.io/nvidia/tensorrt:22.05-py3` Docker image on an Arm SBSA machine, the Arm-specific image is automatically fetched.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 22.05 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh -b main
```

Note: Since the 22.09 release is based on an early access version of TensorRT 8.5, which is not accompanied by the publication of a corresponding TensorRT Open Source Software (OSS) release to GitHub, please specify building from the *main* branch in `install_opensource.sh` until the TensorRT OSS 8.5.1 release is posted.

For more information, see GitHub: TensorRT.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
22.09	20.04	NVIDIA CUDA 11.8.0	TensorRT 8.5 EA
22.08		NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 27. TensorRT Release 22.08

The NVIDIA container image for TensorRT, release 22.08, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.4.2.4.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software:

  - https://github.com/NVIDIA/TensorRT/releases/tag/22.08

The container also includes the following:

- Ubuntu 20.04

  Note: Container image `22.08-py3` contains Python 3.8.

- NVIDIA CUDA® 11.7.1

- NVIDIA cuBLAS 11.10.3.66
- NVIDIA cuDNN 8.5.0.96
- NVIDIA NCCL 2.12.12 (built with CUDA 11.7)

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- OpenMPI 4.1.2rc4+
- OpenUCX 1.12.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.10
- Nsight Compute 2022.1.1.2
- Nsight Systems 2022.1.3.18

Driver Requirements

Release 22.08 is based on CUDA 11.7.1, which requires NVIDIA Driver release 515 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 450.51 (or later R450), 470.57 (or later R470), or 510.47 (or later R510).

The CUDA driver's compatibility package only supports particular drivers. Thus, users should upgrade from all R418, R440, and R460 drivers, which are not forward-compatible with CUDA 11.7. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 22.08 supports CUDA compute capability 3.5 and later. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and NVIDIA Ampere Architecture GPU families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 22.08 is based on TensorRT 8.4.2.4.

  For a list of the features and enhancements that were introduced in TensorRT 8.4.2.4, refer to the TensorRT 8.4.2 release notes.
- Ubuntu 20.04 with July 2022 updates.

Announcements

- Starting with the 22.05 release, the TensorRT container is available for the Arm SBSA platform.

  For example, when you pull the `nvcr.io/nvidia/tensorrt:22.05-py3` Docker image on an Arm SBSA machine, the Arm-specific image is automatically fetched.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 22.05 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information, see GitHub: TensorRT 22.08.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
22.08	20.04	NVIDIA CUDA 11.7.1	TensorRT 8.4.2.4
22.07		NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.03 20.02 20.01		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 28. TensorRT Release 22.07

The NVIDIA container image for TensorRT, release 22.07, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.4.1.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software:
  - https://github.com/NVIDIA/TensorRT/releases/tag/22.07

The container also includes the following:

- Ubuntu 20.04

  Note: Container image `22.07-py3` contains Python 3.8.
- NVIDIA CUDA® 11.7 Update 1 Preview

- NVIDIA cuBLAS 11.10.3.66
- NVIDIA cuDNN 8.4.1
- NVIDIA NCCL 2.12.12 (built with CUDA 11.7)

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- OpenMPI 4.1.2rc4+
- OpenUCX 1.12.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.10
- Nsight Compute 2022.1.1.2
- Nsight Systems 2022.1.3.3

Driver Requirements

Release 22.07 is based on CUDA 11.7 Update 1 Preview, which requires NVIDIA Driver release 515 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 450.51 (or later R450), 470.57 (or later R470), or 510.47 (or later R510).

The CUDA driver's compatibility package only supports particular drivers. Thus, users should upgrade from all R418, R440, and R460 drivers, which are not forward-compatible with CUDA 11.7. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 22.07 supports CUDA compute capability 3.5 and later. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and NVIDIA Ampere Architecture GPU families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 22.07 is based on TensorRT 8.4.1.

  For a list of the features and enhancements that were introduced in TensorRT 8.4.1, refer to the TensorRT 8.4.1 release notes.
- Ubuntu 20.04 with June 2022 updates.

Announcements

- Starting with the 22.05 release, the TensorRT container is available for the Arm SBSA platform.

  For example, when you pull the `nvcr.io/nvidia/tensorrt:22.05-py3` Docker image on an Arm SBSA machine, the Arm-specific image is automatically fetched.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 22.05 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information, see GitHub: TensorRT 22.07.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
22.07	20.04	NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.4.1
22.06			TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.02 20.01			
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 29. TensorRT Release 22.06

The NVIDIA container image for TensorRT, release 22.06, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.2.5.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software:

  - https://github.com/NVIDIA/TensorRT/releases/tag/22.06

The container also includes the following:

- Ubuntu 20.04

  Note: Container image `22.06-py3` contains Python 3.8.

- NVIDIA CUDA® 11.7 Update 1 Preview

- NVIDIA cuBLAS 11.10.3.66
- NVIDIA cuDNN 8.4.1
- NVIDIA NCCL 2.12.12 (built with CUDA 11.7)

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- OpenMPI 4.1.2rc4+
- OpenUCX 1.12.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.10
- Nsight Compute 2022.1.1.2
- Nsight Systems 2022.1.3.3

Driver Requirements

Release 22.06 is based on CUDA 11.7 Update 1 Preview, which requires NVIDIA Driver release 515 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 450.51 (or later R450), 470.57 (or later R470), or 510.47 (or later R510).

The CUDA driver's compatibility package only supports particular drivers. Thus, users should upgrade from all R418, R440, and R460 drivers, which are not forward-compatible with CUDA 11.7. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 22.06 supports CUDA compute capability 3.5 and later. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and NVIDIA Ampere Architecture GPU families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 22.06 is based on TensorRT 8.2.5.

  For a list of the features and enhancements that were introduced in TensorRT 8.2.5, refer to the TensorRT 8.2.5 release notes.
- Ubuntu 20.04 with May 2022 updates.

Announcements

- Starting with the 22.05 release, the TensorRT container is available for the Arm SBSA platform.

  For example, when you pull the `nvcr.io/nvidia/tensorrt:22.05-py3` Docker image on an Arm SBSA machine, the Arm-specific image is automatically fetched.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 22.05 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information, see GitHub: TensorRT 22.06.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
22.06	20.04	NVIDIA CUDA 11.7 Update 1 Preview	TensorRT 8.2.5
22.05		NVIDIA CUDA 11.7.0	
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.02 20.01			
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 30. TensorRT Release 22.05

The NVIDIA container image for TensorRT, release 22.05, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.2.5.1.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software:
  - https://github.com/NVIDIA/TensorRT/releases/tag/22.05

The container also includes the following:

- Ubuntu 20.04

  Note: Container image `22.05-py3` contains Python 3.8.
- NVIDIA CUDA® 11.7.0

- NVIDIA cuBLAS 11.10.1.25
- NVIDIA cuDNN 8.4.0.27
- NVIDIA NCCL 2.12.10 (optimized for NVIDIA NVLink™)

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- OpenMPI 4.1.2rc4+
- OpenUCX 1.12.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.10
- Nsight Compute 2022.1.1.2
- Nsight Systems 2022.1.3.3

Driver Requirements

Release 22.05 is based on CUDA 11.7, which requires NVIDIA Driver release 515 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 450.51 (or later R450), 470.57 (or later R470), or 510.47 (or later R510).

The CUDA driver's compatibility package only supports particular drivers. Thus, users should upgrade from all R418, R440, and R460 drivers, which are not forward-compatible with CUDA 11.7. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 22.05 supports CUDA compute capability 3.5 and later. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and NVIDIA Ampere Architecture GPU families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 22.05 is based on TensorRT 8.2.5.

  For a list of the features and enhancements that were introduced in TensorRT 8.2.5, refer to the TensorRT 8.2.5 release notes.
- Ubuntu 20.04 with April 2022 updates.
- Added Disentangled attention plugin for DeBERTa.
- Added DMHA (`multiscaleDeformableAttnPlugin`) plugin for DDETR.

- Added fp16 support for `pillarScatterPlugin`.
- Removed usage of deprecated TensorRT APIs in samples.

Announcements

- Starting with the 22.05 release, the TensorRT container is available for the Arm SBSA platform.

  For example, when you pull the `nvcr.io/nvidia/tensorrt:22.05-py3` Docker image on an Arm SBSA machine, the Arm-specific image is automatically fetched.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 22.05 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information, see GitHub: TensorRT 22.05.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
22.05	20.04	NVIDIA CUDA 11.7.0	TensorRT 8.2.5.1
22.04		NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03 20.02		NVIDIA CUDA 10.2.89	TensorRT 7.0.0

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.01			
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 31. TensorRT Release 22.04

The NVIDIA container image for TensorRT, release 22.04, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.2.4.2.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software:

  - https://github.com/NVIDIA/TensorRT/releases/tag/22.04

Here are the major updates to the 22.04 TensorRT Open Source Software release:

- Bug fixes and refactored the `PyramidROIAlign` plugin.
- Fixed the `MultilevelCropAndResize` plugin crashes on Windows.
- Added a Detectron2 Mask R-CNN R50-FPN Python sample.
- Removed sampleNMT.

The container also includes the following:

- Ubuntu 20.04

> Note: Container image `22.04-py3` contains Python 3.8.

- NVIDIA CUDA® 11.6.2
- cuBLAS 11.9.3.115
- NVIDIA cuDNN 8.4.0.27
- NVIDIA NCCL 2.12.10 (optimized for NVIDIA NVLink™)

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- OpenMPI 4.1.2rc4+
- OpenUCX 1.12.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.10
- Nsight Compute 2022.1.1.2
- Nsight Systems 2022.2.1.31-5fe97ab

Driver Requirements

Release 22.04 is based on CUDA 11.6.2, which requires NVIDIA Driver release 510 or later. However, if you are running on a Data Center GPU (for example, T4 or any other Tesla board), use NVIDIA driver release 418.40 (or later R418), 440.33 (or later R440), 450.51 (or later R450), 460.27 (or later R460), or 470.57 (or later R470). The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see CUDA Application Compatibility. For more information, see CUDA Compatibility and Upgrades and NVIDIA CUDA and Drivers Support.

GPU Requirements

Release 22.04 supports CUDA compute capability 3.5 and later. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and NVIDIA Ampere Architecture GPU families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 22.04 is based on TensorRT 8.2.4.2.

For a list of the features and enhancements that were introduced in TensorRT 8.2.4.2, refer to the TensorRT 8.2.4.2 release notes.

- Ubuntu 20.04 with March 2022 updates.

Announcements

- Starting with the 21.12 release, a beta version of the TensorRT container is available for the Arm SBSA platform.

  For example, when you pull the `nvcr.io/nvidia/tensorrt:22.04-py3` Docker image on an Arm SBSA machine, the Arm-specific image is automatically fetched.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 22.04 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information, see GitHub: TensorRT 22.04.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.

- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
22.04	20.04	NVIDIA CUDA 11.6.2	TensorRT 8.2.4.2
22.03		NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.01			
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 32. TensorRT Release 22.03

The NVIDIA container image for TensorRT, release 22.03, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation.
  - Build the samples can be by running `make` in the `/workspace/tensorrt/samples` directory.
  - The resulting executables are in the `/workspace/tensorrt/bin` directory.
  - The C++ API documentation is in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation.
  - The Python samples are in the `/workspace/tensorrt/samples/python` directory.

    Refer to the respective README documents for more samples.
  - Many Python samples can be run by using `python <script.py> -d /workspace/tensorrt/data`.

    For example:

    ```
    python onnx_resnet50.py -d /workspace/tensorrt/data
    ```
  - The Python API documentation is in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.2.3.

  The ONNX parser and plug-in libraries that are bundled with this container are built from TensorRT Open Source Software:

  - https://github.com/NVIDIA/TensorRT/releases/tag/22.03

The container also includes the following:

- Ubuntu 20.04

  Note: Container image `22.03-py3` contains Python 3.8.
- NVIDIA CUDA® 11.6.1

- cuBLAS 11.8.1.74
- NVIDIA cuDNN 8.3.3.40
- NVIDIA NCCL 2.12.9 (optimized for NVIDIA NVLink™)

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- OpenMPI 4.1.2rc4+
- OpenUCX 1.12.0
- GDRCopy 2.3
- NVIDIA HPC-X 2.10
- Nsight Compute 2022.1.1.2
- Nsight Systems 2022.1.5.2.53

Driver Requirements

Release 22.03 is based on CUDA 11.6.1, which requires NVIDIA Driver release 510 or later. However, if you are running on a Data Center GPU (for example, T4 or any other Tesla board), use NVIDIA driver release 418.40 (or later R418), 440.33 (or later R440), 450.51 (or later R450), 460.27 (or later R460), or 470.57 (or later R470). The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see CUDA Application Compatibility. For more information, see CUDA Compatibility and Upgrades and NVIDIA CUDA and Drivers Support.

GPU Requirements

Release 22.03 supports CUDA compute capability 3.5 and later. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and NVIDIA Ampere Architecture GPU families. For a list of GPUs to which this compute capability corresponds, see CUDA GPUs. For additional support details, see the Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 22.03 is based on TensorRT 8.2.3.

  For a list of the features and enhancements that were introduced in TensorRT 8.2.3, refer to the TensorRT 8.2.3 release notes.
- Ubuntu 20.04 with February 2202 updates.

Announcements

- Starting with the 21.12 release, a beta version of the TensorRT container is available for the Arm SBSA platform.

  For example, when you pull the Docker image nvcr.io/nvidia/tensorrt:22.03-py3 on an Arm SBSA machine will automatically fetch the Arm-specific image.
- NVIDIA Deep Learning Profiler (DLProf) v1.8, which was included in the 21.12 container, was the last release of DLProf.

  Starting with the 22.01 container, DLProf is longer included. It can still be manually installed by using a pip wheel on the `nvidia-pyindex`.

Obtaining Missing Data Files

Some samples require data files that are not included in the TensorRT container because of licensing restrictions, or because they are too large. Samples that do not include the required data files include a `README.md` file in the corresponding source directory that provides information about how to obtain the necessary data files.

Installing Required Python Modules

- To complete some of the samples, you might want to first run the Python setup script.
- If you need to install the missing Python modules and their dependencies, run the `/opt/tensorrt/python/python_setup.sh` script.

Installing Open Source Components

A script has been added to clone, build, and replace the provided plug-in, the Caffe parser, and the ONNX parser libraries with the open source ones that are based on the 22.03 tag on the official TensorRT open source repository.

To install the open source components in the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information, see GitHub: TensorRT 22.03.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

  Use the TensorFlow Container to accelerate through TF-TRT instead.
- Torch-TensorRT is not supported in the TensorRT containers.

  Use the PyTorch Container to accelerate through Torch-TRT instead.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
22.03	20.04	NVIDIA CUDA 11.6.1	TensorRT 8.2.3
22.02		NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01		NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for Arm SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03 20.02 20.01		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12			TensorRT 6.0.1

Container Version	Ubuntu	CUDA Toolkit	TensorRT
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

None.

Chapter 33. TensorRT Release 22.02

The NVIDIA container image for TensorRT, release 22.02, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Refer to the respective README documents for more samples. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python onnx_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.2.3. Note that the ONNX parser and plugin libraries bundled with this container are built from TensorRT Open Source Software: https://github.com/NVIDIA/TensorRT/releases/tag/22.02https://github.com/NVIDIA/TensorRT/releases/tag/22.02.

The container also includes the following:

- Ubuntu 20.04

> Note: Container image `22.02-py3` contains Python 3.8.

- NVIDIA CUDA 11.6.0
- cuBLAS 11.8.1.74
- NVIDIA cuDNN 8.3.2
- NVIDIA NCCL 2.11.4 (optimized for NVLink™)

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- NVIDIA HPC-X 2.10
- OpenMPI 4.1.2rc4+
- OpenUCX 1.12.0
- GDRCopy 2.3
- Nsight Compute 2021.1.0.18
- Nsight Systems 2021.5.2.53

Driver Requirements

Release 22.02 is based on NVIDIA CUDA 11.6.0, which requires NVIDIA Driver release 510 or later. However, if you are running on a Data Center GPU (for example, T4 or any other Tesla board), you may use NVIDIA driver release 418.40 (or later R418), 440.33 (or later R440), 450.51 (or later R450), 460.27 (or later R460), or 470.57 (or later R470). The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades and NVIDIA CUDA and Drivers Support.

GPU Requirements

Release 22.02 supports CUDA compute capability 3.5 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and NVIDIA Ampere Architecture GPU families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 22.02 is based on TensorRT 8.2.3. For a list of the new features and enhancements introduced in TensorRT 8.2.3 refer to the TensorRT 8.2.3 release notes.
- Ubuntu 20.04 with January 2022 updates.

Announcements

- Starting with the 21.12 release, a beta version of the TensorRT container is available for the ARM SBSA platform. For example, pulling the Docker image nvcr.io/nvidia/tensorrt:22.02-py3 on an ARM SBSA machine will automatically fetch the ARM-specific image.
- DLProf v1.8, which was included in the 21.12 container, was the last release of DLProf. Starting with the 22.01 container, DLProf is longer included. It can still be manually installed via a pip wheel on the nvidia-pyindex.

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 21.10 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information see GitHub: TensorRT 22.02.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers. Please use the TensorFlow Container to accelerate via TF-TRT.
- Torch-TensorRT is not supported in the TensorRT containers. Please use the PyTorch Container to accelerate via Torch-TRT.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
22.02	20.04	NVIDIA CUDA 11.6.0	TensorRT 8.2.3
22.01			TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	

Container Version	Ubuntu	CUDA Toolkit	TensorRT
			TensorRT 8.0.2.2 for ARM SBSA Linux
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

- None.

Chapter 34. TensorRT Release 22.01

The NVIDIA container image for TensorRT, release 22.01, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Refer to the respective README documents for more samples. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python onnx_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.2.2. Note that the ONNX parser and plugin libraries bundled with this container are built from TensorRT Open Source Software: https://github.com/NVIDIA/TensorRT/releases/tag/22.01.

The container also includes the following:

- Ubuntu 20.04

  Note: Container image `22.01-py3` contains Python 3.8.

- NVIDIA CUDA 11.6.0
- cuBLAS 11.8.1.74
- NVIDIA cuDNN 8.3.2
- NVIDIA NCCL 2.11.4 (optimized for NVLink™)

  Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- NVIDIA HPC-X 2.10
- OpenMPI 4.1.2rc4+
- OpenUCX 1.12.0
- GDRCopy 2.3
- Nsight Systems 2021.5.2.53

Driver Requirements

Release 22.01 is based on NVIDIA CUDA 11.6.0, which requires NVIDIA Driver release 510 or later. However, if you are running on a Data Center GPU (for example, T4 or any other Tesla board), you may use NVIDIA driver release 418.40 (or later R418), 440.33 (or later R440), 450.51 (or later R450), 460.27 (or later R460), or 470.57 (or later R470). The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades and NVIDIA CUDA and Drivers Support.

GPU Requirements

Release 22.01 supports CUDA compute capability 3.5 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and NVIDIA Ampere Architecture GPU families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 22.01 is based on TensorRT 8.2.2. For a list of the new features and enhancements introduced in TensorRT 8.2.2 refer to the TensorRT 8.2.2 release notes.
- Ubuntu 20.04 with December 2021 updates.

Announcements

- Starting with the 21.12 release, a beta version of the TensorRT container is available for the ARM SBSA platform. For example, pulling the Docker image nvcr.io/nvidia/tensorrt:22.01-py3 on an ARM SBSA machine will automatically fetch the ARM-specific image.
- DLProf v1.8, which was included in the 21.12 container, was the last release of DLProf. Starting with the 22.01 container, DLProf is longer included. It can still be manually installed via a pip wheel on the nvidia-pyindex.

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 21.10 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information see GitHub: TensorRT 22.01.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers. Please use the TensorFlow Container to accelerate via TF-TRT.
- Torch-TensorRT is not supported in the TensorRT containers. Please use the PyTorch Container to accelerate via Torch-TRT.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
22.01	20.04	NVIDIA CUDA 11.6.0	TensorRT 8.2.2
21.12		NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	

Container Version	Ubuntu	CUDA Toolkit	TensorRT
			TensorRT 8.0.2.2 for ARM SBSA Linux
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

- None.

Chapter 35. TensorRT Release 21.12

The NVIDIA container image for TensorRT, release 21.12, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Refer to the respective README documents for more samples. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python onnx_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.2.1.8. Note that the ONNX parser and plugin libraries bundled with this container are built from TensorRT Open Source Software: https://github.com/NVIDIA/TensorRT/releases/.

The container also includes the following:

- Ubuntu 20.04

  Note: Container image `21.12-py3` contains Python 3.8.

- NVIDIA CUDA 11.5.0
- cuBLAS 11.7.3.1
- NVIDIA cuDNN 8.3.1.22
- NVIDIA NCCL 2.11.4 (optimized for NVLink™)

  Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- OpenMPI 4.1.1+
- OpenUCX 1.11.0rc1
- GDRCopy 2.3
- NVIDIA HPC-X 2.9
- Nsight Systems 2021.3.2.4

Driver Requirements

Release 21.12 is based on NVIDIA CUDA 11.5.0, which requires NVIDIA Driver release 495 or later. However, if you are running on a Data Center GPU (for example, T4 or any other Tesla board), you may use NVIDIA driver release 418.40 (or later R418), 440.33 (or later R440), 450.51 (or later R450), 460.27 (or later R460), or 470.57 (or later R470). The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades and NVIDIA CUDA and Drivers Support.

GPU Requirements

Release 21.12 supports CUDA compute capability 3.5 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and NVIDIA Ampere Architecture GPU families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 21.12 is based on TensorRT 8.2.1.8. For a list of the new features and enhancements introduced in TensorRT 8.2.1 refer to the TensorRT 8.2.1 release notes.
- Ubuntu 20.04 with November 2021 updates.

Announcements

- Starting with the 21.12 release, a beta version of the TensorRT container is available for the ARM SBSA platform. Pulling the Docker image nvcr.io/nvidia/tensorrt:21.12-py3 on an ARM SBSA machine will automatically fetch the ARM-specific image.
- DLProf v1.8, which is included in the 21.12 container, will be the last release of DLProf. Starting with the 22.01 container, DLProf will no longer be included. It can still be manually installed via a pip wheel on the nvidia-pyindex.

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 21.10 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information see GitHub: TensorRT 21.12.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers. Please use the TensorFlow Container to accelerate via TF-TRT.
- Torch-TensorRT is not supported in the TensorRT containers. Please use the PyTorch Container to accelerate via Torch-TRT.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21,12	20.04	NVIDIA CUDA 11.5.0	TensorRT 8.2.1.8
21.11			TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for ARM SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

- None.

Chapter 36. TensorRT Release 21.11

The NVIDIA container image for TensorRT, release 21.11, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Refer to the respective README documents for more samples. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python onnx_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.0.3.4. Note that the ONNX parser and plugin libraries bundled with this container are built from TensorRT Open Source Software: https://github.com/NVIDIA/TensorRT/releases/.

The container also includes the following:

- Ubuntu 20.04

  Note: Container image `21.11-py3` contains Python 3.8.

- NVIDIA CUDA 11.5.0
- cuBLAS 11.7.3.1
- NVIDIA cuDNN 8.3.0.96
- NVIDIA NCCL 2.11.4 (optimized for NVLink™)

  Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- OpenMPI 4.1.1+
- OpenUCX 1.11.0rc1
- GDRCopy 2.3
- NVIDIA HPC-X 2.9
- Nsight Systems 2021.3.2.4

Driver Requirements

Release 21.11 is based on NVIDIA CUDA 11.5.0, which requires NVIDIA Driver release 495 or later. However, if you are running on a Data Center GPU (for example, T4 or any other Tesla board), you may use NVIDIA driver release 418.40 (or later R418), 440.33 (or later R440), 450.51 (or later R450), 460.27 (or later R460), or 470.57 (or later R470). The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades and NVIDIA CUDA and Drivers Support.

GPU Requirements

Release 21.11 supports CUDA compute capability 3.5 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and NVIDIA Ampere Architecture GPU families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 21.11 is based on TensorRT 8.0.3.4. For a list of the new features and enhancements introduced in TensorRT 8.0.3 refer to the TensorRT 8.0.3 release notes.
- Ubuntu 20.04 with October 2021 updates.

Announcements

DLProf v1.8, which will be included in the 21.12 container, will be the last release of DLProf. Starting with the 22.01 container, DLProf will no longer be included. It can still be manually installed via a pip wheel on the nvidia-pyindex.

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 21.10 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information see GitHub: TensorRT 21.11.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.11	20.04	NVIDIA CUDA 11.5.0	TensorRT 8.0.3.4 for x64 Linux TensorRT 8.0.2.2 for ARM SBSA Linux
21.10		NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03 20.02 20.01		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

- None.

Chapter 37. TensorRT Release 21.10

The NVIDIA container image for TensorRT, release 21.10, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Refer to the respective README documents for more samples. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python onnx_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.0.3.4. Note that the ONNX parser and plugin libraries bundled with this container are built from TensorRT Open Source Software: https://github.com/NVIDIA/TensorRT/releases/tag/21.10. Prominent updates to the 21.10 TensorRT Open Source Software release are:
  - Bump TensorRT version to 8.0.3.4
  - demo/BERT enhancements:
    - Added benchmark script for demoBERT-Megatron
    - Use static shape for single batch single sequence inputs
    - Revert to using native FC layer and FCPlugin only for older GPUs
  - Plugin enhancements:
    - Dynamic Input Shape support for EfficientNMS plugin
  - ONNX support enhancements:
    - Update ONNX submodule to v1.8.0
    - Support empty dimensions in ONNX

        - Several bugfixes and documentation updates
    - Updates to TensorRT developer tools:
        - Polygraphy v0.33.0
            - Added various examples, a CLI User Guide and how-to guides.
            - Added experimental support for DLA
            - Added a PluginRefRunner which provides CPU reference implementations for TensorRT plugins
    - Bugfixes and documentation updates in pytorch-quantization toolkit.

The container also includes the following:

- Ubuntu 20.04

> Note: Container image `21.10-py3` contains Python 3.8.

- NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2
- NVIDIA cuDNN 8.2.4.15
- NVIDIA NCCL 2.11.4 (optimized for NVLink™)

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- OpenMPI 4.1.1+
- OpenUCX 1.11.0rc1
- GDRCopy 2.3
- NVIDIA HPC-X 2.9
- Nsight Systems 2021.3.2.4

Driver Requirements

Release 21.10 is based on NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2, which requires NVIDIA Driver release 470 or later. However, if you are running on Data Center GPUs (formerly Tesla), for example, T4, you may use NVIDIA driver release 418.40 (or later R418), 440.33 (or later R440), 450.51 (or later R450), or 460.27 (or later R460). The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades and NVIDIA CUDA and Drivers Support.

GPU Requirements

Release 21.10 supports CUDA compute capability 3.5 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and NVIDIA Ampere Architecture GPU

families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 21.10 is based on TensorRT 8.0.3.4. For a list of the new features and enhancements introduced in TensorRT 8.0.3 refer to the TensorRT 8.0.3 release notes.
- Ubuntu 20.04 with September 2021 updates.

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 21.10 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information see GitHub: TensorRT 21.10.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.10	20.04	NVIDIA CUDA 11.4.2 with cuBLAS 11.6.5.2	TensorRT 8.0.3.4
21.09		NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

- None.

Chapter 38. TensorRT Release 21.09

The NVIDIA container image for TensorRT, release 21.09, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Refer to the respective README documents for more samples. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python onnx_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.0.3. Note that the ONNX parser and plugin libraries bundled with this container are built from TensorRT Open Source Software: https://github.com/NVIDIA/TensorRT/releases/tag/21.09. Prominent updates to the 21.08 TensorRT Open Source Software release are:
  - Added demoBERT and demoBERT-MT (sparsity) benchmark data for TensorRT 8.
  - Added example python notebooks for BERT - Q&A with TensorRT and EfficientNet - Object Detection with TensorRT.
  - Updated samples and plugins directory structure.
  - Updates to TensorRT developer tools:
    - Polygraphy v0.31.1
    - ONNX-GraphSurgeon v0.3.11
    - pytorch-quantization toolkit v2.1.1

The container also includes the following:

- Ubuntu 20.04

> Note: Container image `21.09-py3` contains Python 3.8.

- NVIDIA CUDA 11.4.2
- cuBLAS 11.6.1.51
- NVIDIA cuDNN 8.2.4.15
- NVIDIA NCCL 2.11.4 (optimized for NVLink™)

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- OpenMPI 4.1.1+
- OpenUCX 1.11.0rc1
- GDRCopy 2.3
- NVIDIA HPC-X 2.9
- Nsight Systems 2021.3.1.57
- TensorRT 8.0.3

Driver Requirements

Release 21.09 is based on NVIDIA CUDA 11.4.2, which requires NVIDIA Driver release 470 or later. However, if you are running on Data Center GPUs (formerly Tesla), for example, T4, you may use NVIDIA driver release 418.40 (or later R418), 440.33 (or later R440), 450.51 (or later R450), or 460.27 (or later R460). The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades and NVIDIA CUDA and Drivers Support.

GPU Requirements

Release 21.09 supports CUDA compute capability 3.5 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and NVIDIA Ampere Architecture GPU families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 21.09 is based on TensorRT 8.0.3. For a list of the new features and enhancements introduced in TensorRT 8.0.3 refer to the TensorRT 8.0.3 release notes.
- Ubuntu 20.04 with August 2021 updates.

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 21.09 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information see GitHub: TensorRT 21.09.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.09	20.04	NVIDIA CUDA 11.4.2	TensorRT 8.0.3
21.08		NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

- None.

Chapter 39. TensorRT Release 21.08

The NVIDIA container image for TensorRT, release 21.08, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Refer to the respective README documents for more samples. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python onnx_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.0.1.6. Note that the ONNX parser and plugin libraries bundled with this container are built from TensorRT Open Source Software: https://github.com/NVIDIA/TensorRT/releases/tag/21.08. Prominent updates to the 21.08 TensorRT Open Source Software release are:
  - Added demoBERT and demoBERT-MT (sparsity) benchmark data for TensorRT 8.
  - Added example python notebooks for BERT - Q&A with TensorRT and EfficientNet - Object Detection with TensorRT.
  - Updated samples and plugins directory structure.
  - Updates to TensorRT developer tools:
    - Polygraphy v0.31.1
    - ONNX-GraphSurgeon v0.3.11
    - pytorch-quantization toolkit v2.1.1

The container also includes the following:

- Ubuntu 20.04

> Note: Container image `21.08-py3` contains Python 3.8.

- NVIDIA CUDA 11.4.1
- cuBLAS 11.5.4
- NVIDIA cuDNN 8.2.2.26
- NVIDIA NCCL 2.10.3 (optimized for NVLink™)

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 36.0
- OpenMPI 4.1.1+
- OpenUCX 1.11.0rc1
- GDRCopy 2.2
- NVIDIA HPC-X 2.9
- Nsight Systems 2021.2.4.12
- TensorRT 8.0.1.6

Driver Requirements

Release 21.08 is based on NVIDIA CUDA 11.4.1, which requires NVIDIA Driver release 470 or later. However, if you are running on Data Center GPUs (formerly Tesla), for example, T4, you may use NVIDIA driver release 418.40 (or later R418), 440.33 (or later R440), 450.51 (or later R450), or 460.27 (or later R460). The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades and NVIDIA CUDA and Drivers Support.

GPU Requirements

Release 21.08 supports CUDA compute capability 3.5 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and NVIDIA Ampere Architecture GPU families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 21.08 is based on TensorRT 8.0.1.6. For a list of the new features and enhancements introduced in TensorRT 8.0.1.6 refer to the TensorRT 8.0.1 release notes.
- Ubuntu 20.04 with July 2021 updates.

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 21.08 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information see GitHub: TensorRT 21.08.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.08	20.04	NVIDIA CUDA 11.4.1	TensorRT 8.0.1.6
21.07		NVIDIA CUDA 11.4.0	
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

- None.

Chapter 40. TensorRT Release 21.07

The NVIDIA container image for TensorRT, release 21.07, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 8.0.1.6. Note that the ONNX parser and plugin libraries bundled with this container are built from TensorRT Open Source Software: https://github.com/NVIDIA/TensorRT/releases/tag/21.07. Prominent updates to the 21.07 TensorRT Open Source Software release are:
  - Major upgrade to TensorRT 8.0.1.6 GA.
  - Added support for ONNX operators: Celu, CumSum, EyeLike, GatherElements, GlobalLpPool, GreaterOrEqual, LessOrEqual, LpNormalization, LpPool, ReverseSequence, and SoftmaxCrossEntropyLoss.
  - Enhanced support for ONNX operators: Resize, ConvTranspose, InstanceNormalization, QuantizeLinear, DequantizeLinear, Pad.
  - Added new plugins: EfficientNMS_TRT, EfficientNMS_ONNX_TRT, ScatterND.
  - Added new samples: engine_refit_onnx_bidaf, efficientdet, efficientnet.
  - Added docker build support for Ubuntu20.04 and RedHat/CentOS 8.3.
  - Added Python 3.9 support.

- Updates to ONNX tools: Polygraphy v0.30.3, ONNX-GraphSurgeon v0.3.10, Pytorch Quantization toolkit v2.1.0.
- Removed IPlugin and IPluginFactory interfaces.
- Removed samples: samplePlugin, sampleMovieLens, sampleMovieLensMPS.
- Removed docker build support for Ubuntu 16.04, and PowerPC.

The container also includes the following:

- Ubuntu 20.04

> Note: Container image `21.07-py3` contains Python 3.8.

- NVIDIA CUDA 11.4.0
- cuBLAS 11.5.2.43
- NVIDIA cuDNN 8.2.2.26
- NVIDIA NCCL 2.10.3 (optimized for NVLink™)

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 32.1
- OpenMPI 4.1.1rc1
- OpenUCX 1.10.1
- GDRCopy 2.2
- NVIDIA HPC-X 2.8.2rc3
- Nsight Compute 2021.1.0.18
- Nsight Systems 2021.2.4.12
- TensorRT 8.0.1.6

Driver Requirements

Release 21.07 is based on NVIDIA CUDA 11.4.0, which requires NVIDIA Driver release 470 or later. However, if you are running on Data Center GPUs (formerly Tesla), for example, T4, you may use NVIDIA driver release 418.40 (or later R418), 440.33 (or later R440), 450.51 (or later R450), or 460.27 (or later R460). The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades and NVIDIA CUDA and Drivers Support.

GPU Requirements

Release 21.07 supports CUDA compute capability 3.5 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and NVIDIA Ampere Architecture GPU

families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 21.07 is based on TensorRT 8.0.1.6. For a list of the new features and enhancements introduced in TensorRT 8.0.1.6 refer to the TensorRT 8.0.1 release notes.
- Ubuntu 20.04 with June 2021 updates.

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 21.07 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information see GitHub: TensorRT 21.07.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.07	20.04	NVIDIA CUDA 11.4.0	TensorRT 8.0.1.6
21.06		NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03 20.02 20.01		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

- The 21.07 release includes libsystemd and libudev versions that have a known vulnerability that was discovered late in our QA process. See CVE-2021-33910 for details. This will be fixed in the next release.

Chapter 41. TensorRT Release 21.06

The NVIDIA container image for TensorRT, release 21.06, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 7.2.3.4. Note that the ONNX parser and plugin libraries bundled with this container are built from TensorRT Open Source Software: https://github.com/NVIDIA/TensorRT/releases/tag/21.06

The container also includes the following:

- Ubuntu 20.04

  Note: Container image `21.06-py3` contains Python 3.8.

- NVIDIA CUDA 11.3.1
- cuBLAS 11.5.1.109
- NVIDIA cuDNN 8.2.1
- NVIDIA NCCL 2.9.9 (optimized for NVLink™)

  Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 32.1
- OpenMPI 4.1.1rc1
- OpenUCX 1.10.1
- GDRCopy 2.2
- NVIDIA HPC-X 2.8.2rc3
- Nsight Compute 2021.1.0.18
- Nsight Systems 2021.2.1.58
- TensorRT 7.2.3.4

Driver Requirements

Release 21.06 is based on NVIDIA CUDA 11.3.1, which requires NVIDIA Driver release 465.19.01 or later. However, if you are running on Data Center GPUs (formerly Tesla), for example, T4, you may use NVIDIA driver release 418.40 (or later R418), 440.33 (or later R440), 450.51 (or later R450), or 460.27 (or later R460). The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades and NVIDIA CUDA and Drivers Support.

GPU Requirements

Release 21.06 supports CUDA compute capability 3.5 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and NVIDIA Ampere Architecture GPU families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 21.06 is based on TensorRT 7.2.3.4. For a list of the new features and enhancements introduced in TensorRT 7.2.3.4 refer to the TensorRT 7.2.3 release notes.
- Added missing `model.py` in `uff_custom_plugin` sample.
- Fixed numerical errors for float type in NMS/batchedNMS plugins.
- Removed fcplugin from demoBERT to improve latency.
- Ubuntu 20.04 with May 2021 updates.

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not

include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 21.06 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information see GitHub: TensorRT 21.06.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.06	20.04	NVIDIA CUDA 11.3.1	TensorRT 7.2.3.4
21.05		NVIDIA CUDA 11.3.0	
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03 20.02 20.01		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

There are no known issues in this release.

Chapter 42. TensorRT Release 21.05

The NVIDIA container image for TensorRT, release 21.05, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:

  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```

  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 7.2.3.4. Note that the ONNX parser and plugin libraries bundled with this container are built from TensorRT Open Source Software: https://github.com/NVIDIA/TensorRT/releases/tag/21.05

  Prominent updates to the 21.04 TensorRT Open Source Software release are:

  - Addition of TensorRT Python API bindings.
  - Addition of TensorRT Python samples.
  - Plugin enhancements - FP16 support in batchedNMSPlugin, configurable input sizes for TLT MaskRCNN plugin.
  - ONNX opset13 updates, ResNet example, and documentation updates to PyTorch Quantization toolkit.
  - BERT demo updated to work with Tensorflow 2.x.

The container also includes the following:

- Ubuntu 20.04

> Note: Container image `21.05-py3` contains Python 3.8.

- NVIDIA CUDA 11.3.0
- cuBLAS 11.5.1.101
- NVIDIA cuDNN 8.2.0.51
- NVIDIA NCCL 2.9.8 (optimized for NVLink™)

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 32.1
- OpenMPI 4.1.1rc1
- OpenUCX 1.10.0
- GDRCopy 2.2
- NVIDIA HPC-X 2.8.2rc3
- Nsight Compute 2021.1.0.18
- Nsight Systems 2021.1.3.14
- TensorRT 7.2.3.4

Driver Requirements

Release 21.05 is based on NVIDIA CUDA 11.3.0, which requires NVIDIA Driver release 465.19.01 or later. However, if you are running on Data Center GPUs (formerly Tesla), for example, T4, you may use NVIDIA driver release 418.40 (or later R418), 440.33 (or later R440), 450.51 (or later R450), or 460.27 (or later R460). The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades and NVIDIA CUDA and Drivers Support.

GPU Requirements

Release 21.05 supports CUDA compute capability 3.5 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and NVIDIA Ampere Architecture GPU families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 21.05 is based on TensorRT 7.2.3.4. For a list of the new features and enhancements introduced in TensorRT 7.2.3.4 refer to the TensorRT 7.2.3 release notes.
- Ubuntu 20.04 with April 2021 updates

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 21.05 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information see GitHub: TensorRT 21.05.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.05	20.04	NVIDIA CUDA 11.3.0	TensorRT 7.2.3.4
21.04			
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03 20.02 20.01		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

There are no known issues in this release.

Chapter 43. TensorRT Release 21.04

The NVIDIA container image for TensorRT, release 21.04, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:

  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```

  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 7.2.3.4. Note that the ONNX parser and plugin libraries bundled with this container are built from TensorRT Open Source Software: https://github.com/NVIDIA/TensorRT/releases/tag/21.04

  Prominent updates to the 21.04 TensorRT Open Source Software release are:

  - Addition of TensorRT Python API bindings.
  - Addition of TensorRT Python samples.
  - Plugin enhancements - FP16 support in batchedNMSPlugin, configurable input sizes for TLT MaskRCNN plugin.
  - ONNX opset13 updates, ResNet example, and documentation updates to PyTorch Quantization toolkit.
  - BERT demo updated to work with Tensorflow 2.x.

The container also includes the following:

- Ubuntu 20.04

> Note: Container image `21.04-py3` contains Python 3.8.

- NVIDIA CUDA 11.3.0
- cuBLAS 11.5.1.101
- NVIDIA cuDNN 8.2.0.41
- NVIDIA NCCL 2.9.6 (optimized for NVLink™)

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- rdma-core 32.1
- OpenMPI 4.1.1rc1
- OpenUCX 1.10.0
- GDRCopy 2.2
- NVIDIA HPC-X 2.8.2rc3
- Nsight Compute 2021.1.0.18
- Nsight Systems 2021.1.3.14
- TensorRT 7.2.3.4

Driver Requirements

Release 21.04 is based on NVIDIA CUDA 11.3.0, which requires NVIDIA Driver release 465.19.01 or later. However, if you are running on Data Center GPUs (formerly Tesla), for example, T4, you may use NVIDIA driver release 418.40 (or later R418), 440.33 (or later R440), 450.51 (or later R450), or 460.27 (or later R460). The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades and NVIDIA CUDA and Drivers Support.

GPU Requirements

Release 21.04 supports CUDA compute capability 3.5 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and NVIDIA Ampere Architecture GPU families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 21.04 is based on TensorRT 7.2.3.4. For a list of the new features and enhancements introduced in TensorRT 7.2.3.4 refer to the TensorRT 7.2.3 release notes.
- Ubuntu 20.04 with March 2021 updates

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 21.04 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information see GitHub: TensorRT 21.04.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.04	20.04	NVIDIA CUDA 11.3.0	TensorRT 7.2.3.4
21.03		NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03 20.02 20.01		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

There are no known issues in this release.

Chapter 44. TensorRT Release 21.03

The NVIDIA container image for TensorRT, release 21.03, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:

  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```

  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 7.2.2.3. Note that the ONNX parser and plugin libraries bundled with this container are built from TensorRT Open Source Software: https://github.com/NVIDIA/TensorRT/releases/tag/21.03

  Prominent updates to the 21.03 TensorRT Open Source Software release are:

  - Addition of TensorRT Python API bindings.
  - Addition of TensorRT Python samples.
  - Plugin enhancements - FP16 support in batchedNMSPlugin, configurable input sizes for TLT MaskRCNN plugin.
  - ONNX opset13 updates, ResNet example, and documentation updates to PyTorch Quantization toolkit.
  - BERT demo updated to work with Tensorflow 2.x.

The container also includes the following:

- Ubuntu 20.04

> Note: Container image `21.03-py3` contains Python 3.8.

- NVIDIA CUDA 11.2.1 including cuBLAS 11.4.1.1026
- NVIDIA cuDNN 8.1.1
- NVIDIA NCCL 2.8.4 (optimized for NVLink™)

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- MLNX_OFED 5.1
- OpenMPI 4.0.5
- Nsight Compute 2020.3.0.18
- Nsight Systems 2020.4.3.7

Driver Requirements

Release 21.03 is based on NVIDIA CUDA 11.2.0, which requires NVIDIA Driver release 460.32.03 or later. However, if you are running on Data Center GPUs (formerly Tesla), for example, T4, you may use NVIDIA driver release 418.40 (or later R418), 440.33 (or later R440), 450.51(or later R450). The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades and NVIDIA CUDA and Drivers Support.

GPU Requirements

Release 21.03 supports CUDA compute capability 3.5 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and NVIDIA Ampere Architecture GPU families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 21.03 is based on TensorRT 7.2.2.3. For a list of the new features and enhancements introduced in TensorRT 7.2.2.3 refer to the TensorRT 7.2.2 release notes.
- NVIDIA CUDA 11.2.1 including cuBLAS 11.4.1.1026
- The latest version of NVIDIA cuDNN 8.1.0
- Ubuntu 20.04 with February 2021 updates

Announcements

- Python 2.7 is no longer supported in this TensorRT container release.

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 21.03 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information see GitHub: TensorRT 21.03.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.03	20.04	NVIDIA CUDA 11.2.1	TensorRT 7.2.2.3
21.02		NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03 20.02 20.01		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

There are no known issues in this release.

Chapter 45. TensorRT Release 21.02

The NVIDIA container image for TensorRT, release 21.02, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:

  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```

  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- 7.2.2.3+cuda11.1.1.0.024. Note that the ONNX parser and plugin libraries bundled with this container are built from TensorRT Open Source Software: https://github.com/NVIDIA/TensorRT/releases/tag/21.02

  Prominent updates to the 21.02 TensorRT Open Source Software release are:

  - Addition of TensorRT Python API bindings.
  - Addition of TensorRT Python samples.
  - Plugin enhancements - FP16 support in batchedNMSPlugin, configurable input sizes for TLT MaskRCNN plugin.
  - ONNX opset13 updates, ResNet example, and documentation updates to PyTorch Quantization toolkit.
  - BERT demo updated to work with Tensorflow 2.x.

The container also includes the following:

- Ubuntu 20.04

> Note: Container image `21.02-py3` contains Python 3.8.

- NVIDIA CUDA 11.2.0 including cuBLAS 11.3.1.
- NVIDIA cuDNN 8.1.0
- NVIDIA NCCL 2.8.4 (optimized for NVLink™)

> Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- MLNX_OFED 5.1
- OpenMPI 4.0.5
- Nsight Compute 2020.3.0.18
- Nsight Systems 2020.4.3.7

Driver Requirements

Release 21.02 is based on NVIDIA CUDA 11.2.0, which requires NVIDIA Driver release 460.27.04 or later. However, if you are running on Data Center GPUs (formerly Tesla), for example, T4, you may use NVIDIA driver release 418.40 (or later R418), 440.33 (or later R440), 450.51(or later R450). The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades and NVIDIA CUDA and Drivers Support.

GPU Requirements

Release 21.02 supports CUDA compute capability 3.5 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and NVIDIA Ampere Architecture GPU families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 21.02 is based on TensorRT 7.2.2.3+cuda11.1.0.024. For a list of the new features and enhancements introduced in TensorRT 7.2.2 refer to the TensorRT 7.2.2 release notes.
- NVIDIA CUDA 11.2.0 including cuBLAS 11.3.1.
- The latest version of NVIDIA cuDNN 8.1.0
- The latest version of NVIDIA NCCL 2.8.4

- The latest version of Nsight Compute 2020.3.0.18
- The latest version of Nsight Systems 2020.4.3.7
- Ubuntu 20.04 with January 2021 updates

Announcements

- Python 2.7 is no longer supported in this TensorRT container release.

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 21.02 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information see GitHub: TensorRT 21.02.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
21.02	20.04	NVIDIA CUDA 11.2.0	7.2.2.3+cuda11.1.0.024
20.12		NVIDIA CUDA 11.1.1	TensorRT 7.2.2

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03 20.02 20.01		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

There are no known issues in this release.

Chapter 46. TensorRT Release 21.01

The NVIDIA container image release for TensorRT 21.01 has been canceled. The next release will be the 21.02 release which is expected to be released at the end of February.

Chapter 47. TensorRT Release 20.12

The NVIDIA container image for TensorRT, release 20.12, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 7.2.2. Note that the ONNX parser and plugin libraries bundled with this container are built from TensorRT Open Source Software: https://github.com/NVIDIA/TensorRT/releases/tag/20.12

The container also includes the following:

- Ubuntu 20.04

  Note: Container image `20.12-py3` contains Python 3.8.

- NVIDIA CUDA 11.1.1 including cuBLAS 11.3.0.
- NVIDIA cuDNN 8.0.5
- NVIDIA NCCL 2.8.3 (optimized for NVLink™)

  Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- MLNX_OFED 5.1

- OpenMPI 4.0.5
- Nsight Compute 2020.2.1.8
- Nsight Systems 2020.3.4.32

Driver Requirements

Release 20.12 is based on NVIDIA CUDA 11.1.1, which requires NVIDIA Driver release 455 or later. However, if you are running on Tesla (for example, T4 or any other Tesla board), you may use NVIDIA driver release 418.xx, 440.30, or 450.xx. The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 20.12 supports CUDA compute capability 3.5 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and Ampere Architecture GPU families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 20.12 is based on TensorRT 7.2.2. For a list of the new features and enhancements introduced in TensorRT 7.2.2 refer to the TensorRT 7.2.2 release notes.
- NVIDIA CUDA 11.1.1 including cuBLAS 11.3.0.
- The latest version of NVIDIA cuDNN 8.0.5
- The latest version of NVIDIA NCCL 2.8.3
- The latest version of Nsight Compute 2020.2.1.8
- Ubuntu 20.04 with November 2020 updates

Announcements

- Python 2.7 is no longer supported in this TensorRT container release.

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 20.12 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information see GitHub: TensorRT 20.12.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.12	20.04	NVIDIA CUDA 11.1.1	TensorRT 7.2.2
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03 20.02 20.01		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12			TensorRT 6.0.1

Container Version	Ubuntu	CUDA Toolkit	TensorRT
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

There are no known issues in this release.

Chapter 48. TensorRT Release 20.11

The NVIDIA container image for TensorRT, release 20.11, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 7.2.1. Note that the ONNX parser and plugin libraries bundled with this container are built from TensorRT Open Source Software: https://github.com/NVIDIA/TensorRT/releases/tag/20.11

The container also includes the following:

- Ubuntu 18.04

  Note: Container image `20.11-py3` contains Python 3.6.

- NVIDIA CUDA 11.1.0 including cuBLAS 11.2.1.
- NVIDIA cuDNN 8.0.4
- NVIDIA NCCL 2.8.2 (optimized for NVLink™)

  Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- MLNX_OFED 5.1

- OpenMPI 4.0.5
- Nsight Compute 2020.2.0.18
- Nsight Systems 2020.3.4.32

Driver Requirements

Release 20.11 is based on NVIDIA CUDA 11.1.0, which requires NVIDIA Driver release 455 or later. However, if you are running on Tesla (for example, T4 or any other Tesla board), you may use NVIDIA driver release 418.xx, 440.30, or 450.xx. The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 20.11 supports CUDA compute capability 3.5 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and Ampere Architecture GPU families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 20.11 is based on TensorRT 7.2.1. For a list of the new features and enhancements introduced in TensorRT 7.2.1 refer to the TensorRT 7.2.1 release notes.
- The latest version of NVIDIA NCCL 2.8.2
- Ubuntu 18.04 with October 2020 updates

Announcements

- Python 2.7 is no longer supported in this TensorRT container release.

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the

missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 20.11 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information see GitHub: TensorRT 20.11.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.11	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.10			
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

There are no known issues in this release.

Chapter 49. TensorRT Release 20.10

The NVIDIA container image for TensorRT, release 20.10, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 7.2.1. Note that the ONNX parser and plugin libraries bundled with this container are built from TensorRT Open Source Software: https://github.com/NVIDIA/TensorRT/releases/tag/20.10

The container also includes the following:

- Ubuntu 18.04

  Note: Container image `20.10-py3` contains Python 3.6.

- NVIDIA CUDA 11.1.0 including cuBLAS 11.2.1.
- NVIDIA cuDNN 8.0.4
- NVIDIA NCCL 2.7.8 (optimized for NVLink™)

  Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- MLNX_OFED

- OpenMPI 3.1.6
- Nsight Compute 2020.2.0.18
- Nsight Systems 2020.3.4.32

Driver Requirements

Release 20.10 is based on NVIDIA CUDA 11.1.0, which requires NVIDIA Driver release 455 or later. However, if you are running on Tesla (for example, T4 or any other Tesla board), you may use NVIDIA driver release 418.xx, 440.30, or 450.xx. The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 20.10 supports CUDA compute capability 3.5 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and Ampere Architecture GPU families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 20.10 is based on TensorRT 7.2.1. For a list of the new features and enhancements introduced in TensorRT 7.2.1 refer to the TensorRT 7.2.1 release notes.
- The latest version of NVIDIA CUDA 11.1.0 including cuBLAS 11.2.1
- The latest version of NVIDIA cuDNN 8.0.4
- The latest version of Nsight Compute 2020.2.0.18
- The latest version of Nsight Systems 2020.3.4.32
- Ubuntu 18.04 with September 2020 updates

Announcements

- Python 2.7 is no longer supported in this TensorRT container release.

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 20.10 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information see GitHub: TensorRT 20.10.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.10	18.04	NVIDIA CUDA 11.1.0	TensorRT 7.2.1
20.09		NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03 20.02 20.01		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	

Container Version	Ubuntu	CUDA Toolkit	TensorRT
19.09			
19.08			TensorRT 5.1.5

Known Issues

There are no known issues in this release.

Chapter 50. TensorRT Release 20.09

The NVIDIA container image for TensorRT, release 20.09, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 7.1.3. Note that the ONNX parser and plugin libraries bundled with this container are built from TensorRT Open Source Software: https://github.com/NVIDIA/TensorRT/releases/tag/20.09

The container also includes the following:

- Ubuntu 18.04

  Note: Container image `20.09-py3` contains Python 3.6.

- NVIDIA CUDA 11.0.3 including cuBLAS 11.2.0.
- NVIDIA cuDNN 8.0.4
- NVIDIA NCCL 2.7.8 (optimized for NVLink™)

  Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- MLNX_OFED

- OpenMPI 3.1.6
- Nsight Compute 2020.1.2.4
- Nsight Systems 2020.3.2.6

Driver Requirements

Release 20.09 is based on NVIDIA CUDA 11.0.3, which requires NVIDIA Driver release 450 or later. However, if you are running on Tesla (for example, T4 or any other Tesla board), you may use NVIDIA driver release 418.xx or 440.30. The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 20.09 supports CUDA compute capability 3.5 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and Ampere Architecture GPU families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 20.09 is based on TensorRT 7.1.3.
- The latest version of NVIDIA cuDNN 8.0.4
- Ubuntu 18.04 with August 2020 updates

Announcements

- Python 2.7 is no longer supported in this TensorRT container release.

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 20.09 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information see GitHub: TensorRT 20.09.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.09	18.04	NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.08			
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
20.02			
20.01			
19.12			TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

There are no known issues in this release.

Chapter 51. TensorRT Release 20.08

The NVIDIA container image for TensorRT, release 20.08, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 7.1.3. Note that the ONNX parser and plugin libraries bundled with this container are built from TensorRT Open Source Software: https://github.com/NVIDIA/TensorRT/releases/tag/20.08

The container also includes the following:

- Ubuntu 18.04

  Note: Container image `20.08-py3` contains Python 3.6.

- NVIDIA CUDA 11.0.3 including cuBLAS 11.2.0.
- NVIDIA cuDNN 8.0.2
- NVIDIA NCCL 2.7.8 (optimized for NVLink™)

  Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- MLNX_OFED

- OpenMPI 3.1.6
- Nsight Compute 2020.1.2.4
- Nsight Systems 2020.3.2.6

Driver Requirements

Release 20.08 is based on NVIDIA CUDA 11.0.3, which requires NVIDIA Driver release 450 or later. However, if you are running on Tesla (for example, T4 or any other Tesla board), you may use NVIDIA driver release 418.xx or 440.30. The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 20.08 supports CUDA compute capability 3.5 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and Ampere Architecture GPU families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 20.08 is based on TensorRT 7.1.3.
- The latest version of NVIDIA CUDA 11.0.3 including cuBLAS 11.2.0.
- The latest version of NVIDIA cuDNN 8.0.2
- The latest version of NVIDIA NCCL 2.7.8
- The latest version of Nsight Compute 2020.1.2.4
- Ubuntu 18.04 with July 2020 updates

Announcements

- Python 2.7 is no longer supported in this TensorRT container release.

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the

missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 20.08 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information see GitHub: TensorRT 20.08.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.08	18.04	NVIDIA CUDA 11.0.3	TensorRT 7.1.3
20.07		NVIDIA CUDA 11.0.194	
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03 20.02 20.01		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

There are no known issues in this release.

Chapter 52. TensorRT Release 20.07

The NVIDIA container image for TensorRT, release 20.07, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 7.1.3.4 release.

The container also includes the following:

- Ubuntu 18.04

  Note: Container image `20.07-py3` contains Python 3.6.
- NVIDIA CUDA 11.0.194 including cuBLAS 11.1.0.
- NVIDIA cuDNN 8.0.1
- NVIDIA NCCL 2.7.6 (optimized for NVLink™)

  Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.
- MLNX_OFED
- OpenMPI 3.1.6
- Nsight Compute 2020.1.1.8

- Nsight Systems 2020.3.2.6

Driver Requirements

Release 20.07 is based on NVIDIA CUDA 11.0.194, which requires NVIDIA Driver release 450 or later. However, if you are running on Tesla (for example, T4 or any other Tesla board), you may use NVIDIA driver release 418.xx or 440.30. The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 20.07 supports CUDA compute capability 3.5 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, Turing, and Ampere Architecture GPU families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 20.07 is based on TensorRT 7.1.3.
- The latest version of NVIDIA CUDA 11.0.194 including cuBLAS 11.1.0.
- The latest version of NVIDIA cuDNN 8.0.1
- The latest version of NVIDIA NCCL 2.7.6
- The latest version of TensorRT 7.1.3
- The latest version of Nsight Compute 2020.1.1.8
- The latest version of Nsight Systems 2020.3.2.6
- Ubuntu 18.04 with June 2020 updates

Announcements

- Python 2.7 is no longer supported in this TensorRT container release.

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the

missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 20.07 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following script:

```
/opt/tensorrt/install_opensource.sh
```

For more information see GitHub: TensorRT 20.07.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.07	18.04	NVIDIA CUDA 11.0.194	TensorRT 7.1.3
20.06		NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03 20.02 20.01		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

There are no known issues in this release.

Chapter 53. TensorRT Release 20.06

The NVIDIA container image for TensorRT, release 20.06, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 7.1.2

The container also includes the following:

- Ubuntu 18.04

  Note: Container image `20.06-py3` contains Python 3.6.
- NVIDIA CUDA 11.0.167 including cuBLAS 11.1.0.
- NVIDIA cuDNN 8.0.1
- NVIDIA NCCL 2.7.5 (optimized for NVLink™)

  Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.
- MLNX_OFED
- OpenMPI 3.1.6
- Nsight Compute 2020.1.0.33

- Nsight Systems 2020.2.5.8

Driver Requirements

Release 20.06 is based on NVIDIA CUDA 11.0.167, which requires NVIDIA Driver release 450 or later. However, if you are running on Tesla (for example, T4 or any other Tesla board), you may use NVIDIA driver release 418.xx or 440.30. The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 20.06 supports CUDA compute capability 3.5 and higher. This corresponds to GPUs in the Pascal, Volta, Turing, and Ampere Architecture GPU families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT container release includes the following key features and enhancements.

- TensorRT container image version 20.06 is based on TensorRT 7.1.2.
- The latest version of NVIDIA CUDA 11.0.167 including cuBLAS 11.1.0.
- The latest version of NVIDIA cuDNN 8.0.1
- The latest version of NVIDIA NCCL 2.7.5
- The latest version of TensorRT 7.1.2
- The latest version of Nsight Compute 2020.1.0.33
- The latest version of Nsight Systems 2020.2.5.8
- The latest version of NVIDIA NCCL 2.7.5
- The latest version of OpenMPI 3.1.6
- Ubuntu 18.04 with May 2020 updates

Announcements

- Python 2.7 is no longer supported in this TensorRT container release.

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 20.06 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following commands to install the required prerequisites and run the installation script:

- ```
  apt-get update && apt-get install libcurl4-openssl-dev zlib1g-dev
      pkg-config
  ```
- ```
  curl -L -k -o /opt/cmake-3.14.4-Linux-x86_64.tar.gz
          https://github.com/Kitware/CMake/releases/download/v3.14.4/cmake-3.14.4-
  Linux-x86_64.tar.gz
          && pushd /opt && tar -xzf cmake-3.14.4-Linux-x86_64.tar.gz && rm
          cmake-3.14.4-Linux-x86_64.tar.gz && popd && export
          PATH=/opt/cmake-3.14.4-Linux-x86_64/bin/:$PATH
  ```
- ```
  chmod +x /opt/tensorrt/install_opensource.sh &&
        /opt/tensorrt/install_opensource.sh
  ```

For more information see GitHub: TensorRT 20.06.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.06	18.04	NVIDIA CUDA 11.0.167	TensorRT 7.1.2
20.03 20.02 20.01		NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12 19.11			TensorRT 6.0.1
  ```

Container Version	Ubuntu	CUDA Toolkit	TensorRT
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

There are no known issues in this release.

Chapter 54. TensorRT Release 20.03

The NVIDIA container image for TensorRT, release 20.03, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 7.0.0

The container also includes the following:

- Ubuntu 18.04

  Note: Container image `20.03-py3` contains Python 3.6.

- NVIDIA CUDA 10.2.89 including cuBLAS 10.2.2.89
- NVIDIA cuDNN 7.6.5
- NVIDIA NCCL 2.6.3 (optimized for NVLink™)

  Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- MLNX_OFED
- OpenMPI 3.1.4
- Nsight Compute 2019.5.0

- Nsight Systems 2020.1.1

Driver Requirements

Release 20.03 is based on NVIDIA CUDA 10.2.89, which requires NVIDIA Driver release 440.33.01. However, if you are running on Tesla (for example, T4 or any other Tesla board), you may use NVIDIA driver release 396, 384.111+, 410, 418.xx or 440.30. The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 20.03 supports CUDA compute capability 3.0 and higher. This corresponds to GPUs in the Pascal, Volta, and Turing families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 20.03 is based on TensorRT 7.0.0.
- Ubuntu 18.04 with February 2020 updates

Announcements

- Python 2.7 is no longer supported in this TensorRT container release.

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 20.03 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following commands to install the required prerequisites and run the installation script:

- ```
  apt-get update && apt-get install libcurl4-openssl-dev zlib1g-dev
      pkg-config
  ```
- ```
  curl -L -k -o /opt/cmake-3.14.4-Linux-x86_64.tar.gz
          https://github.com/Kitware/CMake/releases/download/v3.14.4/cmake-3.14.4-
  Linux-x86_64.tar.gz
          && pushd /opt && tar -xzf cmake-3.14.4-Linux-x86_64.tar.gz && rm
          cmake-3.14.4-Linux-x86_64.tar.gz && popd && export
          PATH=/opt/cmake-3.14.4-Linux-x86_64/bin/:$PATH
  ```
- ```
  chmod +x /opt/tensorrt/install_opensource.sh &&
        /opt/tensorrt/install_opensource.sh
  ```

For more information see GitHub: TensorRT 20.03.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.03 20.02 20.01	18.04 16.04	NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

There are no known issues in this release.
  ```

Chapter 55. TensorRT Release 20.02

The NVIDIA container image for TensorRT, release 20.02, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 7.0.0

The container also includes the following:

- Ubuntu 18.04

  Note: Container image `20.02-py3` contains Python 3.6.

- NVIDIA CUDA 10.2.89 including cuBLAS 10.2.2.89
- NVIDIA cuDNN 7.6.5
- NVIDIA NCCL 2.5.6 (optimized for NVLink™)

  Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.

- MLNX_OFED
- OpenMPI 3.1.4
- Nsight Compute 2019.5.0

- Nsight Systems 2020.1.1

Driver Requirements

Release 20.02 is based on NVIDIA CUDA 10.2.89, which requires NVIDIA Driver release 440.33.01. However, if you are running on Tesla (for example, T4 or any other Tesla board), you may use NVIDIA driver release 396, 384.111+, 410, 418.xx or 440.30. The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 20.02 supports CUDA compute capability 3.0 and higher. This corresponds to GPUs in the Pascal, Volta, and Turing families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 20.02 is based on TensorRT 7.0.0.
- The latest version of Nsight Systems 2020.1.1
- Ubuntu 18.04 with January 2020 updates

Announcements

- Python 2.7 is no longer supported in this TensorRT container release.

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 20.02 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following commands to install the required prerequisites and run the installation script:

- ```
  apt-get update && apt-get install libcurl4-openssl-dev zlib1g-dev
      pkg-config
  ```
- ```
  curl -L -k -o /opt/cmake-3.14.4-Linux-x86_64.tar.gz
          https://github.com/Kitware/CMake/releases/download/v3.14.4/cmake-3.14.4-
  Linux-x86_64.tar.gz
          && pushd /opt && tar -xzf cmake-3.14.4-Linux-x86_64.tar.gz && rm
          cmake-3.14.4-Linux-x86_64.tar.gz && popd && export
          PATH=/opt/cmake-3.14.4-Linux-x86_64/bin/:$PATH
  ```
- ```
  chmod +x /opt/tensorrt/install_opensource.sh &&
        /opt/tensorrt/install_opensource.sh
  ```

For more information see GitHub: TensorRT 20.02.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.02 20.01	18.04 16.04	NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12 19.11			TensorRT 6.0.1
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

There are no known issues in this release.
  ```

Chapter 56. TensorRT Release 20.01

The NVIDIA container image for TensorRT, release 20.01, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/cpp` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 7.0.0

The container also includes the following:

- Ubuntu 18.04

  Note: Container image `20.01-py2` contains Python 2.7; `20.01-py3` contains Python 3.6.
- NVIDIA CUDA 10.2.89 including cuBLAS 10.2.2.89
- NVIDIA cuDNN 7.6.5
- NVIDIA NCCL 2.5.6 (optimized for NVLink™)

  Note: Although NCCL is packaged in the container, it does not affect TensorRT nor inferencing in any way.
- MLNX_OFED
- OpenMPI 3.1.4
- Nsight Compute 2019.5.0

- Nsight Systems 2019.6.1

Driver Requirements

Release 20.01 is based on NVIDIA CUDA 10.2.89, which requires NVIDIA Driver release 440.33.01. However, if you are running on Tesla (for example, T4 or any other Tesla board), you may use NVIDIA driver release 396, 384.111+, 410, 418.xx or 440.30. The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 20.01 supports CUDA compute capability 3.0 and higher. This corresponds to GPUs in the Pascal, Volta, and Turing families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 20.01 is based on TensorRT 7.0.0.
- Ubuntu 18.04 with December 2019 updates

Announcements

- We will stop support for Python 2.7 in the next TensorRT container release.

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 20.01 tag on the official TensorRT open source repository.

To install the open source components inside the container, run the following commands to install the required prerequisites and run the installation script:

- ```
  apt-get update && apt-get install libcurl4-openssl-dev zlib1g-dev
      pkg-config
  ```
- ```
  curl -L -k -o /opt/cmake-3.14.4-Linux-x86_64.tar.gz
          https://github.com/Kitware/CMake/releases/download/v3.14.4/cmake-3.14.4-
  Linux-x86_64.tar.gz
          && pushd /opt && tar -xzf cmake-3.14.4-Linux-x86_64.tar.gz && rm
          cmake-3.14.4-Linux-x86_64.tar.gz && popd && export
          PATH=/opt/cmake-3.14.4-Linux-x86_64/bin/:$PATH
  ```
- ```
  chmod +x /opt/tensorrt/install_opensource.sh &&
        /opt/tensorrt/install_opensource.sh
  ```

For more information see GitHub: TensorRT 20.01.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

NVIDIA TensorRT Container Versions

The following table shows what versions of Ubuntu, CUDA, and TensorRT are supported in each of the NVIDIA containers for TensorRT. For older container versions, refer to the Frameworks Support Matrix.

Container Version	Ubuntu	CUDA Toolkit	TensorRT
20.01	18.04	NVIDIA CUDA 10.2.89	TensorRT 7.0.0
19.12	16.04		TensorRT 6.0.1
19.11			
19.10		NVIDIA CUDA 10.1.243	
19.09			
19.08			TensorRT 5.1.5

Known Issues

There are no known issues in this release.
  ```

Chapter 57. TensorRT Release 19.12

The NVIDIA container image for TensorRT, release 19.12, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/html` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 6.0.1

The container also includes the following:

- Ubuntu 18.04

  Note: Container image `19.12-py2` contains Python 2.7; `19.12-py3` contains Python 3.6.
- NVIDIA CUDA 10.2.89 including cuBLAS 10.2.2.89
- NVIDIA cuDNN 7.6.5
- NVIDIA NCCL 2.5.6 (optimized for NVLink™)

  Note: Although NCCL is packaged in the container, it does not effect TensorRT nor inferencing in any way.
- MLNX_OFED
- OpenMPI 3.1.4
- Nsight Compute 2019.5.0

- Nsight Systems 2019.6.1

Driver Requirements

Release 19.12 is based on NVIDIA CUDA 10.2.89, which requires NVIDIA Driver release 440.33.01. However, if you are running on Tesla (for example, T4 or any other Tesla board), you may use NVIDIA driver release 396, 384.111+, 410, 418.xx or 440.30. The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 19.12 supports CUDA compute capability 3.0 and higher. This corresponds to GPUs in the Pascal, Volta, and Turing families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 19.12 is based on TensorRT 6.0.1.
- Latest version of Nsight Systems 2019.6.1
- Ubuntu 18.04 with November 2019 updates

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Installing Open Source Components

A script has been added to clone, build and replace the provided plugin, Caffe parser, and ONNX parser libraries with the open source ones based off the 19.12 tag on the official TensorRT open source repository. The script is found at `/opt/tensorrt/install_opensource.sh`. For more information, and for instructions to build the open source samples, see GitHub: TensorRT 19.12.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

Known Issues

There are no known issues in this release.

Chapter 58. TensorRT Release 19.11

The NVIDIA container image for TensorRT, release 19.11, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/html` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 6.0.1

The container also includes the following:

- Ubuntu 18.04

  Note: Container image `19.11-py2` contains Python 2.7; `19.11-py3` contains Python 3.6.
- NVIDIA CUDA 10.2.89 including cuBLAS 10.2.2.89
- NVIDIA cuDNN 7.6.5
- NVIDIA NCCL 2.5.6 (optimized for NVLink™)

  Note: Although NCCL is packaged in the container, it does not effect TensorRT nor inferencing in any way.
- MLNX_OFED
- OpenMPI 3.1.4
- Nsight Compute 2019.5.0

- Nsight Systems 2019.5.2

Driver Requirements

Release 19.11 is based on NVIDIA CUDA 10.2.89, which requires NVIDIA Driver release 440.30. However, if you are running on Tesla (for example, T4 or any other Tesla board), you may use NVIDIA driver release 396, 384.111+, 410 or 418.xx. The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 19.11 supports CUDA compute capability 3.0 and higher. This corresponds to GPUs in the Pascal, Volta, and Turing families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 19.11 is based on TensorRT 6.0.1.
- Latest version of NVIDIA CUDA 10.2.89 including cuBLAS 10.2.2.89
- Latest version of NVIDIA cuDNN 7.6.5
- Latest version of NVIDIA NCCL 2.5.6
- Latest version of Nsight Compute 2019.5.0
- Latest version of Nsight Systems 2019.5.2
- Ubuntu 18.04 with October 2019 updates

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

Known Issues

There are no known issues in this release.

Chapter 59. TensorRT Release 19.10

The NVIDIA container image for TensorRT, release 19.10, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/html` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 6.0.1

The container also includes the following:

- Ubuntu 18.04

  Note: Container image `19.10-py2` contains Python 2.7; `19.10-py3` contains Python 3.6.
- NVIDIA CUDA 10.1.243 including cuBLAS 10.2.1.243
- NVIDIA cuDNN 7.6.4
- NVIDIA NCCL 2.4.8 (optimized for NVLink™)

  Note: Although NCCL is packaged in the container, it does not effect TensorRT nor inferencing in any way.
- MLNX_OFED
- OpenMPI 3.1.4
- Nsight Compute 2019.4.0

- Nsight Systems 2019.5.1

Driver Requirements

Release 19.10 is based on NVIDIA CUDA 10.1.243, which requires NVIDIA Driver release 418.xx. However, if you are running on Tesla (for example, T4 or any other Tesla board), you may use NVIDIA driver release 396, 384.111+ or 410. The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 19.10 supports CUDA compute capability 3.0 and higher. This corresponds to GPUs in the Pascal, Volta, and Turing families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 19.10 is based on TensorRT 6.0.1.
- Latest version of NVIDIA cuDNN 7.6.4
- Latest versions of Nsight Systems 2019.5.1
- Ubuntu 18.04 with September 2019 updates

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

Known Issues

There are no known issues in this release.

Chapter 60. TensorRT Release 19.09

The NVIDIA container image for TensorRT, release 19.09, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/html` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 6.0.1

The container also includes the following:

- Ubuntu 18.04

  Note: Container image `19.09-py2` contains Python 2.7; `19.09-py3` contains Python 3.6.
- NVIDIA CUDA 10.1.243 including cuBLAS 10.2.1.243
- NVIDIA cuDNN 7.6.3
- NVIDIA NCCL 2.4.8 (optimized for NVLink™)

  Note: Although NCCL is packaged in the container, it does not effect TensorRT nor inferencing in any way.
- MLNX_OFED
- OpenMPI 3.1.4
- Nsight Compute 2019.4.0

- Nsight Systems 2019.4.2

Driver Requirements

Release 19.09 is based on NVIDIA CUDA 10.1.243, which requires NVIDIA Driver release 418.xx. However, if you are running on Tesla (for example, T4 or any other Tesla board), you may use NVIDIA driver release 396, 384.111+ or 410. The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 19.09 supports CUDA compute capability 3.0 and higher. This corresponds to GPUs in the Pascal, Volta, and Turing families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 19.09 is based on TensorRT 6.0.1.
- Latest version of NVIDIA cuDNN 7.6.3
- Latest versions of Nsight Compute 2019.4.0 and Nsight Systems 2019.4.2
- Ubuntu 18.04 with August 2019 updates

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

Known Issues

- There is a known issue when running the `/opt/tensorrt/python/python_setup.sh` script. This scripts does not work due to the UFF converter not supporting TensorFlow version 2.0. To workaround this issue, to install TensorFlow version 1.15 or 1.14. This will be resolved in a future container.

Chapter 61. TensorRT Release 19.08

The NVIDIA container image for TensorRT, release 19.08, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/html` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 5.1.5

The container also includes the following:

- Ubuntu 18.04

  Note: Container image `19.08-py2` contains Python 2.7; `19.08-py3` contains Python 3.6.
- NVIDIA CUDA 10.1.243 including cuBLAS 10.2.1.243
- NVIDIA cuDNN 7.6.2
- NVIDIA NCCL 2.4.8 (optimized for NVLink™)

  Note: Although NCCL is packaged in the container, it does not effect TensorRT nor inferencing in any way.
- MLNX_OFED +4.0
- OpenMPI 3.1.4
- Nsight Compute 10.1.168

- Nsight Systems 2019.3.7.9

Driver Requirements

Release 19.08 is based on NVIDIA CUDA 10.1.243, which requires NVIDIA Driver release 418.87. However, if you are running on Tesla (Tesla V100, Tesla P4, Tesla P40, or Tesla P100), you may use NVIDIA driver release 384.111+ or 410. The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 19.08 supports CUDA compute capability 3.0 and higher. This corresponds to GPUs in the Pascal, Volta, and Turing families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 19.08 is based on TensorRT 5.1.5.
- Upgraded to Python 3.6.
- Latest version of NVIDIA CUDA 10.1.243 including cuBLAS 10.2.1.243
- Latest version of NVIDIA cuDNN 7.6.2
- Latest version of NVIDIA NCCL 2.4.8
- Latest version of MLNX_OFED +4.0
- Latest version of OpenMPI 3.1.4
- Latest version of Nsight Systems 2019.3.7.9
- Ubuntu 18.04 with July 2019 updates

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

Known Issues

There are no known issues in this release.

Chapter 62. TensorRT Release 19.07

The NVIDIA container image for TensorRT, release 19.07, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/html` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/data`. For example:
  ```
  python caffe_resnet50.py -d /workspace/tensorrt/data
  ```
  The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 5.1.5

The container also includes the following:

- Ubuntu 18.04

  Note: Container image `19.07-py2` contains Python 2.7; `19.07-py3` contains Python 3.5.
- NVIDIA CUDA 10.1.168 including cuBLAS 10.2.0.168
- NVIDIA cuDNN 7.6.1
- NVIDIA NCCL 2.4.7 (optimized for NVLink™)

  Note: Although NCCL is packaged in the container, it does not effect TensorRT nor inferencing in any way.
- MLNX_OFED +3.4
- OpenMPI 3.1.3

Driver Requirements

Release 19.07 is based on NVIDIA CUDA 10.1.168, which requires NVIDIA Driver release 418.67. However, if you are running on Tesla (Tesla V100, Tesla P4, Tesla P40, or Tesla P100), you may use NVIDIA driver release 384.111+ or 410. The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 19.07 supports CUDA compute capability 3.0 and higher. This corresponds to GPUs in the Pascal, Volta, and Turing families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 19.07 is based on TensorRT 5.1.5.
- Latest version of NVIDIA cuDNN 7.6.1
- Latest version of MLNX_OFED +3.4
- Latest version of Ubuntu 18.04

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

Known Issues

There are no known issues in this release.

Chapter 63. TensorRT Release 19.06

The NVIDIA container image for TensorRT, release 19.06, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/html` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/python/data`. The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 5.1.5

The container also includes the following:

- Ubuntu 16.04

> Note: Container image `19.06-py2` contains Python 2.7; `19.06-py3` contains Python 3.5.

- NVIDIA CUDA 10.1.168 including cuBLAS 10.2.0.168
- NVIDIA cuDNN 7.6.0
- NVIDIA NCCL 2.4.7 (optimized for NVLink™)

> Note: Although NCCL is packaged in the container, it does not effect TensorRT nor inferencing in any way.

- OpenMPI 3.1.3

Driver Requirements

Release 19.06 is based on NVIDIA CUDA 10.1.168, which requires NVIDIA Driver release 418.xx. However, if you are running on Tesla (Tesla V100, Tesla P4, Tesla P40,

or Tesla P100), you may use NVIDIA driver release 384.111+ or 410. The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 19.06 supports CUDA compute capability 3.0 and higher. This corresponds to GPUs in the Pascal, Volta, and Turing families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 19.06 is based on TensorRT 5.1.5.
- Latest version of NVIDIA CUDA 10.1.168 including cuBLAS 10.2.0.168
- Latest version of NVIDIA NCCL 2.4.7
- Ubuntu 16.04 with May 2019 updates (see Announcements)

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Announcements

In the next release, we will no longer support Ubuntu 16.04. Release 19.07 will instead support Ubuntu 18.04.

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

Known Issues

There are no known issues in this release.

Chapter 64. TensorRT Release 19.05

The NVIDIA container image for TensorRT, release 19.05, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/html` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/python/data`. The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 5.1.5

The container also includes the following:

- Ubuntu 16.04

> Note: Container image `19.05-py2` contains Python 2.7; `19.05-py3` contains Python 3.5.

- NVIDIA CUDA 10.1 Update 1 including cuBLAS 10.1 Update 1
- NVIDIA cuDNN 7.6.0
- NVIDIA NCCL 2.4.6 (optimized for NVLink™)

> Note: Although NCCL is packaged in the container, it does not effect TensorRT nor inferencing in any way.

- OpenMPI 3.1.3

Driver Requirements

Release 19.05 is based on CUDA 10.1 Update 1, which requires NVIDIA Driver release 418.xx. However, if you are running on Tesla (Tesla V100, Tesla P4, Tesla P40, or

Tesla P100), you may use NVIDIA driver release 384.111+ or 410. The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 19.05 supports CUDA compute capability 3.0 and higher. This corresponds to GPUs in the Pascal, Volta, and Turing families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 19.05 is based on TensorRT 5.1.5.
- Latest version of NVIDIA CUDA 10.1 Update 1 including cuBLAS 10.1 Update 1
- Latest version of NVIDIA cuDNN 7.6.0
- Latest version of TensorRT 5.1.5
- Ubuntu 16.04 with April 2019 updates

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

Known Issues

There are no known issues in this release.

Chapter 65. TensorRT Release 19.04

The NVIDIA container image for TensorRT, release 19.04, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/html` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/python/data`. The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 5.1.2

The container also includes the following:

- Ubuntu 16.04

> Note: Container image `19.04-py2` contains Python 2.7; `19.04-py3` contains Python 3.5.

- NVIDIA CUDA 10.1.105 including cuBLAS 10.1.0.105
- NVIDIA cuDNN 7.5.0
- NVIDIA NCCL 2.4.6 (optimized for NVLink™)

> Note: Although NCCL is packaged in the container, it does not effect TensorRT nor inferencing in any way.

- OpenMPI 3.1.3

Driver Requirements

Release 19.04 is based on CUDA 10.1, which requires NVIDIA Driver release 418.xx.x+. However, if you are running on Tesla (Tesla V100, Tesla P4, Tesla P40, or Tesla P100),

you may use NVIDIA driver release 384.111+ or 410. The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 19.04 supports CUDA compute capability 3.0 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, and Turing families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 19.04 is based on TensorRT 5.1.2 RC.
- Latest version of NVIDIA NCCL 2.4.6
- Latest version of cuBLAS 10.1.0.105
- Ubuntu 16.04 with March 2019 updates

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

Known Issues

There are no known issues in this release.

Chapter 66. TensorRT Release 19.03

The NVIDIA container image for TensorRT, release 19.03, is available on NGC.

Contents of the TensorRT container

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/html` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/python/data`. The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.
- TensorRT 5.1.2

The container also includes the following:

- Ubuntu 16.04

> Note: Container image `19.03-py2` contains Python 2.7; `19.03-py3` contains Python 3.5.

- NVIDIA CUDA 10.1.105 including cuBLAS 10.1.105
- NVIDIA cuDNN 7.5.0
- NVIDIA NCCL 2.4.3 (optimized for NVLink™)

> Note: Although NCCL is packaged in the container, it does not effect TensorRT nor inferencing in any way.

- OpenMPI 3.1.3

Driver Requirements

Release 19.03 is based on CUDA 10.1, which requires NVIDIA Driver release 418.xx+. However, if you are running on Tesla (Tesla V100, Tesla P4, Tesla P40, or Tesla P100),

you may use NVIDIA driver release 384.111+ or 410. The CUDA driver's compatibility package only supports particular drivers. For a complete list of supported drivers, see the CUDA Application Compatibility topic. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 19.03 supports CUDA compute capability 3.0 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, and Turing families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 19.03 is based on TensorRT 5.1.2 RC.
- Latest version of NVIDIA CUDA 10.1.105 including cuBLAS 10.1.105
- Latest version of NVIDIA cuDNN 7.5.0
- Latest version of NVIDIA NCCL 2.4.3
- Ubuntu 16.04 with February 2019 updates

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

Known Issues

- If using or upgrading to a 3-part-version driver, for example, a driver that takes the format of `xxx.yy.zz`, you will receive a `Failed to detect NVIDIA driver version.` message. This is due to a known bug in the entry point script's parsing of 3-part

driver versions. This message is non-fatal and can be ignored. This will be fixed in the 19.04 release.

Chapter 67. TensorRT Release 19.02

The NVIDIA container image for TensorRT, release 19.02, is available.

Contents of TensorRT

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/html` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/python/data`. The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.

The container also includes the following:

- Ubuntu 16.04

> Note: Container image `19.02-py2` contains Python 2.7; `19.02-py3` contains Python 3.5.

- NVIDIA CUDA 10.0.130 including CUDA® Basic Linear Algebra Subroutines library™ (cuBLAS) 10.0.130
- NVIDIA CUDA® Deep Neural Network library™ (cuDNN) 7.4.2
- NVIDIA Collective Communications Library (NCCL) 2.3.7 (optimized for NVLink™)

> Note: Although NCCL is packaged in the container, it does not effect TensorRT nor inferencing in any way.

- OpenMPI 3.1.3

Driver Requirements

Release 19.02 is based on CUDA 10, which requires NVIDIA Driver release 410.xx. However, if you are running on Tesla (Tesla V100, Tesla P4, Tesla P40, or Tesla P100), you

may use NVIDIA driver release 384. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 19.02 supports CUDA compute capability 3.0 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, and Turing families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 19.02 is based on TensorRT 5.0.2.
- Ubuntu 16.04 with January 2019 updates

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.txt` or `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

Known Issues

- If using or upgrading to a 3-part-version driver, for example, a driver that takes the format of `xxx.yy.zz`, you will receive a `Failed to detect NVIDIA driver version.` message. This is due to a known bug in the entry point script's parsing of 3-part driver versions. This message is non-fatal and can be ignored. This will be fixed in the 19.04 release.

Chapter 68. TensorRT Release 19.01

The NVIDIA container image for TensorRT, release 19.01, is available.

Contents of TensorRT

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/html` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/python/data`. The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.

The container also includes the following:

- Ubuntu 16.04

> Note: Container image `19.01-py2` contains Python 2.7; `19.01-py3` contains Python 3.5.

- NVIDIA CUDA 10.0.130 including CUDA® Basic Linear Algebra Subroutines library™ (cuBLAS) 10.0.130
- NVIDIA CUDA® Deep Neural Network library™ (cuDNN) 7.4.2
- NCCL 2.3.7 (optimized for NVLink™)

> Note: Although NCCL is packaged in the container, it does not effect TensorRT nor inferencing in any way.

- OpenMPI 3.1.3

Driver Requirements

Release 19.01 is based on CUDA 10, which requires NVIDIA Driver release 410.xx. However, if you are running on Tesla (Tesla V100, Tesla P4, Tesla P40, or Tesla P100), you

may use NVIDIA driver release 384. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 19.01 supports CUDA compute capability 3.0 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, and Turing families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 19.01 is based on TensorRT 5.0.2.
- Latest version of OpenMPI 3.1.3
- Ubuntu 16.04 with December 2018 updates

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.txt` or `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Limitations

- Accelerating Inference In TensorFlow with TensorRT (TF-TRT) is not supported in the TensorRT containers.

Known Issues

- If using or upgrading to a 3-part-version driver, for example, a driver that takes the format of `xxx.yy.zz`, you will receive a `Failed to detect NVIDIA driver version.` message. This is due to a known bug in the entry point script's parsing of 3-part driver versions. This message is non-fatal and can be ignored. This will be fixed in the 19.04 release.

Chapter 69. TensorRT Release 18.12

The NVIDIA container image for TensorRT, release 18.12, is available.

Contents of TensorRT

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/html` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/python/data`. The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.

The container also includes the following:

- Ubuntu 16.04

> Note: Container image `18.12-py2` contains Python 2.7; `18.12-py3` contains Python 3.5.

- NVIDIA CUDA 10.0.130 including CUDA® Basic Linear Algebra Subroutines library™ (cuBLAS) 10.0.130
- NVIDIA CUDA® Deep Neural Network library™ (cuDNN) 7.4.1
- NCCL 2.3.7 (optimized for NVLink™)

> Note: Although NCCL is packaged in the container, it does not effect TensorRT nor inferencing in any way.

- OpenMPI 3.1.2

Driver Requirements

Release 18.12 is based on CUDA 10, which requires NVIDIA Driver release 410.xx. However, if you are running on Tesla (Tesla V100, Tesla P4, Tesla P40, or Tesla P100), you

may use NVIDIA driver release 384. For more information, see CUDA Compatibility and Upgrades.

GPU Requirements

Release 18.12 supports CUDA compute capability 3.0 and higher. This corresponds to GPUs in the Kepler, Maxwell, Pascal, Volta, and Turing families. Specifically, for a list of GPUs that this compute capability corresponds to, see CUDA GPUs. For additional support details, see Deep Learning Frameworks Support Matrix.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 18.12 is based on TensorRT 5.0.2.
- Ubuntu 16.04 with November 2018 updates

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.txt` or `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Known Issues

There are no known issues in this release.

Chapter 70. TensorRT Release 18.11

The NVIDIA container image for TensorRT, release 18.11, is available.

Contents of TensorRT

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/html` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/python/data`. The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.

The container also includes the following:

- Ubuntu 16.04

> Note: Container image `18.11-py2` contains Python 2.7; `18.11-py3` contains Python 3.5.

- NVIDIA CUDA 10.0.130 including CUDA® Basic Linear Algebra Subroutines library™ (cuBLAS) 10.0.130
- NVIDIA CUDA® Deep Neural Network library™ (cuDNN) 7.4.1
- NCCL 2.3.7 (optimized for NVLink™)
- OpenMPI 3.1.2

Driver Requirements

Release 18.11 is based on CUDA 10, which requires NVIDIA Driver release 410.xx. However, if you are running on Tesla (Tesla V100, Tesla P4, Tesla P40, or Tesla P100), you may use NVIDIA driver release 384. For more information, see CUDA Compatibility and Upgrades.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 18.11 is based on TensorRT 5.0.2.
- Latest version of NCCL 2.3.7.
- Latest version of NVIDIA cuDNN 7.4.1.
- Ubuntu 16.04 with October 2018 updates

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.txt` or `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install the missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Known Issues

There are no known issues in this release.

Chapter 71. TensorRT Release 18.10

The NVIDIA container image of TensorRT, release 18.10, is available.

Contents of TensorRT

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/html` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/python/data`. The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.

The container also includes the following:

- Ubuntu 16.04

> Note: Container image `18.10-py2` contains Python 2.7; `18.10-py3` contains Python 3.5.

- NVIDIA CUDA 10.0.130 including CUDA® Basic Linear Algebra Subroutines library™ (cuBLAS) 10.0.130
- NVIDIA CUDA® Deep Neural Network library™ (cuDNN) 7.4.0
- NCCL 2.3.6 (optimized for NVLink™)
- OpenMPI 3.1.2

Driver Requirements

Release 18.10 is based on CUDA 10, which requires NVIDIA Driver release 410.xx. However, if you are running on Tesla (Tesla V100, Tesla P4, Tesla P40, or Tesla P100), you may use NVIDIA driver release 384. For more information, see CUDA Compatibility and Upgrades.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 18.10 is based on TensorRT 5.0.0 RC.
- Latest version of NCCL 2.3.6.
- Added support for OpenMPI 3.1.2.
- Ubuntu 16.04 with September 2018 updates

Obtaining Missing Data Files

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. Samples which do not include all the required data files include a `README.txt` or `README.md` file in the corresponding source directory informing you how to obtain the necessary data files.

Installing Required Python Modules

You may need to first run the Python setup script in order to complete some of the samples. The following script has been added to the container to install missing Python modules and their dependencies if desired: `/opt/tensorrt/python/python_setup.sh`

Known Issues

There are no known issues in this release.

Chapter 72. TensorRT Release 18.09

The NVIDIA container image of TensorRT, release 18.09, is available.

Contents of TensorRT

This container includes the following:

- The TensorRT C++ samples and C++ API documentation. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory. The C++ API documentation can be found in the `/workspace/tensorrt/doc/html` directory.
- The TensorRT Python samples and Python API documentation. The Python samples can be found in the `/workspace/tensorrt/samples/python` directory. Many Python samples can be run using `python <script.py> -d /workspace/tensorrt/python/data`. The Python API documentation can be found in the `/workspace/tensorrt/doc/python` directory.

The container also includes the following:

- Ubuntu 16.04

> Note: Container image `18.09-py2` contains Python 2.7; `18.09-py3` contains Python 3.5.

- NVIDIA CUDA 10.0.130 including CUDA® Basic Linear Algebra Subroutines library™ (cuBLAS) 10.0.130
- NVIDIA CUDA® Deep Neural Network library™ (cuDNN) 7.3.0
- NCCL 2.3.4 (optimized for NVLink™)

Driver Requirements

Release 18.09 is based on CUDA 10, which requires NVIDIA Driver release 410.xx. However, if you are running on Tesla (Tesla V100, Tesla P4, Tesla P40, or Tesla P100), you may use NVIDIA driver release 384. For more information, see CUDA Compatibility and Upgrades.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 18.09 is based on TensorRT 5.0.0 RC.
- Latest version of cuDNN 7.3.0.
- Latest version of CUDA 10.0.130 which includes support for DGX-2, Turing, and Jetson Xavier.
- Latest version of cuBLAS 10.0.130.
- Latest version of NCCL 2.3.4.
- Ubuntu 16.04 with August 2018 updates

Installing Required Python Modules

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. The following script has been added to the container to install these missing Python modules and their dependencies if desired: **`/opt/tensorrt/python/python_setup.sh`**

Samples which do not include all the required data files include a `README.txt` file in the corresponding source directory informing you how to obtain the necessary data files. You may need to first run the Python setup script in order to complete some of the samples.

Known Issues

The TensorRT Release Notes (`TensorRT-Release-Notes.pdf`) is missing from the container. Refer to the online TensorRT Release Notes instead.

Chapter 73. TensorRT Release 18.08

The NVIDIA container image of TensorRT, release 18.08, is available.

Contents of TensorRT

This container includes the following:

- The TensorRT documentation and C++ samples. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory.
- The TensorRT Python examples. The Python examples can be found in the `/workspace/tensorrt/python/examples` directory. Most Python examples can be run using `python <script.py> /workspace/tensorrt/python/data`. The Python API documentation can be found in the `/usr/lib/python<x.y>/dist-packages/docs` directory.

The container also includes the following:

- Ubuntu 16.04

> Note: Container image `18.08-py2` contains Python 2.7; `18.08-py3` contains Python 3.5.

- NVIDIA CUDA 9.0.176 (see Errata section and 2.1) including CUDA® Basic Linear Algebra Subroutines library™ (cuBLAS) 9.0.425
- NVIDIA CUDA® Deep Neural Network library™ (cuDNN) 7.2.1
- NCCL 2.2.13 (optimized for NVLink™)

Driver Requirements

Release 18.08 is based on CUDA 9, which requires NVIDIA Driver release 384.xx.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 18.08 is based on TensorRT 4.0.1.
- Latest version of cuDNN 7.2.1.

- A new script has been added to the container that will install `uff`, `graphsurgeon`, as well as other Python modules that are required to execute all of the Python examples.
- Ubuntu 16.04 with July 2018 updates

Installing Required Python Modules

Some samples require data files that are not included within the TensorRT container either due to licensing restrictions or because they are too large. The following script has been added to the container to install these missing Python modules and their dependencies if desired: **`/opt/tensorrt/python/python_setup.sh`**

Samples which do not include all the required data files include a `README.txt` file in the corresponding source directory informing you how to obtain the necessary data files. You may need to first run the Python setup script in order to complete some of the samples.

Known Issues

There are no known issues in this release.

Chapter 74. TensorRT Release 18.07

The NVIDIA container image of TensorRT, release 18.07, is available.

Contents of TensorRT

This container includes the following:

- The TensorRT documentation and C++ samples. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory.
- The TensorRT Python examples. The Python examples can be found in the `/workspace/tensorrt/python/examples` directory. Most Python examples can be run using `python <script.py> /workspace/tensorrt/python/data`. The Python API documentation can be found in the `/usr/lib/python2.7/dist-packages/docs` directory.

The container also includes the following:

- Ubuntu 16.04

> Note: Container image `18.07-py2` contains Python 2.7; `18.07-py3` contains Python 3.5.

- NVIDIA CUDA 9.0.176 (see Errata section and 2.1) including CUDA® Basic Linear Algebra Subroutines library™ (cuBLAS) 9.0.425
- NVIDIA CUDA® Deep Neural Network library™ (cuDNN) 7.1.4
- NCCL 2.2.13 (optimized for NVLink™)

Driver Requirements

Release 18.07 is based on CUDA 9, which requires NVIDIA Driver release 384.xx.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 18.07 is based on TensorRT 4.0.1.
- Latest version of CUDA® Basic Linear Algebra Subroutines library™ (cuBLAS) 9.0.425.

- Ubuntu 16.04 with June 2018 updates

Known Issues

Some samples require data files that are not included within the TensorRT container either due to licensing concerns or because they are too large. Samples which do not include all the required data files instead include a `README.txt` file in the corresponding source directory informing you how to obtain the necessary data files. The data files required for the samples `sampleNMT` and `sampleUffSSD` cannot be easily created within the TensorRT container using the default packages. You should instead prepare the data files for these samples outside the container and then use `docker cp` to copy the necessary files into the TensorRT container or use a mount point when running the TensorRT container.

Chapter 75. TensorRT Release 18.06

The NVIDIA container image of TensorRT, release 18.06, is available.

Contents of TensorRT

This container includes the following:

- The TensorRT documentation and C++ samples. The samples can be built by running `make` in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory.
- The TensorRT Python examples. The Python examples can be found in the `/workspace/tensorrt/python/examples` directory. Most Python examples can be run using `python <script.py> /workspace/tensorrt/python/data`. The Python API documentation can be found in the `/usr/lib/python2.7/dist-packages/docs` directory.

The container also includes the following:

- Ubuntu 16.04

> Note: Container image `18.06-py2` contains Python 2.7; `18.06-py3` contains Python 3.5.

- NVIDIA CUDA 9.0.176 (see Errata section and 2.1) including CUDA® Basic Linear Algebra Subroutines library™ (cuBLAS) 9.0.333 (see section 2.3.1)
- NVIDIA CUDA® Deep Neural Network library™ (cuDNN) 7.1.4
- NCCL 2.2.13 (optimized for NVLink™)

Driver Requirements

Release 18.06 is based on CUDA 9, which requires NVIDIA Driver release 384.xx.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 18.06 is based on TensorRT 4.0.1.
- Ubuntu 16.04 with May 2018 updates

Known Issues

Some samples require data files that are not included within the TensorRT container either due to licensing concerns or because they are too large. Samples which do not include all the required data files instead include a `README.txt` file in the corresponding source directory informing you how to obtain the necessary data files. The data files required for the samples `sampleNMT` and `sampleUffSSD` cannot be easily created within the TensorRT container using the default packages. You should instead prepare the data files for these samples outside the container and then use `docker cp` to copy the necessary files into the TensorRT container or use a mount point when running the TensorRT container.

Chapter 76. TensorRT Release 18.05

The NVIDIA container image of TensorRT, release 18.05, is available.

Contents of TensorRT

This container image contains an example deployment strategy using TensorRT inference exposed via a REST server. Three trained models, NVCaffe, ONNX and TensorFlow, are included to demonstrate the inference REST server. You can also perform inference using your own NVCaffe, ONNX and TensorFlow models via the REST server.

This container also include the following:

- The TensorRT documentation and samples. The samples can be built by running make in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory.
- The example NVCaffe MNIST model and the `caffe_mnist` script are located in the `/workspace/tensorrt_server` directory. The script runs the REST server to provide inference for that model via an HTTP endpoint.
- The example Inception-v1 ONNX model and the `onnx_inception_v1` script are also located in the `/workspace/tensorrt_server` directory. This example and script runs the REST server to provide inference for that model via an HTTP endpoint.
- The example ResNet-152 TensorFlow model and the `tensorflow_resnet` script are also located in the `/workspace/tensorrt_server` directory. This example and script runs the REST server to provide inference for that model via an HTTP endpoint.

The container also includes the following:

- Ubuntu 16.04 including Python 2.7 environment
- NVIDIA CUDA 9.0.176 (see Errata section and 2.1) including CUDA® Basic Linear Algebra Subroutines library™ (cuBLAS) 9.0.333 (see section 2.3.1)
- NVIDIA CUDA® Deep Neural Network library™ (cuDNN) 7.1.2
- NCCL 2.1.15 (optimized for NVLink™)

Driver Requirements

Release 18.05 is based on CUDA 9, which requires NVIDIA Driver release 384.xx.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 18.05 is based on TensorRT 3.0.4.
- Fixed an issue with INT8 deconvolution bias. If you have seen an issue with deconvolution INT8 accuracy especially regarding TensorRT 2.1, then this fix should solve the issue.
- Fixed an accuracy issue in FP16 mode for NVCaffe models.
- Ubuntu 16.04 with April 2018 updates

Known Issues

There are no known issues in this release.

Chapter 77. TensorRT Release 18.04

The NVIDIA container image of TensorRT, release 18.04, is available.

Contents of TensorRT

This container image contains an example deployment strategy using TensorRT inference exposed via a REST server. Three trained models, NVCaffe, ONNX and TensorFlow, are included to demonstrate the inference REST server. You can also perform inference using your own NVCaffe, ONNX and TensorFlow models via the REST server.

This container also include the following:

- The TensorRT documentation and samples. The samples can be built by running make in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory.
- The example NVCaffe MNIST model and the `caffe_mnist` script are located in the `/workspace/tensorrt_server` directory. The script runs the REST server to provide inference for that model via an HTTP endpoint.
- The example Inception-v1 ONNX model and the `onnx_inception_v1` script are also located in the `/workspace/tensorrt_server` directory. This example and script runs the REST server to provide inference for that model via an HTTP endpoint.
- The example ResNet-152 TensorFlow model and the `tensorflow_resnet` script are also located in the `/workspace/tensorrt_server` directory. This example and script runs the REST server to provide inference for that model via an HTTP endpoint.

The container also includes the following:

- Ubuntu 16.04 including Python 2.7 environment
- NVIDIA CUDA 9.0.176 (see Errata section and 2.1) including CUDA® Basic Linear Algebra Subroutines library™ (cuBLAS) 9.0.333 (see section 2.3.1)
- NVIDIA CUDA® Deep Neural Network library™ (cuDNN) 7.1.1
- NCCL 2.1.15 (optimized for NVLink™)

Driver Requirements

Release 18.04 is based on CUDA 9, which requires NVIDIA Driver release 384.xx.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 18.04 is based on TensorRT 3.0.4.
- Fixed an issue with INT8 deconvolution bias. If you have seen an issue with deconvolution INT8 accuracy especially regarding TensorRT. 2.1, then this fix should solve the issue.
- Fixed an accuracy issue in FP16 mode for NVCaffe models.
- Latest version of NCCL 2.1.15
- Ubuntu 16.04 with March 2018 updates

Known Issues

There are no known issues in this release.

Chapter 78. TensorRT Release 18.03

The NVIDIA container image of TensorRT, release 18.03, is available.

Contents of TensorRT

This container image contains an example deployment strategy using TensorRT inference exposed via a REST server. Three trained models, NVCaffe, ONNX and TensorFlow, are included to demonstrate the inference REST server. You can also perform inference using your own NVCaffe, ONNX and TensorFlow models via the REST server.

This container also include the following:

- The TensorRT documentation and samples. The samples can be built by running make in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory.
- The example NVCaffe MNIST model and the `caffe_mnist` script are located in the `/workspace/tensorrt_server` directory. The script runs the REST server to provide inference for that model via an HTTP endpoint.
- The example Inception-v1 ONNX model and the `onnx_inception_v1` script are also located in the `/workspace/tensorrt_server` directory. This example and script runs the REST server to provide inference for that model via an HTTP endpoint.
- The example ResNet-152 TensorFlow model and the `tensorflow_resnet` script are also located in the `/workspace/tensorrt_server` directory. This example and script runs the REST server to provide inference for that model via an HTTP endpoint.

The container also includes the following:

- Ubuntu 16.04 including Python 2.7 environment
- NVIDIA CUDA 9.0.176 (see Errata section and 2.1) including CUDA® Basic Linear Algebra Subroutines library™ (cuBLAS) 9.0.333 (see section 2.3.1)
- NVIDIA CUDA® Deep Neural Network library™ (cuDNN) 7.1.1
- NCCL 2.1.2 (optimized for NVLink™)

Driver Requirements

Release 18.03 is based on CUDA 9, which requires NVIDIA Driver release 384.xx.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- TensorRT container image version 18.03 is based on TensorRT 3.0.4.
- Fixed an issue with INT8 deconvolution bias. If you have seen an issue with deconvolution INT8 accuracy especially regarding TensorRT. 2.1, then this fix should solve the issue.
- Fixed an accuracy issue in FP16 mode for NVCaffe models.
- Latest version of cuBLAS 9.0.333
- Latest version of cuDNN 7.1.1
- Ubuntu 16.04 with February 2018 updates

Known Issues

There are no known issues in this release.

Chapter 79. TensorRT Release 18.02

The NVIDIA container image of TensorRT, release 18.02, is available.

TensorRT container image version 18.02 is based on TensorRT 3.0.4.

Contents of TensorRT

This container image contains an example deployment strategy using TensorRT inference exposed via a REST server. Three trained models, NVCaffe, ONNX and TensorFlow, are included to demonstrate the inference REST server. You can also perform inference using your own NVCaffe, ONNX and TensorFlow models via the REST server.

This container also include the following:

- The TensorRT documentation and samples. The samples can be built by running make in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory.
- The example NVCaffe MNIST model and the `caffe_mnist` script are located in the `/workspace/tensorrt_server` directory. The script runs the REST server to provide inference for that model via an HTTP endpoint.
- The example Inception-v1 ONNX model and the `onnx_inception_v1` script are also located in the `/workspace/tensorrt_server` directory. This example and script runs the REST server to provide inference for that model via an HTTP endpoint.
- The example ResNet-152 TensorFlow model and the `tensorflow_resnet` script are also located in the `/workspace/tensorrt_server` directory. This example and script runs the REST server to provide inference for that model via an HTTP endpoint.

The container also includes the following:

- Ubuntu 16.04 including Python 2.7 environment
- NVIDIA CUDA 9.0.176 including:
  - CUDA® Basic Linear Algebra Subroutines library™ (cuBLAS) 9.0.282 Patch 2 which is installed by default
  - cuBLAS 9.0.234 Patch 1 as a debian file. Installing Patch 1 by issuing the `dpkg -i /opt/cuda-cublas-9-0_9.0.234-1_amd64.deb` command is the workaround for the known issue described below.

- NVIDIA CUDA® Deep Neural Network library™ (cuDNN) 7.0.5
- NCCL 2.1.2 (optimized for NVLink™)

Driver Requirements

Release 18.02 is based on CUDA 9, which requires NVIDIA Driver release 384.xx.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- Latest version of cuBLAS
- Ubuntu 16.04 with January 2018 updates

Known Issues

cuBLAS 9.0.282 regresses RNN seq2seq FP16 performance for a small subset of input sizes. This issue should be fixed in the next update. As a workaround, install cuBLAS 9.0.234 Patch 1 by issuing the `dpkg -i /opt/cuda-cublas-9-0_9.0.234-1_amd64.deb` command.

Chapter 80. TensorRT Release 18.01

The NVIDIA container image of TensorRT, release 18.01, is available.

TensorRT container image version 18.01 is based on TensorRT 3.0.1.

Contents of TensorRT

This container image contains an example deployment strategy using TensorRT inference exposed via a REST server. Three trained models, NVCaffe, ONNX and TensorFlow, are included to demonstrate the inference REST server. You can also perform inference using your own NVCaffe, ONNX and TensorFlow models via the REST server.

This container also include the following:

- The TensorRT documentation and samples. The samples can be built by running make in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory.
- The example NVCaffe MNIST model and the `caffe_mnist` script are located in the `/workspace/tensorrt_server` directory. The script runs the REST server to provide inference for that model via an HTTP endpoint.
- The example Inception-v1 ONNX model and the `onnx_inception_v1` script are also located in the `/workspace/tensorrt_server` directory. This example and script runs the REST server to provide inference for that model via an HTTP endpoint.
- The example ResNet-152 TensorFlow model and the `tensorflow_resnet` script are also located in the `/workspace/tensorrt_server` directory. This example and script runs the REST server to provide inference for that model via an HTTP endpoint.

The container also includes the following:

- Ubuntu 16.04 including Python 2.7 environment
- NVIDIA CUDA 9.0.176 including CUDA® Basic Linear Algebra Subroutines library™ (cuBLAS) 9.0.282
- NVIDIA CUDA® Deep Neural Network library™ (cuDNN) 7.0.5
- NCCL 2.1.2 (optimized for NVLink™)

Driver Requirements

Release 18.01 is based on CUDA 9, which requires NVIDIA Driver release 384.xx.

Key Features and Enhancements

This TensorRT release includes the following key features and enhancements.

- Latest version of cuBLAS
- Latest version of cuDNN
- Latest version of NCCL
- Ubuntu 16.04 with December 2017 updates

Known Issues

cuBLAS 9.0.282 regresses RNN seq2seq FP16 performance for a small subset of input sizes. As a workaround, revert back to the 11.12 container.

Chapter 81. TensorRT Release 17.12

The NVIDIA container image of TensorRT, release 17.12, is available.

Contents of TensorRT

This container image contains an example deployment strategy using TensorRT inference exposed via a REST server. Three trained models, NVCaffe, ONNX and TensorFlow, are included to demonstrate the inference REST server. You can also perform inference using your own NVCaffe, ONNX and TensorFlow models via the REST server.

This container also include the following:

- The TensorRT documentation and samples. The samples can be built by running make in the `/workspace/tensorrt/samples` directory. The resulting executables are in the `/workspace/tensorrt/bin` directory.
- The example NVCaffe MNIST model and the `caffe_mnist` script are located in the `/workspace/tensorrt_server` directory. The script runs the REST server to provide inference for that model via an HTTP endpoint.
- The example Inception-v1 ONNX model and the `onnx_inception_v1` script are also located in the `/workspace/tensorrt_server` directory. This example and script runs the REST server to provide inference for that model via an HTTP endpoint.
- The example ResNet-152 TensorFlow model and the `tensorflow_resnet` script are also located in the `/workspace/tensorrt_server` directory. This example and script runs the REST server to provide inference for that model via an HTTP endpoint.

The container also includes the following:

- Ubuntu 16.04
- NVIDIA CUDA 9.0.176 including CUDA® Basic Linear Algebra Subroutines library™ (cuBLAS) 9.0.234
- NVIDIA CUDA® Deep Neural Network library™ (cuDNN) 7.0.5
- NCCL 2.1.2 (optimized for NVLink™)

Driver Requirements

Release 17.12 is based on CUDA 9, which requires NVIDIA Driver release 384.xx.

Key Features and Enhancements

This is the first TensorRT container release.

Known Issues

There are no known issues in this release.

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.

Arm

Arm, AMBA and Arm Powered are registered trademarks of Arm Limited. Cortex, MPCore and Mali are trademarks of Arm Limited. "Arm" is used to represent Arm Holdings plc; its operating company Arm Limited; and the regional subsidiaries Arm Inc.; Arm KK; Arm Korea Limited.; Arm Taiwan Limited; Arm France SAS; Arm Consulting (Shanghai) Co. Ltd.; Arm Germany GmbH; Arm Embedded Technologies Pvt. Ltd.; Arm Norway, AS and Arm Sweden AB.

HDMI

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

Blackberry/QNX

Copyright © 2020 BlackBerry Limited. All rights reserved.

Trademarks, including but not limited to BLACKBERRY, EMBLEM Design, QNX, AVIAGE, MOMENTICS, NEUTRINO and QNX CAR are the trademarks or registered trademarks of BlackBerry Limited, used under license, and the exclusive rights to such trademarks are expressly reserved.

Google

Android, Android TV, Google Play and the Google Play logo are trademarks of Google, Inc.

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

Trademarks

NVIDIA, the NVIDIA logo, and BlueField, CUDA, DALI, DRIVE, Hopper, JetPack, Jetson AGX Xavier, Jetson Nano, Maxwell, NGC, Nsight, Orin, Pascal, Quadro, Tegra, TensorRT, Triton, Turing and Volta are trademarks and/or registered trademarks of NVIDIA Corporation in the United States and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

© 2017-2024 NVIDIA Corporation & affiliates. All rights reserved.

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