INFERENCE SERVER

RN-08995-001 _v18.10 | October 2018

Release Notes

TABLE OF CONTENTS

Chapter 1. TENSORRT INFERENCE SERVER OVERVIEW

The NVIDIA TensorRT inference server provides a cloud inferencing solution optimized for NVIDIA GPUs.

The TensorRT inference server provides an inference service via an HTTP endpoint, allowing remote clients to request inferencing for any model that is being managed by the server.

The TensorRT inference server itself is included in the TensorRT inference server container. External to the container, there are additional C++ and Python client libraries, and additional documentation at GitHub: Inference Server.

This document describes the key features, software enhancements and improvements, any known issues, and how to run this container.

Chapter 2. PULLING A CONTAINER

Before you can pull a container from the NGC container registry, you must have Docker and nvidia-docker installed. For DGX users, this is explained in Preparing to use NVIDIA Containers Getting Started Guide.

For users other than DGX, follow the NVIDIA® GPU Cloud™ (NGC) container registry nvidia-docker installation documentation based on your platform.

You must also have access and logged into the NGC container registry as explained in the NGC Getting Started Guide.

There are the four repositories where you can find the NGC docker containers.

`nvcr.io/nvidia/`

The deep learning framework containers are stored in the **`nvcr.io/nvidia/`** repository.

`nvcr.io/hpc`

The HPC containers are stored in the **`nvcr.io/hpc`** repository.

`nvcr.io/nvidia-hpcvis`

The HPC visualization containers are stored in the **`nvcr.io/nvidia-hpcvis`** repository.

`nvcr.io/partner`

The partner containers are stored in the **`nvcr.io/partner`** repository. Currently the partner containers are focused on Deep Learning or Machine Learning, but that doesn’t mean they are limited to those types of containers.

Chapter 3. RUNNING THE TENSORRT INFERENCE SERVER

Before running the TensorRT inference server, you must first set up a model store containing the models that the server will make available for inferencing. The TensorRT Inference Server User Guide, describes how to create a model store. For this example, assume the model store is created on the host system directory **/path/to/model/store**.

For 18.09 and later releases, the following command will launch the TensorRT inference server using that model store.

```
$ nvidia-docker run --rm --shm-size=1g --ulimit memlock=-1 --ulimit
 stack=67108864 -p8000:8000 -p8001:8001 -v/path/to/model/store:/tmp/models
 inferenceserver:18.xx-py3 /opt/tensorrtserver/bin/trtserver --model-store=/tmp/
models
```

For 18.08 and earlier releases, use the following command:

```
$ nvidia-docker run --rm --shm-size=1g --ulimit memlock=-1 --ulimit
 stack=67108864 -p8000:8000 -p8001:8001 -v/path/to/model/store:/tmp/models
 inferenceserver:18.xx-py<x> /opt/inference_server/bin/inference_server --model-
store=/tmp/models
```

Where **inferenceserver:18.xx-py3** is the container that was pulled from the NGC container registry as described in https://docs.nvidia.com/deeplearning/dgx/inference-user-guide/index.html#pullcontainer. Additionally, in 18.08 and earlier releases, where **py<x>** is the version of Python that you are using.

The **nvidia-docker -v** option maps **/path/to/model/store** on the host into the container at **/tmp/models**, and the **--model-store** option to the inference server is used to point to **/tmp/models** as the model store.

The TensorRT inference server listens on port 8000 and the above command uses the **-p** flag to map container port 8000 to host port 8000. A different host port can be used by modifying the **-p** flag, for example **-p9000:8000** will cause the TensorRT inference server to be available on host port 9000.

The **--shm-size** and **--ulimit** flags are recommended to improve TensorRT inference server performance. For **--shm-size** the minimum recommended size is 1g but larger sizes may be necessary depending on the number and size of models being served.

After starting, the TensorRT inference server will log initialization information to the console. Initialization is complete and the server is ready to accept requests after the console shows the following:

```
Starting server listening on :8000
```

Additionally, C++ and Python client libraries and examples are available at GitHub: TensorRT inference server. These libraries and examples demonstrate how to communicate with the TensorRT inference server container from a C++ or Python application.

Chapter 4. TENSORRT INFERENCE SERVER RELEASE 18.10 BETA

The TensorRT inference server container image, previously referred to as inference server, release 18.10, is available as a beta release.

Contents of the TensorRT inference server

This container image contains the TensorRT inference server executable in **`/opt/tensorrtserver`**.

The container also includes the following:

- Ubuntu 16.04 including Python 3.5
- NVIDIA CUDA 10.0 including CUDA® Basic Linear Algebra Subroutines library™ (cuBLAS) 10.0
- NVIDIA CUDA® Deep Neural Network library™ (cuDNN) 7.3.0
- NCCL 2.3.5 (optimized for NVLink™)
- OpenMPI 2.0
- TensorRT 5.0.0 RC

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 can use NVIDIA driver release 384. For more information, see CUDA Compatibility and Upgrades.

Key Features and Enhancements

This TensorRT inference server release includes the following key features and enhancements.

- The TensorRT inference server container image version 18.10 is based on NVIDIA TensorRT inference server 0.7.0 beta, TensorFlow 1.10.0, and Caffe2 0.8.1.
- Latest version of NCCL 2.3.5.
- Dynamic batching support is added for all model types. Dynamic batching can be enabled and configured on a per-model bases.
- An improved inference request scheduler provides better handling of inference requests.
- Added new metrics to indicate GPU power limit, GPU utilization, and model executions (which is useful for determining the impact of dynamic batching).
- Prometheus metrics are now tagged with GPU UUID, model name, and model version as appropriate, so that metric values can be correlated to specific GPUs and models.
- Request latencies reported by status API and metrics are more clear in what they report, for example total request time, queuing time, and inference compute time are now reported.
- Ubuntu 16.04 with September 2018 updates

Known Issues

This is a beta release of the TensorRT inference server. All features are expected to be available, however, some aspects of functionality and performance will likely be limited compared to a non-beta release.

Chapter 5. TENSORRT INFERENCE SERVER RELEASE 18.09 BETA

The TensorRT inference server container image, previously referred to as inference server, release 18.09, is available as a beta release.

Contents of the TensorRT inference server

This container image contains the TensorRT inference server executable in **`/opt/tensorrtserver`**.

The container also includes the following:

- Ubuntu 16.04 including Python 3.5
- NVIDIA CUDA 10.0 including CUDA® Basic Linear Algebra Subroutines library™ (cuBLAS) 10.0
- NVIDIA CUDA® Deep Neural Network library™ (cuDNN) 7.3.0
- NCCL 2.3.4 (optimized for NVLink™)
- OpenMPI 2.0
- TensorRT 5.0.0 RC

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 can use NVIDIA driver release 384. For more information, see CUDA Compatibility and Upgrades.

Key Features and Enhancements

This TensorRT inference server release includes the following key features and enhancements.

- The TensorRT inference server container image version 18.09 is based on NVIDIA TensorRT inference server 0.6.0 beta, TensorFlow 1.10.0, and Caffe2 0.8.1.
- Latest version of cuDNN 7.3.0.
- Latest version of CUDA 10.0 which includes support for DGX-2, Turing, and Jetson Xavier.
- Latest version of cuBLAS 10.0.
- Latest version of NCCL 2.3.4.
- Latest version of TensorRT 5.0.0 RC.
- Google Cloud Storage paths are now allowed when specifying the location of the model store. For example, **`--model-store=gs://<bucket>/<mode store path>`**.
- Additional Prometheus metrics are exposed on the metrics endpoint: GPU power usage; GPU power limit; per-model request, queue and compute time.
- The C++ and Python client API now supports asynchronous requests.
- Ubuntu 16.04 with August 2018 updates

Known Issues

- This is a beta release of the TensorRT inference server. All features are expected to be available, however, some aspects of functionality and performance will likely be limited compared to a non-beta release.
- Starting with the 18.09 release, the directory holding the TensorRT inference server components has changed from **`/opt/inference_server`** to **`/opt/tensorrtserver`** and the TensorRT inference server executable name has changed from **`inference_server`** to **`trtserver`**.

Chapter 6. INFERENCE SERVER RELEASE 18.08 BETA

The NVIDIA container image of the inference server, release 18.08, is available as a beta release.

Contents of the inference server

This container image contains the inference server executable in **`/opt/inference_server`**.

The container also includes the following:

- Ubuntu 16.04

> 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™)
- TensorRT 4.0.1

Driver Requirements

Release 18.08 is based on CUDA 9, which requires NVIDIA Driver release 384.xx.

Key Features and Enhancements

This inference server release includes the following key features and enhancements.

- The inference server container image version 18.08 is based on NVIDIA inference server 0.5.0 beta, TensorFlow 1.9.0, and Caffe2 0.8.1.
- Latest version of cuDNN 7.2.1.
- Added support for Kubernetes compatible ready and live endpoints.

- Added support for Prometheus metrics. Load metric is reported that can be used for Kubernetes-style auto-scaling.
- Enhance example **`perf_client`** application to generate latency vs. inferences/ second results.
- Improve performance of TensorRT models by allowing multiple TensorRT model instances to execute simultaneously.
- Improve HTTP client performance by reusing connections for multiple inference requests.
- Ubuntu 16.04 with July 2018 updates

Known Issues

- This is a beta release of the inference server. All features are expected to be available, however, some aspects of functionality and performance will likely be limited compared to a non-beta release.
- There is a known performance regression in the inference benchmarks for ResNet-50. We haven't seen this regression in the inference benchmarks for VGG or training benchmarks for any network. The cause of the regression is still under investigation.

Chapter 7. INFERENCE SERVER RELEASE 18.07 BETA

The NVIDIA container image of the inference server, release 18.07, is available as a beta release.

Contents of the inference server

This container image contains the inference server executable in **`/opt/inference_server`**.

The container also includes the following:

- Ubuntu 16.04

> 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™)
- TensorRT 4.0.1

Driver Requirements

Release 18.07 is based on CUDA 9, which requires NVIDIA Driver release 384.xx.

Key Features and Enhancements

This inference server release includes the following key features and enhancements.

- The inference server container image version 18.07 is based on NVIDIA inference server 0.4.0 beta, TensorFlow 1.8.0, and Caffe2 0.8.1.
- Latest version of CUDA® Basic Linear Algebra Subroutines library™ (cuBLAS) 9.0.425.

- Support added for TensorFlow SavedModel format.
- Support added for gRPC in addition to existing HTTP REST.
- Ubuntu 16.04 with June 2018 updates

Known Issues

This is a beta release of the inference server. All features are expected to be available, however, some aspects of functionality and performance will likely be limited compared to a non-beta release.

Chapter 8. INFERENCE SERVER RELEASE 18.06 BETA

The NVIDIA container image of the inference server, release 18.06, is available as a beta release.

Contents of the inference server

This container image contains the inference server executable in **`/opt/inference_server`**.

The container also includes the following:

- Ubuntu 16.04

> 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™)
- TensorRT 4.0.1

Driver Requirements

Release 18.06 is based on CUDA 9, which requires NVIDIA Driver release 384.xx.

Key Features and Enhancements

This inference server release includes the following key features and enhancements.

- The inference server container image version 18.06 is based on NVIDIA inference server 0.3.0 beta, TensorFlow 1.8.0, and Caffe2 0.8.1.
- Support added for Caffe2 NetDef models.

- Support added for CPU-only servers in addition to servers that have one or more GPUs. The inference server can simultaneously use both CPUs and GPUs for inferencing.
- Logging format and control is unified across all inferencing backends: TensorFlow, TensorRT, and Caffe2.
- Gracefully exits upon receiving SIGTERM or SIGINT. Any in-flight inferences are allowed to complete before exiting, subject to a timeout.
- Server status is enhanced to report the readiness and availability of the server and of each model (and model version).
- Ubuntu 16.04 with May 2018 updates

Known Issues

This is a beta release of the inference server. All features are expected to be available, however, some aspects of functionality and performance will likely be limited compared to a non-beta release.

Chapter 9. INFERENCE SERVER RELEASE 18.05 BETA

The NVIDIA container image of the inference server, release 18.05, is available as a beta release.

Contents of the inference server

This container image contains the inference server executable in `/opt/inference_server`.

The container also includes the following:

- Ubuntu 16.04

  Container image `18.05-py2` contains Python 2.7; `18.05-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) 7.1.2
- NCCL 2.1.15 (optimized for NVLink™)
- TensorRT 3.0.4

Driver Requirements

Release 18.05 is based on CUDA 9, which requires NVIDIA Driver release 384.xx.

Key Features and Enhancements

This inference server release includes the following key features and enhancements.

- The inference server container image version 18.05 is based on NVIDIA inference server 0.2.0 beta and TensorFlow 1.7.0.
- Multiple model support. The inference server can manage any number and mix of TensorFlow to TensorRT models (limited by system disk and memory resources).

- TensorFlow to TensorRT integrated model support. The inference server can manage TensorFlow models that have been optimized with TensorRT.
- Multi-GPU support. The Inference Server can distribute inferencing across all system GPUs. Systems with heterogeneous GPUs are supported.
- Multi-tenancy support. Multiple models (or multiple instances of the same model) can run simultaneously on the same GPU.
- Batching support
- Ubuntu 16.04 with April 2018 updates

Known Issues

This is a beta release of the inference server. All features are expected to be available, however, some aspects of functionality and performance will likely be limited compared to a non-beta release.

Chapter 10. INFERENCE SERVER RELEASE 18.04 BETA

The NVIDIA container image of the inference server, release 18.04, is available as a beta release.

Contents of the inference server

This container image contains the inference server executable in **`/opt/inference_server`**.

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 inference server release includes the following key features and enhancements.

- This is the beta release of the inference server container.
- The inference server container image version 18.04 is based on NVIDIA inference server 0.1.0 beta.
- Multiple model support. The inference server can manage any number and mix of models (limited by system disk and memory resources). Supports TensorRT and TensorFlow GraphDef model formats.
- Multi-GPU support. The server can distribute inferencing across all system GPUs.
- Multi-tenancy support. Multiple models (or multiple instances of the same model) can run simultaneously on the same GPU.

- Batching support.
- Latest version of NCCL 2.1.15
- Ubuntu 16.04 with March 2018 updates

Known Issues

This is a beta release of the inference server. All features are expected to be available, however, some aspects of functionality and performance will likely be limited compared to a non-beta release.

Notice

Trademarks

Copyright

www.nvidia.com