INFERENCE SERVER

RN-08995-001 _v19.07 | July 2019

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 or GRPC 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 be logged into the NGC container registry as explained in the NGC Getting Started Guide.

There are 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 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 documentation 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**.

The following command will launch the 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
 tensorrtserver:19.xx-py3 /opt/tensorrtserver/bin/trtserver --model-store=/tmp/
models
```

Where **tensorrtserver:19.xx-py3** is the container that was pulled from the NGC container registry as described in Installing PreBuilt Containers.

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 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 inference server to be available on host port 9000.

The **--shm-size** and **--ulimit** flags are recommended to improve 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.

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 Inference Server container from a C++ or Python application.

Chapter 4.
TENSORRT INFERENCE SERVER RELEASE 19.07

The TensorRT Inference Server container image, release 19.07, is available on NGC and is open source on GitHub.

Contents of the TensorRT inference server container

The TensorRT Inference Server Docker image contains the inference server executable and related shared libraries in **/opt/tensorrtserver**.

The container also includes the following:

▸ Ubuntu 18.04
▸ NVIDIA CUDA 10.1.168 including cuBLAS 10.2.0.168
▸ NVIDIA cuDNN 7.6.1
▸ NVIDIA NCCL 2.4.7 (optimized for NVLink™)
▸ MLNX_OFED +3.4
▸ OpenMPI 3.1.3
▸ TensorRT 5.1.5

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 6.0 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 Inference Server release includes the following key features and enhancements.

▸ The inference server container image version 19.07 is based on NVIDIA TensorRT Inference Server 1.4.0, TensorFlow 1.14.0, ONNX Runtime 0.4.0, and Caffe2 0.8.2.

▸ Added `libtorch` as a new backend. PyTorch models manually decorated or automatically traced to produce TorchScript can now be run directly by the inference server.

▸ Build system converted from bazel to CMake. The new CMake-based build system is more transparent, portable and modular.

▸ To simplify the creation of custom backends, a Custom Backend SDK and improved documentation is now available.

▸ Improved AsyncRun API in C++ and Python client libraries.

▸ `perf_client` can now use user-supplied input data (previously used random or zero input data).

▸ `perf_client` now reports latency at multiple confidence percentiles (p50, p90, p95, p99) as well as a user-supplied percentile that is also used to stabilize latency results.

▸ Improvements to automatic model configuration creation (`--strict-model-config=false`).

▸ C++ and Python client libraries now allow additional HTTP headers to be specified when using the HTTP protocol.

▸ Latest version of NVIDIA cuDNN 7.6.1

▸ Latest version of MLNX_OFED +3.4

▸ Latest version of Ubuntu 18.04

Known Issues

There are no known issues in this release.

Chapter 5.
TENSORRT INFERENCE SERVER RELEASE 19.06

The TensorRT Inference Server container image, release 19.06, is available on NGC and is open source on GitHub.

Contents of the TensorRT inference server container

The TensorRT Inference Server Docker image contains the inference server executable and related shared libraries in **/opt/tensorrtserver**.

The container also includes the following:

▸ Ubuntu 16.04
▸ NVIDIA CUDA 10.1.168 including cuBLAS 10.2.0.168
▸ NVIDIA cuDNN 7.6.0
▸ NVIDIA NCCL 2.4.7 (optimized for NVLink™)
▸ OpenMPI 3.1.3
▸ TensorRT 5.1.5

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 6.0 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 Inference Server release includes the following key features and enhancements.

▸ The inference server container image version 19.06 is based on NVIDIA TensorRT Inference Server 1.3.0, TensorFlow 1.13.1, ONNX Runtime 0.4.0, and Caffe2 0.8.2.

▸ Latest version of NVIDIA CUDA 10.1.168 including cuBLAS 10.2.0.168

▸ Latest version of NVIDIA NCCL 2.4.7

▸ Added ONNX Runtime as a new backend. The ONNX Runtime backend allows the inference server to directly run ONNX models without requiring conversion to Caffe2 or TensorRT.

▸ HTTP health port may be specified independently of inference and status HTTP port with `--http-health-port` flag.

▸ Fixed bug in `perf_client` that caused high CPU usage by client that could lower the measured inference/sec in some cases.

▸ Ubuntu 16.04 with May 2019 updates (see Announcements)

Announcements

In the next release, we will no longer support Ubuntu 16.04. Release 19.07 will instead support Ubuntu 18.04.

Known Issues

▸ Google Cloud Storage (GCS) support is not available in this release. Support for GCS will be re-enabled in the 19.07 release.

Chapter 6.
TENSORRT INFERENCE SERVER RELEASE 19.05

The TensorRT Inference Server container image, release 19.05, is available on NGC and is open source on GitHub.

Contents of the TensorRT inference server container

The TensorRT Inference Server Docker image contains the inference server executable and related shared libraries in **/opt/tensorrtserver**.

The container also includes the following:

▸ Ubuntu 16.04
▸ 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™)
▸ OpenMPI 3.1.3
▸ TensorRT 5.1.5

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 6.0 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 Inference Server release includes the following key features and enhancements.

▸ The inference server container image version 19.05 is based on NVIDIA TensorRT Inference Server 1.2.0, TensorFlow 1.13.1, and Caffe2 0.8.2.

▸ 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

▸ Ensembling is now available. An ensemble represents a pipeline of one or more models and the connection of input and output tensors between those models. A single inference request to an ensemble will trigger the execution of the entire pipeline.

▸ Added a Helm chart that deploys a single TensorRT Inference Server into a Kubernetes cluster.

▸ The client **Makefile** now supports building for both Ubuntu 18.04 and Ubuntu 16.04. The Python wheel produced from the build is now compatible with both Python2 and Python3.

▸ The **perf_client** application now has a **--percentile** flag that can be used to report latencies instead of reporting average latency (which remains the default). For example, using **--percentile=99** causes **perf_client** to report the 99th percentile latency.

▸ The **perf_client** application now has a **-z** option to use zero-valued input tensors instead of random values.

▸ Improved error reporting of incorrect input/output tensor names for TensorRT models.

▸ Added **--allow-gpu-metrics** option to enable/disable reporting of GPU metrics.

Known Issues

There are no known issues in this release.

Chapter 7.
TENSORRT INFERENCE SERVER RELEASE 19.04

The TensorRT Inference Server container image, release 19.04, is available on NGC and is open source on GitHub.

Contents of the TensorRT inference server container

The TensorRT Inference Server Docker image contains the inference server executable and related shared libraries in **/opt/tensorrtserver**.

The container also includes the following:

▸ Ubuntu 16.04
▸ NVIDIA CUDA 10.1.105 including cuBLAS 10.1.0.105
▸ NVIDIA cuDNN 7.5.0
▸ NVIDIA NCCL 2.4.6 (optimized for NVLink™)
▸ OpenMPI 3.1.3
▸ TensorRT 5.1.2 RC

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 6.0 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 Inference Server release includes the following key features and enhancements.

▸ The inference server container image version 19.04 is based on NVIDIA TensorRT Inference Server 1.1.0, TensorFlow 1.13.1, and Caffe2 0.8.2.

▸ Latest version of NVIDIA NCCL 2.4.6

▸ Latest version of cuBLAS 10.1.0.105

▸ Client libraries and examples now build with a separate Makefile (a Dockerfile is also included for convenience).

▸ Input or output tensors with variable-size dimensions (indicated by -1 in the model configuration) can now represent tensors where the variable dimension has value 0 (zero).

▸ Zero-sized input and output tensors are now supported for batching models. This enables the inference server to support models that require inputs and outputs that have shape `[ batch-size ]`.

▸ TensorFlow custom operations (C++) can now be built into the inference server. An example and documentation are included in this release.

▸ Ubuntu 16.04 with March 2019 updates

Known Issues

There are no known issues in this release.

Chapter 8.
TENSORRT INFERENCE SERVER RELEASE 19.03

The TensorRT Inference Server container image, release 19.03, is available on NGC and is open source on GitHub.

Contents of the TensorRT inference server container

The TensorRT Inference Server Docker image contains the inference server executable and related shared libraries in **/opt/tensorrtserver**.

The container also includes the following:

▶ Ubuntu 16.04
▶ NVIDIA CUDA 10.1.105 including cuBLAS 10.1.105
▶ NVIDIA cuDNN 7.5.0
▶ NVIDIA NCCL 2.4.3 (optimized for NVLink™)
▶ OpenMPI 3.1.3
▶ TensorRT 5.1.2 RC

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 6.0 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 Inference Server release includes the following key features and enhancements.

- ▸ The inference server container image version 19.03 is based on NVIDIA TensorRT Inference Server 1.0.0, TensorFlow 1.13.1, and Caffe2 0.8.2.
- ▸ 19.03 is the first GA release of TensorRT Inference Server. See the README at the GitHub project for information on backwards-compatibility guarantees for this and future releases.
- ▸ Added support for "stateful" models and backends that require multiple inference requests be routed to the same model instance/batch slot. The new *sequence batcher* provides scheduling and batching capabilities for this class of models.
- ▸ Added GRPC streaming protocol support for inference requests.
- ▸ HTTP front-end is now asynchronous to enable lower-latency and higher-throughput handling of inference requests.
- ▸ Enhanced `perf_client` to support "stateful"/sequence models and backends.
- ▸ 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
- ▸ Latest version of TensorRT 5.1.2 RC
- ▸ Ubuntu 16.04 with February 2019 updates

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 9.
TENSORRT INFERENCE SERVER RELEASE 19.02 BETA

The TensorRT Inference Server container image, release 19.02, is available as a beta release and is open source on GitHub.

Contents of the TensorRT inference server

This container image contains the TensorRT Inference Server executable and related shared libraries in **/opt/tensorrtserver**.

The container also includes the following:

▸ Ubuntu 16.04
▸ 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™)
▸ OpenMPI 3.1.3
▸ TensorRT 5.0.2

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 6.0 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 Inference Server release includes the following key features and enhancements.

▸ The inference server container image version 19.02 is based on NVIDIA TensorRT Inference Server 0.11.0 beta, TensorFlow 1.13.0-rc0, and Caffe2 0.8.2.

▸ Variable-size input and output tensors are now supported.

▸ **STRING** datatype is now supported for input and output tensors for TensorFlow models and custom backends.

▸ The inference server can now be run on systems without GPUs or that do not have CUDA installed.

▸ Ubuntu 16.04 with January 2019 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.

▸ 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 10.
TENSORRT INFERENCE SERVER RELEASE 19.01 BETA

The TensorRT Inference Server container image, release 19.01, is available as a beta release and is open source on GitHub.

Contents of the TensorRT inference server

This container image contains the TensorRT Inference Server executable and related shared libraries in **/opt/tensorrtserver**.

The container also includes the following:

▸ Ubuntu 16.04
▸ 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™)
▸ OpenMPI 3.1.3
▸ TensorRT 5.0.2

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 6.0 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 Inference Server release includes the following key features and enhancements.

▸ The inference server container image version 19.01 is based on NVIDIA TensorRT Inference Server 0.10.0 beta, TensorFlow 1.12.0, and Caffe2 0.8.2.

▸ Latest version of NVIDIA cuDNN 7.4.2

▸ Latest version of OpenMPI 3.1.3

▸ Custom backend support. The inference server allows individual models to be implemented with custom backends instead of by a deep learning framework. With a custom backend, a model can implement any logic desired, while still benefiting from the GPU support, concurrent execution, dynamic batching and other features provided by the inference server.

▸ Ubuntu 16.04 with December 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.

▸ 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 11.
TENSORRT INFERENCE SERVER RELEASE 18.12 BETA

The TensorRT Inference Server container image, previously referred to as Inference Server, release 18.12, is available as a beta release.

Contents of the TensorRT inference server

This container image contains the TensorRT Inference Server (TRTIS) executable and related shared libraries in **/opt/tensorrtserver**.

The container also includes the following:

▸ Ubuntu 16.04 including 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
▸ TensorRT 5.0.2

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 6.0 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 Inference Server release includes the following key features and enhancements.

▸ The Inference Server container image version 18.12 is based on NVIDIA Inference Server 0.9.0 beta, TensorFlow 1.12.0, and Caffe2 0.8.2.

▸ TensorRT inference server is now open source. For more information, see GitHub.

▸ TRTIS now monitors the model repository for any change and dynamically reloads the model when necessary, without requiring a server restart. It is now possible to add and remove model versions, add/remove entire models, modify the model configuration, and modify the model labels while the server is running.

▸ Added a model priority parameter to the model configuration. Currently the model priority controls the CPU thread priority when executing the model and for TensorRT models also controls the CUDA stream priority.

▸ Fixed a bug in GRPC API: changed the model version parameter from string to int. *This is a non-backwards compatible change.*

▸ Added `--strict-model-config=false` option to allow some model configuration properties to be derived automatically. For some model types, this removes the need to specify the `config.pbtxt` file.

▸ Improved performance from an asynchronous GRPC frontend.

▸ Ubuntu 16.04 with November 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 12.
TENSORRT INFERENCE SERVER RELEASE 18.11 BETA

The inference server container image, previously referred to as Inference Server, release 18.11, 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 10.0.130 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
- ▸ TensorRT 5.0.2

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 Inference Server release includes the following key features and enhancements.

- ▸ The Inference Server container image version 18.11 is based on NVIDIA Inference Server 0.8.0 beta, TensorFlow 1.12.0-rc2, and Caffe2 0.8.2.

- Models may now be added to and removed from the model repository without requiring an inference server restart.
- Fixed an issue with models that don't support batching. For models that don't support batching, set the model configuration to `max_batch_size = 0`.
- Added a metric to indicate GPU energy consumption.
- Latest version of NCCL 2.3.7.
- Latest version of NVIDIA cuDNN 7.4.1.
- Latest version of TensorRT 5.0.2
- Ubuntu 16.04 with October 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 13.
TENSORRT INFERENCE SERVER RELEASE 18.10 BETA

The 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 10.0.130 including CUDA® Basic Linear Algebra Subroutines library™ (cuBLAS) 10.0.130
▸ NVIDIA CUDA® Deep Neural Network library™ (cuDNN) 7.3.0
▸ NCCL 2.3.6 (optimized for NVLink™)
▸ OpenMPI 3.1.2
▸ 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 may use NVIDIA driver release 384. For more information, see CUDA Compatibility and Upgrades.

Key Features and Enhancements

This Inference Server release includes the following key features and enhancements.

▸ The Inference Server container image version 18.10 is based on NVIDIA Inference Server 0.7.0 beta, TensorFlow 1.10.0, and Caffe2 0.8.2.

- ▸ Latest version of NCCL 2.3.6.
- ▸ Latest version of OpenMPI 3.1.2.
- ▸ 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 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 14.
TENSORRT INFERENCE SERVER RELEASE 18.09 BETA

The 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 10.0.130 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™)
▸ 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 may use NVIDIA driver release 384. For more information, see CUDA Compatibility and Upgrades.

Key Features and Enhancements

This Inference Server release includes the following key features and enhancements.

▸ The Inference Server container image version 18.09 is based on NVIDIA 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 10.0.130 which includes support for DGX-2, Turing, and Jetson Xavier.
- ▶ Latest version of cuBLAS 10.0.130.
- ▶ Latest version of NCCL 2.3.4.
- ▶ 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 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 15.
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 16.
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 17.
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 18.
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 19.
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.