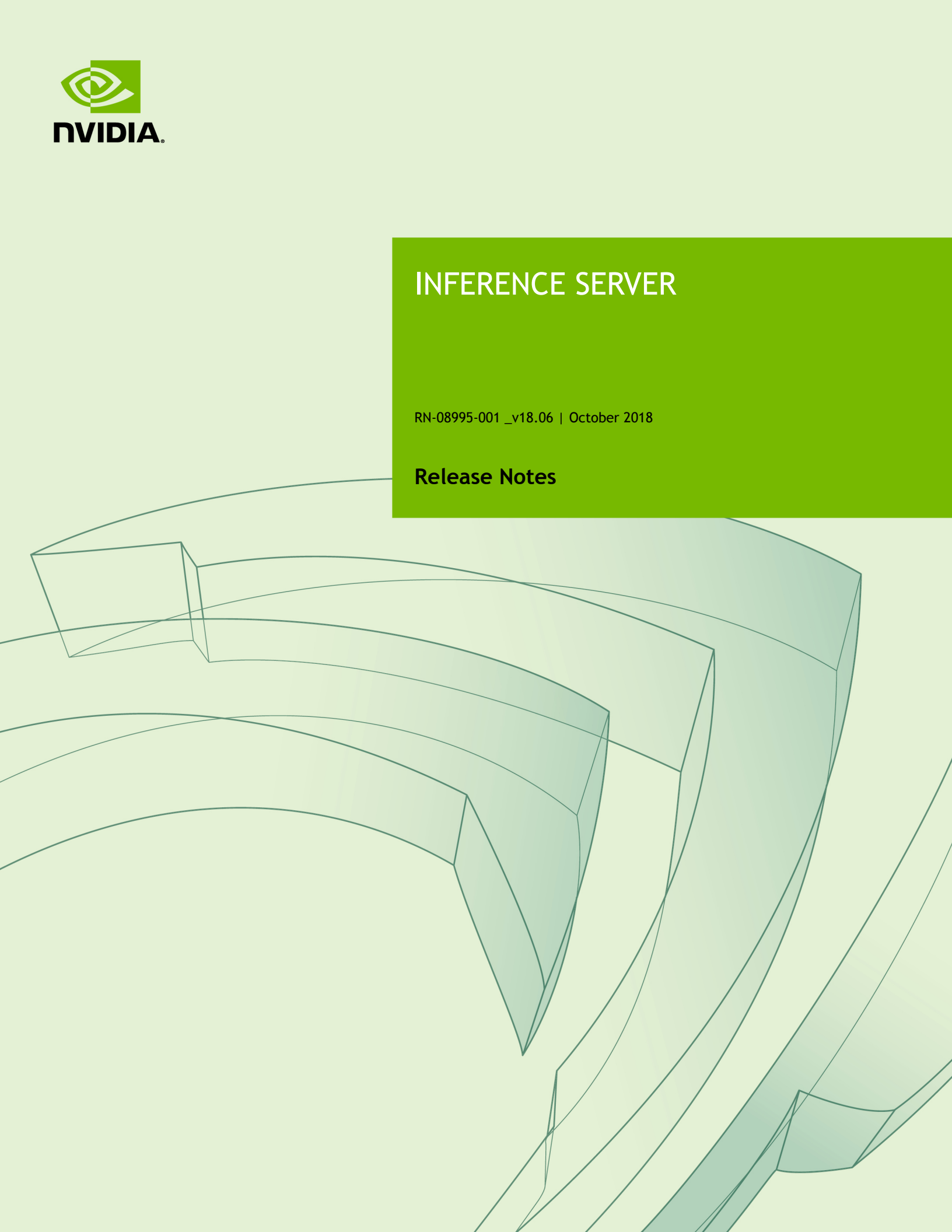

INFERENCE SERVER

RN-08995-001 _v18.06 | October 2018

Release Notes

TABLE OF CONTENTS

Chapter 1. Inference Server Overview..... 1
Chapter 2. Pulling A Container.....2
Chapter 3. Running The Inference Server..... 3
Chapter 4. Inference Server Release 18.06 Beta..... 5
Chapter 5. Inference Server Release 18.05 Beta..... 7
Chapter 6. Inference Server Release 18.04 Beta..... 9

Chapter 1. INFERENCE SERVER OVERVIEW

The NVIDIA® TensorRT™ Inference Server provides a cloud inferencing solution optimized for NVIDIA GPUs. The 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 Inference Server itself is included in the 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

You can access NVIDIA's GPU accelerated containers for all three products, the NVIDIA DGX-1™, NVIDIA DGX Station™, and the NVIDIA® GPU Cloud™ (NGC). If you own a DGX-1 or DGX Station then you should use the NVIDIA® DGX™ container registry at https://compute.nvidia.com. This is a web interface to the Docker hub, **`nvcr.io`** (NVIDIA DGX container registry). You can pull the containers from there and you can also push containers there into your own account in the registry.

If you are accessing the NVIDIA containers from the NVIDIA® GPU Cloud™ (NGC) container registry via a cloud services provider such as Amazon® Web Services™ (AWS), then you should use the NGC container registry at https://ngc.nvidia.com. This is also a web interface to the same Docker repository as for the DGX-1 and DGX Station. After you create an account, the commands to pull containers are the same as if you had a DGX-1 in your own data center. However, currently, you cannot save any containers to the NGC container registry. Instead, you have to save the containers to your own Docker repository that is either on-premise or in the Cloud.

The containers are exactly the same, whether you pull them from the NVIDIA DGX container registry or the NGC container registry.

Before you can pull a container from the NGC container registry, you must have Docker and nvidia-docker installed as explained in Preparing to use NVIDIA Containers Getting Started Guide. You must also have access and logged into the NGC container registry as explained in the NGC Getting Started Guide.

For step-by-step instructions, see Container User Guide.

Chapter 3. RUNNING THE 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 Inference Server User Guide - Model Store, 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=6710s8864 -p8000:8000 -v/path/to/model/store:/tmp/models
<container> /opt/inference_server/bin/inference_server
--model-store=/tmp/models
```

Where **<container>** is the name of the docker container that was pulled from the NVIDIA DGX or NGC container registry as described in https://docs.nvidia.com/deeplearning/dgx/inference-user-guide/index.html#pullcontainer.

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.

After starting, the 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: Inference Server. These libraries and examples demonstrate how to communicate with the Inference Server container from a C++ or Python application.

Chapter 4.
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 0.3 beta and 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

There are no known issues in this release.

Chapter 5. 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 0.2 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

There are no known issues in this release.

Chapter 6. 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 0.1 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