TECHNICAL OVERVIEW

NVIDIA GPU CLOUD DEEP LEARNING FRAMEWORKS

A Guide to the Optimized Framework Containers on NVIDIA GPU Cloud

NVIDIA

Introduction

Artificial intelligence is helping to solve some of the most complex problems facing humankind: providing early detection and finding cures for infectious diseases, reducing traffic fatalities, finding imperfections in critical infrastructure before they create safety hazards, and much more. Two of the biggest hurdles in using AI and deep learning are maximizing performance and managing the constant rate of churn in the underlying technologies.

NVIDIA solves these problems with NVIDIA GPU Cloud (NGC). NGC empowers AI researchers with performance-engineered deep learning framework containers, allowing them to spend less time on IT, and more time experimenting, gaining insights, and driving results.

Deep Learning with NVIDIA AI

NVIDIA GPU Cloud is a GPU-accelerated cloud platform optimized for deep learning. NGC manages a catalog of fully integrated and optimized deep learning framework containers that take full advantage of NVIDIA GPUs. These framework containers are delivered ready-to-run, including all necessary dependencies such as CUDA runtime, NVIDIA libraries, and an operating system. They are tuned, tested, and certified by NVIDIA to run on Amazon EC2 P3 instances with NVIDIA Volta™ (with additional cloud providers coming soon) and NVIDIA DGX Systems. NVIDIA updates these containers monthly to ensure they continue to provide peak performance.

GET UP AND RUNNING WHEN AND WHERE YOU WANT

Available through NVIDIA GPU Cloud, the NGC container registry makes it simple to tap into the power of the latest NVIDIA GPUs. Users can now easily create deep neural networks (DNNs) using pre-integrated containers that leverage the full power of NVIDIA GPUs. This makes it easier than ever for data scientists, researchers, and engineers to tackle challenges with AI that were once thought impossible, whether they are working in a fully outfitted lab, or leveraging cloud infrastructure.

- **Innovate in Minutes, Not Weeks** - The top deep learning frameworks like TensorFlow, PyTorch, MXNet, and more, are tuned, tested, and certified by NVIDIA for maximum performance on Amazon EC2 P3 instances with NVIDIA Tesla V100 GPUs, and NVIDIA DGX Systems. The frameworks are delivered in pre-integrated,

easy to use containers so users can start doing deep learning jobs immediately, eliminating time-consuming and difficult do-it-yourself software integration.

- **Deep Learning Across Platforms** - Data scientists and researchers can rapidly build, train, and deploy deep neural network models on NVIDIA GPUs on the desktop, in the datacenter, and in the cloud. NGC gives them the flexibility to work in the environment that is best for them, and provides immediate scalability when needed, helping them address some of the most complicated AI challenges.
- **Always Up to Date** - Containers available on NGC benefit from continuous NVIDIA development, ensuring each deep learning framework is tuned for the fastest training possible on the latest NVIDIA GPUs. NVIDIA engineers continually optimize libraries, drivers, and containers, delivering monthly updates to ensure that users' deep learning investments reap greater returns over time.

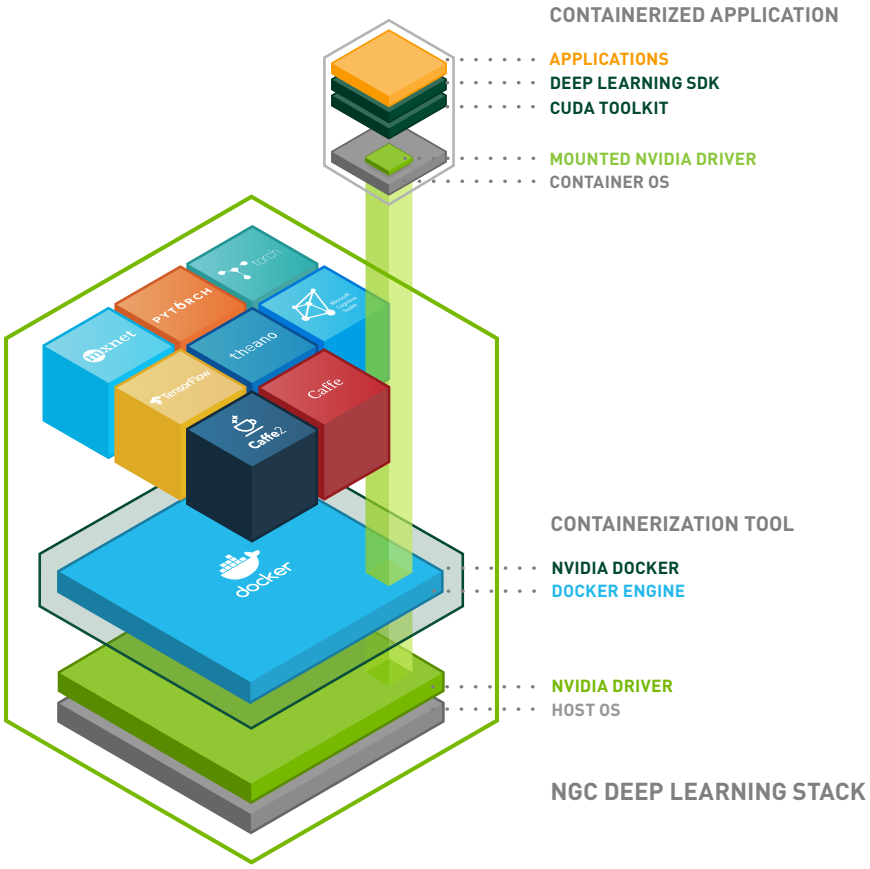
NGC Container Registry

The NGC container registry is a catalog of GPU-accelerated deep learning software. It includes CUDA Toolkit, DIGITS workflow, and the following deep learning frameworks: NVCaffe, Caffe2, Microsoft Cognitive Toolkit (CNTK), MXNet, PyTorch, TensorFlow, Theano, and Torch.

The NGC container registry provides containerized versions of these frameworks. These frameworks, including all necessary dependencies, form the NGC Deep Learning Stack. For users who need more flexibility to build custom deep learning solutions, each framework container image also includes the framework source code to enable custom modifications and enhancements, along with the complete software development stack.

The design of the platform software is centered around a minimal OS and driver install on the server and provisioning of all application and SDK software in NVIDIA Docker containers through NVIDIA Docker Registry. Figure 1 presents a graphical layout of the layers of the NGC Deep Learning Stack.

Figure 1: NVIDIA Docker mounts the user mode components of the NVIDIA driver and the GPUs into the Docker container at launch.



To enable portability in Docker images that leverage GPUs, NVIDIA developed NVIDIA Docker, an open source project that provides a command line tool to mount the user mode components of the NVIDIA driver and the GPUs into the Docker container at launch. nv-docker is essentially a wrapper around Docker that transparently provisions a container with the necessary components to execute code on the GPU. A Docker container is a mechanism for bundling a Linux application with all of its libraries, configuration files, and environment variables so that the execution environment is always the same, on whatever Linux system it runs and between instances on the same host. Docker containers are user-mode only, so all kernel calls from the container are handled by the host system kernel.

A LAYERED APPROACH

A deep learning framework is part of a software stack that consists of several layers. Each layer depends on the layer below it in the stack. This software architecture has many advantages:

- Because each deep learning framework is in a separate container, each framework can use different versions of libraries such as libc, cuDNN, and others, and not interfere with each other.

- A key reason for having layered containers is that one can target the experience for what the user requires.
- As deep learning frameworks are improved for performance or bug fixes, new versions of the containers are made available in the registry.
- The system is easy to maintain, and the OS image stays clean since applications are not installed directly on the OS.
- Security updates, driver updates, and OS patches can be delivered seamlessly.

WHY USE A FRAMEWORK?

Frameworks have been created to make researching and applying deep learning more accessible and efficient. The key benefits of using frameworks include:

- Frameworks provide highly optimized GPU enabled code specific to the computations required for training deep neural networks (DNNs).
- NVIDIA's frameworks are tuned and tested for the best possible GPU performance.
- Frameworks provide access to code through a simple command line or scripting language interfaces such as Python.
- Many powerful DNNs can be trained and deployed using these frameworks without ever having to write any GPU or complex compiled code, but while still benefiting from the training speed-up afforded by GPU acceleration.

The NGC Deep Learning Stack Containers

Caffe

NVCAFFE

Caffe is a deep learning framework made with flexibility, speed, and modularity in mind. It was originally developed by the Berkeley Vision and Learning Center (BVLC) and by community contributors.

NVCaffe is an NVIDIA-maintained fork of BVLC Caffe tuned for NVIDIA GPUs, particularly in multi-GPU configurations. NVCaffe includes:

- Mixed-precision support. It allows to store and/or compute data in either 64, 32 or 16-bit formats. Precision can be defined on each

layer (forward and backward phases might be different too), or it can be set to default for the whole Net.

- Integration with cuDNN v6.
- Automatic selection of the best cuDNN convolution algorithm.
- Integration with v1.3.4 of NCCL library for improved multi-GPU scaling.
- Optimized GPU memory management for data and parameters storage, I/O buffers and workspace for convolutional layers.
- Parallel data parser and transformer for improved I/O performance.
- Parallel back-propagation and gradient reduction on multi-GPU systems.
- Fast solvers implementation with fused CUDA kernels for weights and history update.
- Multi-GPU test phase for even memory load across multiple GPUs.
- Backward compatibility with BVLC Caffe and NVCaffe 0.15.
- Extended set of optimized models (including 16-bit floating point examples).

CAFFE2

Caffe2 is a deep-learning framework designed to easily express all model types, for example, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and more, in a friendly Python-based API, and execute them using a highly efficient C++ and CUDA backend.

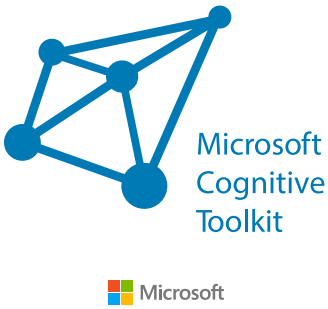
It allows a large amount of flexibility for the user to assemble their model, whether for inference or training, using combinations of high-level and expressive operations, before running through the same Python interface allowing for easy visualization, or serializing the created model and directly using the underlying C++ implementation.

Caffe2 supports single and multi-GPU execution, along with support for multi-node execution.

The following list summarizes the NGC Deep Learning Stack Caffe2 optimizations and changes:

- Use of the latest cuDNN release
- Performance fine-tuning
- GPU-accelerated image input pipeline

- Automatic selection of the best convolution algorithm



MICROSOFT COGNITIVE TOOLKIT

The Microsoft Cognitive Toolkit (previously known as CNTK), is a unified deep learning toolkit that allows users to easily realize and combine popular model types such as feed-forward deep neural networks (DNNs), CNNs, and RNNs.

The Microsoft Cognitive Toolkit implements Stochastic Gradient Descent (SGD) learning with automatic differentiation and parallelization across multiple GPUs and servers. The Microsoft Cognitive Toolkit can be called as a library from Python or C++ applications or executed as a standalone tool using the BrainScript model description language.

The following list summarizes the NGC Deep Learning Stack Microsoft Cognitive Toolkit optimizations and changes:

- Use of the latest cuDNN release
- Integration of the latest version of NCCL with NVLink support for improved multi-GPU scaling. NCCL with NVLink boosts the training performance of ResNet-50 by 2x when using data parallel SGD.
- Image reader pipeline improvements allow AlexNet to train at over 12,000 images/second
- Reduced GPU memory overhead for multi-GPU training by up to 2 GB per GPU
- Dilated convolution support
- Optimizations reducing the memory footprint needed for cuDNN workspaces

MXNET

MXNet is a deep learning framework designed for both efficiency and flexibility, which allows you to mix the symbolic and imperative programming to maximize efficiency and productivity.

At the core of MXNet is a dynamic dependency scheduler that automatically parallelizes both symbolic and imperative operations on the fly. A graph optimization layer on top of the scheduler makes symbolic execution fast and memory efficient. MXNet is portable and lightweight and scales to multiple GPUs and multiple machines.

The following list summarizes the NGC Deep Learning Stack MXNet optimizations and changes:

- Use of the latest cuDNN release
- Improved input pipeline for image processing
- Optimized embedding layer CUDA kernels
- Optimized tensor broadcast and reduction CUDA kernels

PYTORCH

PyTorch is a Python package that provides two high-level features:

- Tensor computation (like numpy) with strong GPU acceleration
- Deep Neural Networks built on a tape-based autograd system

You can reuse your favorite Python packages such as numpy, scipy and Cython to extend PyTorch when needed.

The following list summarizes the NGC Deep Learning Stack PyTorch optimizations and changes:

- Use of latest cuDNN release
- Integration of the latest version of NCCL with NVLink support
- Buffering of parameters to be communicated by NCCL to reduce latency overhead
- Dilated convolution support
- Optimizations to avoid unnecessary copies of data and zeroing of buffers

TENSORFLOW

TensorFlow is an open-source software library for numerical computation using data flow graphs. Nodes in the graph represent mathematical operations, while the graph edges represent the multidimensional data arrays (tensors) that flow between them. This flexible architecture lets you deploy computation to one or more CPUs or GPUs in a desktop, server, or mobile device without rewriting code.

TensorFlow was originally developed by researchers and engineers working on the Google Brain team within Google's Machine Intelligence research organization to conduct machine learning and deep neural networks research. The system is general enough to be applicable in a wide variety of other domains, as well.

For visualizing TensorFlow results, the TensorFlow Docker image also contains TensorBoard. TensorBoard is a suite of visualization tools. For

example, you can view the training histories as well as what the model looks like.

The following list summarizes the NGC Deep Learning Stack TensorFlow optimizations and changes:

- Use of the latest cuDNN release
- Integration of the latest version of NCCL with NVLink support for improved multi-GPU scaling. NCCL with NVLink boosts the training performance of ResNet-50 by 2x when using data parallel SGD.
- Support for fused color adjustment kernels by default
- Support for the use of non-fused Winograd convolution algorithms by default

THEANO

Theano is a Python library that allows you to efficiently define, optimize, and evaluate mathematical expressions involving multi-dimensional arrays. Theano has been powering large-scale computationally intensive scientific investigations since 2007.

The following list summarizes the NGC Deep Learning Stack Theano optimizations and changes:

- Use of the latest cuDNN release
- Runtime code generation: evaluate expressions faster
- Extensive unit-testing and self-verification: detect and diagnose many types of errors

TORCH

Torch is a scientific computing framework with wide support for deep learning algorithms. Torch is easy to use and efficient, thanks to an easy and fast scripting language, Lua, and an underlying C/CUDA implementation.

Torch offers popular neural network and optimization libraries that are easy to use yet provide maximum flexibility to build complex neural network topologies.

The following list summarizes the NGC Deep Learning Stack Torch optimizations and changes:

- Use of the latest cuDNN release

- Integration on the latest version of NCCL with NVLink support for improved multi-GPU scaling. NCCL with NVLink boosts the training performance of ResNet-50 by 2x when using data parallel SGD.
- Buffering of parameters to be communicated by NCCL to reduce latency overhead
- cuDNN bindings for re-currents networks (RNN, GRU, LSTM), including persistent versions, which greatly improving the performance of small batch training
- Dilated convolution support
- Support for 16- and 32-bit floating point (FP16 and FP32) data input to cuDNN routines
- Support for operations on FP16 tensors (using FP32 arithmetic)

DIGITS

DIGITS

The NVIDIA Deep Learning GPU Training System (DIGITS) puts the power of deep learning into the hands of engineers and data scientists.

DIGITS is not a framework. DIGITS is a wrapper for Caffe and Torch; which provides a graphical web interface to those frameworks rather than dealing with them directly on the command-line.

DIGITS can be used to rapidly train highly accurate deep neural network (DNNs) for image classification, segmentation, and object detection tasks. DIGITS simplifies common deep learning tasks such as managing data, designing and training neural networks on multi-GPU systems, monitoring performance in real time with advanced visualizations, and selecting the best performing model from the results browser for deployment. DIGITS is completely interactive so that data scientists can focus on designing and training networks rather than programming and debugging.

Accelerate AI with NVIDIA GPU Cloud

NVIDIA GPU Cloud features a comprehensive catalog of integrated and optimized deep learning software. NVIDIA leverages its years of research and development in AI to deliver ready to run, performance engineered software for everyone in the NGC container registry and contributes its enhancements to the deep learning frameworks back to the open source community.

As new versions of frameworks, drivers, and hardware are created, NVIDIA makes continual improvements and updates to ensure everything works together optimally and with maximum performance, removing the ongoing burden of testing and integration from the user. The frameworks available from the NGC container registry allow data scientists and researchers to deliver deep learning breakthroughs in nearly every discipline and industry and help them to solve some of the greatest challenges of our time with AI.

To learn more about NGC and watch a getting started video, visit:
www.NVIDIA.com/cloud

To sign up for NGC, visit:
www.NVIDIA.com/ngcsignup

NVIDIA