TECHNICAL OVERVIEW

NVIDIA GPU CLOUD DEEP LEARNING SOFTWARE

A Guide to the Optimized Deep Learning Containers Available from NVIDIA GPU Cloud

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 software 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 cloud-based container registry for GPU-accelerated software. NGC manages a catalog of fully integrated and optimized deep learning software containers that take full advantage of NVIDIA GPUs. These containers are delivered ready-to-run, including all necessary dependencies such as NVIDIA® CUDA® Toolkit, NVIDIA deep learning libraries, and an operating system. They're tuned, tested, and certified by NVIDIA to run on NVIDIA DGX™ Systems, NVIDIA TITAN (powered by NVIDIA Volta and NVIDIA Pascal™), NVIDIA Quadro® GV100, GP100, and P6000, and supported public cloud providers. NGC containers are available to users of supported NVIDIA GPUs on Amazon EC2, Google Cloud Platform, Microsoft Azure, and Oracle Cloud Infrastructure. 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 high-performance, 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're working on a desktop, working in a fully outfitted lab, leveraging cloud infrastructure, or using any combination of these.

- **Innovate in Minutes, Not Weeks** - The top deep learning software, like TensorFlow, PyTorch, MXNet, NVIDIA TensorRT™, and more, are tuned, tested, and certified by NVIDIA for maximum performance on NVIDIA DGX systems, NVIDIA TITAN (powered by NVIDIA Volta and NVIDIA Pascal), NVIDIA Quadro GV100, GP100, and P6000, and supported public cloud providers. NGC containers are available to users of supported NVIDIA GPUs on Amazon EC2, Google Cloud Platform, Microsoft Azure, and Oracle Cloud Infrastructure. The software is 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 data center, 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** - The deep learning containers available on NGC benefit from continuous NVIDIA development. NVIDIA engineers optimize libraries, drivers, and containers, delivering monthly updates to ensure that users' deep learning investments reap greater returns over time.

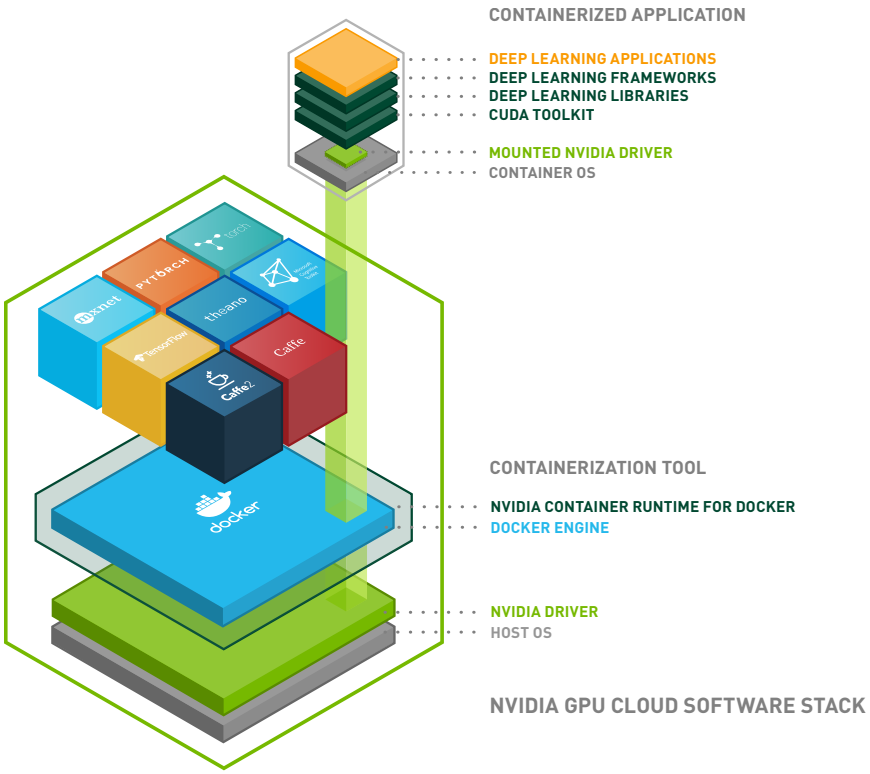
NGC Container Registry

The NGC container registry is a catalog of GPU-accelerated deep learning software. It includes the NVIDIA Cuda Toolkit, NVIDIA DIGITS™, NVIDIA TensorRT, 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 this software, including all necessary dependencies. This optimized set of software inside the containers is called the NVIDIA GPU Cloud Software 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 is centered around a minimal OS and driver install on the server and provisioning of all application and software development kit (SDK) software in containers through a registry. Figure 1 presents a graphical layout of the layers of the NGC Software Stack.

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



To enable portability in container images that leverage GPUs, NVIDIA developed NVIDIA Container Runtime for 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 or application is in a separate container, they 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 and applications 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 frameworks or 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 Containers

This section describes the deep learning software containers available on NGC. Each container is updated monthly to include the latest NVIDIA deep learning library integrations with cuDNN, cuBLAS, and NCCL.

The NVIDIA CUDA Deep Neural Network library (cuDNN) is a GPU-accelerated library of primitives for deep neural networks. cuDNN provides highly tuned implementations for standard routines such as forward and backward convolution, pooling, normalization, and activation layers.

The NVIDIA cuBLAS library is a GPU-accelerated implementation of the standard basic linear algebra subroutines (BLAS). Using cuBLAS APIs, you can speed up your applications by deploying compute-intensive operations to a single GPU or scale up and distribute work across multi-GPU configurations efficiently.

The NVIDIA Collective Communications Library (NCCL) implements multi-GPU and multi-node collective communication primitives that are performance optimized for NVIDIA GPUs. NCCL provides routines, such as all-gather, all-reduce, broadcast, reduce, reduce-scatter, that are optimized to achieve high bandwidth over PCIe and NVLink high-speed interconnect.

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. For more information on the latest enhancements, please see the **NVCaffe container release notes**.

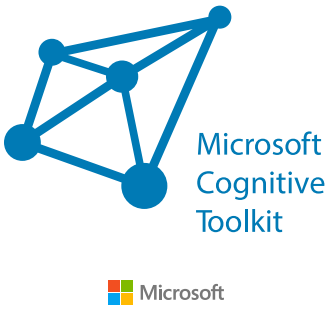
CAFFE2

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 application programming interface (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.

For more information on the latest enhancements, please see the **Caffe2 container release notes**.



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.

For more information on the latest enhancements, please see the **Microsoft Cognitive Toolkit container release notes**.

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.

For more information on the latest enhancements, please see the **MXNet container release notes**.

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.

For more information on the latest enhancements, please see the **PyTorch container release notes**.

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 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.

For more information on the latest enhancements, please see the **TensorFlow container release notes**.

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.

For more information on the latest enhancements, please see the **Theano container release notes**.

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.

For more information on the latest enhancements, please see the **Torch container release notes**.

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.

For more information on the latest enhancements, please see the **DIGITS container release notes**.

TensorRT

TENSORRT

NVIDIA TensorRT is a C++ library that facilitates high-performance inference on NVIDIA GPUs. TensorRT takes a network definition and optimizes it by merging tensors and layers, transforming weights, choosing efficient intermediate data formats, and selecting from a large kernel catalog based on layer parameters and measured performance.

TensorRT includes an infrastructure that allows you to leverage the high-speed, reduced-precision capabilities of Pascal and Volta GPUs as an optional optimization.

The TensorRT container provides an easy-to-use container for TensorRT development. It allows for the TensorRT samples to be built, modified, and executed. For more information about the TensorRT samples, see **NVIDIA Deep Learning SDK documentation**.

For more information on the latest enhancements, please see the **TensorRT container release notes**.

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 deep learning software available from the NGC container registry allows data scientists and researchers to deliver breakthroughs in nearly every discipline and industry and helps them to solve some of the greatest challenges of our time with AI.

To learn more about NGC, visit:
www.NVIDIA.com/cloud

To sign up and use NGC deep learning containers at no charge, visit:
www.NVIDIA.com/ngcsignup