Introduction
TABLE OF CONTENTS

Chapter 1. Introduction To Deep Learning SDK............................................................1
Chapter 1.
INTRODUCTION TO DEEP LEARNING SDK

The NVIDIA Deep Learning SDK provides powerful tools and libraries for designing and deploying GPU-accelerated deep learning applications. It includes libraries for deep learning primitives, inference, video analytics, linear algebra, sparse matrices, and multi-GPU communications.

Figure 1 NVIDIA Deep Learning SDK

Deep Learning Primitives (CUDA® Deep Neural Network library™ (cuDNN))
High-performance building blocks for deep neural network applications including convolutions, activation functions, and tensor transformations. The cuDNN product page can be found here. The cuDNN documentation page can be found here.

Deep Learning Inference Engine (TensorRT™)
High-performance deep learning inference runtime for production deployment. The TensorRT product page can be found here. The TensorRT documentation page can be found here.

Deep Learning for Video Analytics (NVIDIA DeepStream™ SDK)
High-level C++ API and runtime for GPU-accelerated transcoding and deep learning inference. The DeepStream SDK product page is located here.

Linear Algebra (CUDA® Basic Linear Algebra Subroutines library™ (cuBLAS))
GPU-accelerated BLAS functionality that delivers 6x to 17x faster performance than CPU-only BLAS libraries. The cuBLAS product page is located here. The cuBLAS documentation page is located here.

Sparse Matrix Operations (NVIDIA CUDA® Sparse Matrix library™ (cuSPARSE))
GPU-accelerated linear algebra subroutines for sparse matrices that deliver up to 8x faster performance than CPU BLAS (MKL), ideal for applications such as natural language processing and graph analytics.
language processing. The cuSPARSE product page is located [here](#). The cuSPARSE documentation is located [here](#).

**Multi-GPU Communication (NVIDIA® Collective Communications Library™ (NCCL))**

Collective communication routines, such as all-gather, reduce, and broadcast that accelerate multi-GPU deep learning training on up to eight GPUs. The NCCL product page is located [here](#). The NCCL documentation is located [here](#).

The Deep Learning SDK requires the **CUDA® Toolkit™**, which offers a comprehensive development environment for building new GPU-accelerated deep learning algorithms, and dramatically increasing the performance of existing applications.
Notice

THE INFORMATION IN THIS GUIDE AND ALL OTHER INFORMATION CONTAINED IN NVIDIA DOCUMENTATION REFERENCED IN THIS GUIDE IS PROVIDED “AS IS.” NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE INFORMATION FOR THE PRODUCT, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA’s aggregate and cumulative liability towards customer for the product described in this guide shall be limited in accordance with the NVIDIA terms and conditions of sale for the product.

THE NVIDIA PRODUCT DESCRIBED IN THIS GUIDE IS NOT FAULT TOLERANT AND IS NOT DESIGNED, MANUFACTURED OR INTENDED FOR USE IN CONNECTION WITH THE DESIGN, CONSTRUCTION, MAINTENANCE, AND/OR OPERATION OF ANY SYSTEM WHERE THE USE OR A FAILURE OF SUCH SYSTEM COULD RESULT IN A SITUATION THAT THREATENS THE SAFETY OF HUMAN LIFE OR SEVERE PHYSICAL HARM OR PROPERTY DAMAGE (INCLUDING, FOR EXAMPLE, USE IN CONNECTION WITH ANY NUCLEAR, AVIONICS, LIFE SUPPORT OR OTHER LIFE CRITICAL APPLICATION). NVIDIA EXPRESSLY DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY OF FITNESS FOR SUCH HIGH RISK USES. NVIDIA SHALL NOT BE LIABLE TO CUSTOMER OR ANY THIRD PARTY, IN WHOLE OR IN PART, FOR ANY CLAIMS OR DAMAGES ARISING FROM SUCH HIGH RISK USES.

NVIDIA makes no representation or warranty that the product described in this guide will be suitable for any specified use without further testing or modification. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer’s sole responsibility to ensure the product is suitable and fit for the application planned by customer and to do the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer’s product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this guide. NVIDIA does not accept any liability related to any default, damage, costs or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this guide, or (ii) customer product designs.

Other than the right for customer to use the information in this guide with the product, no other license, either expressed or implied, is hereby granted by NVIDIA under this guide. Reproduction of information in this guide is permissible only if reproduction is approved by NVIDIA in writing, is reproduced without alteration, and is accompanied by all associated conditions, limitations, and notices.

Trademarks

NVIDIA, the NVIDIA logo, and cuBLAS, CUDA, cuDNN, cuFFT, cuSPARSE, DALI, DIGITS, DGX, DGX-1, Jetson, Kepler, NVIDIA Maxwell, NCCL, NVLink, Pascal, Tegra, TensorRT, and Tesla are trademarks and/or registered trademarks of NVIDIA Corporation in the United States and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

© 2018 NVIDIA Corporation. All rights reserved.

www.nvidia.com