

Name: Deep Ghadiyali

Report: NVIDIA CUDA X GPU-Accelerated

Roll Number: CS17B011

Libraries

NVIDIA CUDA-X, built on top of NVIDIA CUDA, is a collection of libraries, tools, and technologies that deliver dramatically higher performance compared to CPU-only alternatives across multiple application domains, from artificial intelligence (AI) to high performance computing (HPC).

These libraries are easy to use and provides highly optimised GPU accelerations and developer can use this optimisation to utilize all the resources available on computer to speed up the process.

Components:

➤ **Math Libraries:**

- GPU-accelerated math libraries lay the foundation for compute-intensive applications in areas such as molecular dynamics, computational fluid dynamics, computational chemistry, medical imaging, and seismic exploration.
- cuBLAS: GPU-accelerated basic linear algebra (BLAS) library.
- cuFFT: GPU-accelerated library for Fast Fourier Transforms.
- CUDA Math library: GPU-accelerated standard mathematical functions.
- cuRAND: GPU-accelerated random number generation.
- cuSOLVER: GPU-accelerated dense and sparse direct solvers.
- cuSPARSE: GPU-accelerated BLAS for sparse matrices.
- cuTensor: GPU-accelerated tensor linear algebra library.
- AmgX: GPU-accelerated linear solvers for simulations and implicit unstructured methods.

➤ **Parallel Algorithms:**

- Thrust: GPU-accelerated library of C++ parallel algorithms and data structures.

➤ **Image and Video Libraries:**

- nvJPEG: High performance GPU-accelerated library for JPEG decoding
- Performance Primitives: Provides GPU-accelerated image, video, and signal processing functions.
- Video Codec SDK: For video encoding and decoding.
- Optical Flow SDK: For computing relative motion of pixels between images.

➤ **Communication Libraries:**

- Performance-optimized multi-GPU and multi-node communication primitives.
- NVSHMEM: OPenSHMEM standard for GPU memory, with extensions for improved performance.
- NCCL: Open-source library for fast multi-GPU, multi-node communications that maximizes bandwidth while maintaining low latency.

➤ **Deep Learning:**

- GPU-accelerated libraries for Deep Learning applications that leverage CUDA and specialized hardware components of GPUs.
- cuDNN: GPU-accelerated library of primitives for deep neural.

- networks TensorRT: High-performance deep learning inference optimizer and runtime for production deployment.
- Jarvis: Platform for developing engaging and contextual AI-powered conversation apps.
- DeepStream SDK: Real-time streaming analytics toolkit for AI-based video understanding and multi-sensor processing.
- DALI: Portable, open-source library for decoding and augmenting images and videos to accelerate deep learning applications.

➤ **Partner Libraries:**

- OpenCV: GPU-accelerated open-source library for computer vision, image processing, and machine learning, now supporting real-time operation.
- FFmpeg: Open-source multimedia framework with a library of plugins for audio and video processing.
- ArrayFire: GPU-accelerated open-source library for matrix, signal, and image processing.
- MAGMA: GPU-accelerated linear algebra routines for heterogeneous architectures, by Magma.
- IMSL Fortran Numerical Library: GPU-accelerated open-source Fortran library with functions for math, signal, and image processing, statistics, by RogueWave.
- Gunrock: Library for graph-processing designed specifically for the GPU.
- CHOLMOD: GPU-accelerated functions for sparse direct solvers, included in the SuiteSparse linear algebra package, authored by Prof.
- Triton Ocean SDK: Real-time visual simulation of oceans, water bodies in games, simulation, and training applications, by Triton..
- CUVilib: Primitives for accelerating imaging applications from medical, industrial, and defense domains.