GPU/CUDA Programming for DNN

Bin ZHOU

Jan. 2015

Lecturer


- Bin ZHOU Ph.D. synosy@gmail.com
- NVIDIA CUDA Fellow、USTC Adjunct Research Prof
- Chief Scientist and Director of Marine Remote Sensing & Information Processing Lab, SDIOI.

- Major: Electronics and Computer Engineering
- Research: Signal, Image & Video Processing, Data Analysis, Cryptography and Crypto-Analysis, UAS
- Other Fields:
- Numerical Methods for Meteorology, Bio-Informatics, Search Engine and Mobile systems.

- Tag: GPU, HPC, UAS...

Prerequisites & References

- Basics
  - 1) Computer Architecture Basics 2) C Programming Language
  - 3) Numerical Methods | Analysis 4) Neural Network
- Materials (Provided)
- 1. CUDA C Programming Guide, NVIDIA Corp.
- 2. CUDA Best Practice Guide, NVIDIA Corp.
- 3. Programming Massively Parallel Processors, 2010, David Kirk and Wen-mei Hwu
- 4. cuDNN references
- References
  - 1.Patrick Cozzi, CIS 565, University of Pennsylvania
  - 2. Udacity CS 344 Intro to Parallel Computing

Contents

- 1) Basics of CUDA (1.5 hour)
- 2) Debugging, Profiling & Tools for CUDA/GPU (1 hour, with Lab. Contents)

- 3) DNN with GPU/CUDA (1.5 hours, with Lab. Contents)
- 4) CUDA Optimization for DNN(1 hour)
- 5) Advanced Topics with Multi-GPU and more. (0.5 hours)

Basics of CUDA

- 1) CPU Architecture Review (done)
- 2) Very Brief Review of Parallel Computing
- 3) Development Environment Configuration & Tools
- 4) GPU Architecture Review
- 5) GPU/CUDA Programming & Memory Model
- 6) CUDA Programming By Examples

Debugging, Profiling & Tools (Lab.)

- 1) Programming, Compiling
- 2) Debugging under windows & Linux
- 3) Profiling for Performance
- 4) Library and Tools

DNN with GPU/CUDA

- 1) Simple neural network with CUDA
- 2) cuDNN and caffe
- 3) Hands-on work for NN, cuDNN

CUDA Optimization for DNN

- General Optimization Procedure & Consideration
- Efficient CUDA Programming Skills
- Memory Throughput Optimization
- DNN Analytical Optimization

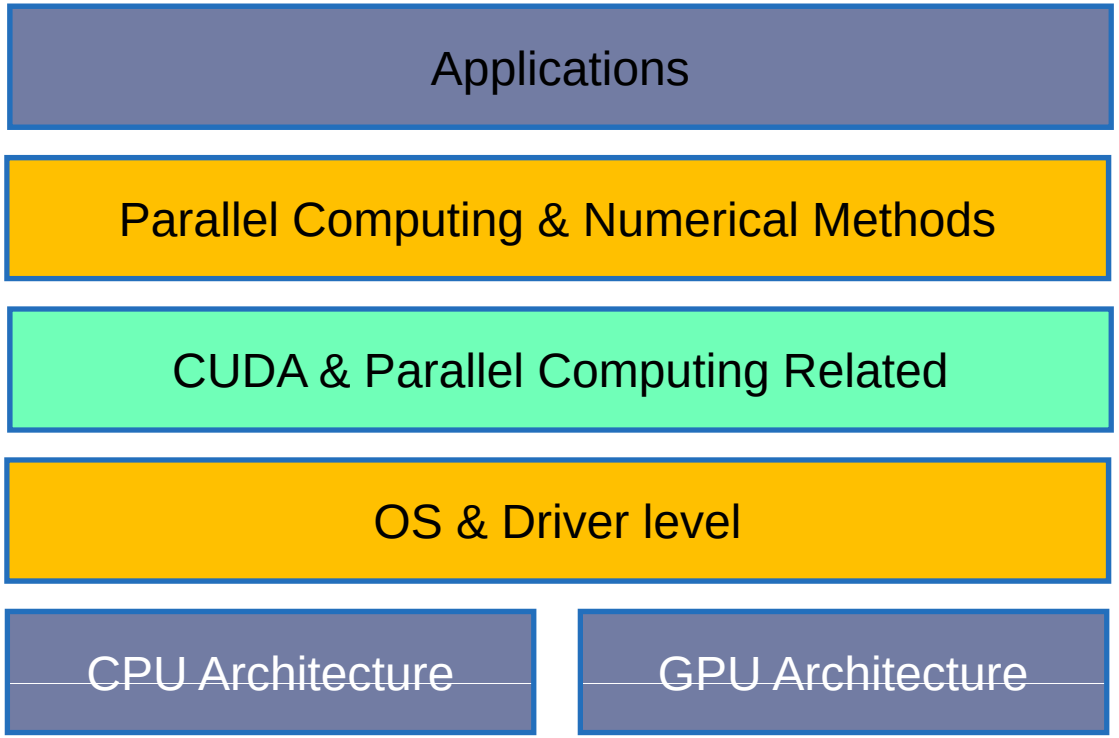
Advanced Topics with Multi-GPU and more

- Multi-GPU, Multi-Node
- RDMA and GPUDirect
- Hyper-Q
- Dynamic Parallelization
- Tegra K1
- …

We’re dealing with GPU/CUDA contents….

- Programming Model?
- Memory Model?
- WARP?
- Occupancy?
- Optimization
  - Compute Bound or memory Bound?
- Others
  - CUDA-GDB
  - Parallel Nsight?

GPU Ecosystems



Thanks and QA…

Let’s CUDA!