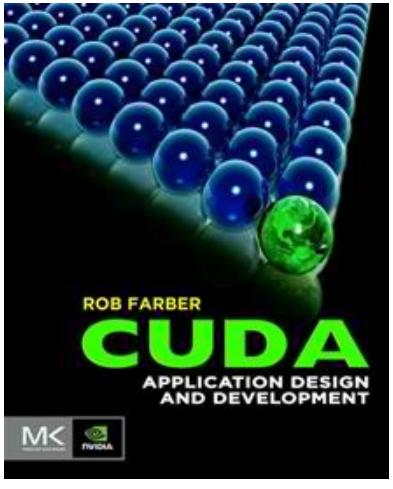

CUDA Application Design and Development

Rob Farber

Book review by Chris Jang

Three questions

1. *Why* read this book?
   - Overview of CUDA technology *platform* and ecosystem
   - Machine learning, what is that?
2. *Who* is the audience?
   - Scientists and engineers familiar with GPGPU
   - Interested in GPGPU as an *applications* platform
   - Not beginners
3. *What* is the value?
   - How to think about stream processing
   - Mind bending, new concepts and ideas

Three cultures

- CUDA technology platform
  - GPGPU development using CUDA and Thrust
  - Third party software package support
- Stream processing on the GPU
  - Thread scheduling
  - Memory hierarchy
  - Map-reduce and locality
- Machine learning
  - Neural networks (cybernetics before big data)
  - Statistical learning theory (big data + linear models)

Three parts

1. Quick start introduction (1, 2, 3)
2. Scheduling and memory (4, 5, 6, 7)
3. Technology stacks and culture (8, 9, 10, 11, 12)

Act 1: Quick start introduction

Three chapters...

1. First Programs and How to Think in CUDA
2. CUDA for Machine Learning and Optimization
3. The CUDA Tool Suite: Profiling a PCA/NLPCA Functor

Act 2: Scheduling and memory

Four chapters. . .

4. The CUDA Execution Model
5. CUDA Memory
6. Efficiently Using GPU Memory
7. Techniques to Increase Parallelism

Act 3: Technology stacks and culture

Five chapters...

8. CUDA for All GPU and CPU Applications
9. Mixing CUDA and Rendering
10. CUDA in a Cloud and Cluster Environments
11. CUDA for Real Problems
12. Application Focus on Live Streaming Video

Theme: GPU architecture

Chapters 4, 5, 6, 7...

- Warp divergence and SIMT
- TLP versus ILP, occupancy versus registers
- Little's law and latency hiding
- GPU memory hierarchy (L2 cache!)
- Memory access pattern for structure of arrays
- Automatic data movement with memory address spaces
- Concurrency and streams

Theme: Application design

Chapters 9, 10, 12...

- Primitive restart, OpenGL renders CUDA output
- MPI: Message Passing Interface
- Image processing filters

Theme: Machine learning

Chapters 2, 3, 11...

- Model families and cost functions
- Nonlinear optimization
- Supervised training set
- Unsupervised goodness metric
- Curse of Dimensionality
- Factor, neighborhood, and network models

Technology summary

About CUDA and NVIDIA GPUs...

- A technology *platform*, more than a shader language
- OpenCL and CUDA are not really comparable
- Mature with good tool and package support
- Architecture emphasizes memory bandwidth

Book summary

About the book...

- The whole world of CUDA in one book
- Helpful for making technology choices
- Multiple technology cultures may be confusing
- Not a how-to guide