Applied Parallel Computing LLC
http://parallel-computing.pro
+41 78 925 90 90
contact@parallel-computing.pro

2-day Course on GPU Computing at the University of Offenburg

(18–19 September 2017)

Exploiting the potential of GPU computing is inevitable for any modern HPC application. As the leader in the application of compute & deep learning technologies, NVIDIA sets up the quality standard for massively parallel hardware, development tools and GPU-accelerated libraries. This course provides essential practical experience for scientists to develop, debug and optimize fast and efficient research codes with NVIDIA CUDA.

Applied Parallel Computing LLC is delivering GPU training courses since 2009. Several dozens of courses have been organized all over Europe, both for commercial and academic customers. We work in close partnership with NVIDIA, CUDA Centers of Excellence and Tesla Preferred Partners. In addition to trainings, our company provides GPU porting/optimization services and CUDA certification.

All corresponding presentations and code samples will be available to attendees in printed handouts.

Day 1: Introduction to CUDA and GPU libraries

Morning (09:00-12:30)

09:00-10:30: lecture

- CUDA principles and CUDA implementation for C++
- Analogies between MPI+OpenMP and CUDA programming models
- The first CUDA program explained
- CUDA compute grid, examples
- Realistic CUDA application example (wave propagation code)
- Understanding GPU compute capabilities, *deviceQuery*
- Basic optimization techniques
- Overview of CUDA applications development using Visual Studio 2015

11:00-13:00: Hands-on session

- Example of *vector addition* in CUDA, compared to OpenACC implementation
- **Hands-on:** Write & deploy a simple CUDA program
- **Hands-on:** More control on CUDA compute grid

13:00-14:00: Lunch

14:00-15:30: Hands-on session

- **Hands-on:** Write & deploy bilinear image interpolation in CUDA

15:30-16:30: GPU-enabled libraries

- Thrust – the C++ library of GPU-enabled parallel algorithms

- CUBLAS, MAGMA, CUBLAS-XT, CUSPARSE, CUFFT and CURAND
- CUSP and AmgX – Krylov and mutigrid solvers
- CUDNN – Deep Neural Network library

16:45-18:00: Hands-on session

- **Hands-on:** solving Poisson equation with CUFFT

Day 2: GPU memory hierarchy, advanced CUDA, optimization & profiling

09:00-10:30: GPU memory hierarchy

- GPU memory types
- Shared memory
- GPU caches hierarchy and mode switches
- Automatic texture cache (Kepler GK110)
- Unified virtual address space (UVA) in CUDA 7.5
- Streams and asynchronous data transfers

10:45-13:00: Hands-on session

- **Hands-on:** "fill-in" exercise on reduction with and without shared memory
- **Hands-on:** getting additional performance using automatic texture cache

13:00-14:00: Lunch

14:00-15:30: Advanced CUDA

- CUDA 9 cooperative groups
- Warp-synchronous programming in CUDA 9
- Dynamic parallelism
- Warp shuffle instruction. Optimizing reduction with shuffles.
- CUDA C++ compiler pipeline, PTX assembler, SASS
- Understanding "-Xptxas -v" reports

15:30-16:30: GPU code optimization

- The cost of global memory allocation
- PCI-E optimizations: streams, asynchronous data transfers
- An overview of Kepler, Maxwell, Pascal and Volta GPU architectures
- GPU optimizations: compute grid, coalescing, divergence, unrolling, vectorization, maxrregcount, aligning, floating-point constants
- Overview of *NVIDIA Visual Profiler*
- Overview of *nvprof* (command line profiler)
- Common practices of identifying performance hazards in GPU application using NVIDIA Visual Profiler

16:45-18:00: Hands-on session

- **Hands-on:** profile and optimize the bilinear interpolation kernel