CUDA GPU Computing Tutorial

Organisers and presenters:

Dominik Göddeke (University of Stuttgart, Germany)

Robert Strzodka (Heidelberg University, Germany)

Manuel Ujaldón (University of Malaga, Spain, and Nvidia CUDA Fellow)

Malte Schirwon (University of Stuttgart, Germany)

The presenters have long-standing research expertise in GPU Computing, dating back to the first era of GPU computing in which graphics APIs were used exclusively for general-purpose computations. They have presented numerous courses, lectures and tutorials on the topic, in academia and industry.

Tutorial description:

In the past 10+ years, GPU Computing has evolved from an obscure niche to a pervasive area of (high performance) computing in both academia and industry. Today's GPUs are high-performance many-core processors capable of very high computation and data throughput. Applications from hundreds of scientific domains have been successfully ported to GPUs with substantial speedups, and several of the world's fastest supercomputers rely on GPUs for their outstanding performance and energy efficiency. This success story has been (and continues to be) enabled by a joint development of the hardware and novel algorithmic techniques for fine-grained parallelism, along with new programming environments, libraries, toolchains, accessible programming interfaces and industry-standard languages such as C. NVIDIA CUDA is the most wide-spread and mature GPU Computing ecosystem.

In short, GPU Computing is no longer an emerging, but rather a pervasive field: GPU Computing today is easier and more accessible than ever, but the challenge remains to achieve a high fraction of the peak hardware performance. Despite ongoing efforts for "drop-in" library support, programmers need to adapt core application-specific functionality specifically for the "new" hardware platform.

The goal of this tutorial is to provide and establish an overview of best practices and generally applicable strategies of "programming for performance", based on instructive examples and a comparatively large practical session.

Schedule:

The 6-hour tutorial comprises three parts, with ample time for discussion at the included coffee breaks:

- Part 1 briefly covers the CUDA ecosystem and the programming model, including but not limited to kernels, grids of blocks of threads, elementary C/C++ language extensions and the CUDA "high level" API. This part requires no a-priori knowledge except a certain degree of fluency in C/C++.
- Part 2 focuses on generally applicable CUDA best practices. Topics covered include a (feasible) programmer's "mental model" of the hardware independent of application domains, as well as techniques like execution configurations, kernel design patterns,

memory access, concurrency optimisation etc. This part targets a mainly intermediate audience, while beginners are not excluded.

- Part 3 allows all participants to "get their hands dirty" by experimenting with the concepts introduced so far, and beyond. We run a "CUDA challenge" where different optimisations can be deployed at beginner, intermediate and advance levels so that everyone can compete on a game to maximize GPU performance. This hands-on will use Amazon EC2 instances (cloud computing) freely provided by NVIDIA. Participants should bring their own laptops with a working installation of an ssh client for command-line access to a

Sessions:

11:00 - 11:40 The CUDA paradigm: Hardware and programing models.

11:40 - 11:50 Break

11:50 - 12:50 Basic optimizations: Warps, blocks, kernels, concurrency.

12:50 - 14:00 Lunch

14:00 - 15:00 Memory optimizations: Shared memory, memory banks, coalescing.

15:00 - 15:10 Break

15:10 - 16:00 Memory innovations: Unified memory (Maxwell), 3D-RAM (Pascal).

16:00 – 16:10 Break

16:10 - 18:00 Hands-on in the cloud (practical individual training).

The time slot for the hands-on is flexible according to everyone's skills and effort. We will provide a maximum time frame of 4 hours on Amazon EC2 instances, and assignments will be unveiled for basic, intermediate and advanced programmers. Attendees are free to leave earlier or resume its session in the hotel room. A quiz will be posed to all programming levels, with participants challenged to maximize GFLOPS on a common GPU model.

The winner will get the prize of a Tesla K40 GPU generously donated by Nvidia for the conference given its role of PPAM sponsor (retail value: 5000 USD).