TUTORIAL: Modern GPU Computing

Organizers and Presenters:

Dominik Göddeke http://www.ians.uni-stuttgart.de/cmcs/
Robert Strzodka https://asc.ziti.uni-heidelberg.de/en
Manuel Ujaldón http://manuel.ujaldon.es/

Abstract:

GPU Computing has evolved from an obscure niche to a commonplace technique in the arsenal of simulation scientists. With the recent introduction of the concept of unified memory and changes to the memory systems in current hardware generations, GPU programming has achieved another important step towards usability, expressability and performance. This is particularly true when combined with modern C++ features such as template metaprogramming, that enable complex custom data structures and algorithms. The goal of this tutorial is to provide a hands-on overview on these recent advances, with a focus on simulation workloads. Each session incorporates a detailed practical with small exercises and experiments to get a feeling for the new features. Attendees not familiar with CUDA yet will have the opportunity to obtain access to a CUDA introduction online course generously provided by NVIDIA's Deep Learning Institute, with a certificate awarded upon successful completion in the weeks after the conference.

Target audience:

The afternoon talks aim at an audience already familiar with modern C++ and NVIDIA CUDA. Participants are expected to bring their own laptops in order to connect to the provided GPU servers with ssh.

Schedule:

Session 1: **11:30 – 13:00**
hardware capabilities of the Pascal, Volta and Turing architectures, brief CUDA roundup
Introduction to the NVIDIA DLI course and the certificate

Lunch: **13:00 – 13:30**

Session 2: **13:30 – 15:30**
high-level improvements: unified memory, scheduling, using GPUs for multiple applications simultaneously

Break: **15:30 – 16:00**

Session 3: **16:00 – 18:00**
low-level improvements: warp synchronisation, interoperability, integer and tensor operations
summary and roundup