

Parallel Computing with GPUs

Anne C. ELSTER ^{a,1} and Stéphane REQUENA ^b

^a *Norwegian University of Science and Technology (NTNU), Trondheim, Norway*

^b *Genci, Paris, France*

Abstract.

The success of the gaming industry is now pushing processor technology like we have never seen before. Since recent graphics processors (GPU's) have been improving both their programmability as well as have been adding more and more floating point processing, it makes them very appealing as accelerators for general-purpose computing. This minisymposium gives an overview of some of these advancements by bringing together experts working on the development of techniques and tools that improve the programmability of GPU's as well as the experts interested in utilizing the computational power of GPU' scientific applications. This first EuroGPU Minisymposium brought together several experts working on the development of techniques and tools that improve the programmability of GPU's as well as the experts interested in utilizing the computational power of GPU's for scientific applications. This short summary thus gives a very useful, but quick overview of some of the major recent advancement in modern GPU computing.

The minisymposium started with a short history and overview as seen by Anne C. Elster, one of the organizers. her talk went directly into her work with her graduate student Rune J. Hovland on *Throughput Computing on Future GPUs*, both overview papers included here. The rest of the minisymposium was divided into two parts, an industrial track described below and an academic track which 4 papers are included here.

1. Industry-related track

This track contained several interesting talks that will be summarized, but not be published here. However, copies of the slides from each can be found on the web at:

<http://www.idi.ntnu.no/~elster/eurogpu09> and eventually also at
<http://www.eurogpu.org/eurogpu09>

OpenCL, a new standard for GPU programming by Francois Bodin(Caps Enterprice) generated a lot of interest. His presentation gave an overview of OpenCL for programming Graphics Processing Units (GPUs). OpenCL is an initiative launched by Apple to ensure application portability across various types of GPUs. It aims at being an open standard (royalty free and vendor neutral) developed by the Khronos OpenCL working group (<http://www.khronos.org>). OpenCL which is based on the ISO C99, shares many features with CUDA and exposes data and task parallelism.

¹ A big thank you to Guillaume Colin de Verdiere from CEA for stepping in for Stephane Requena at the last minute at the ParCo 2009 conference!

Heterogeneous Multicore Parallel Programming by Stephane Bihan (Caps Entreprise, Rennes) presented HMPP, a Heterogeneous Multicore Parallel Programming workbench with compilers, developed by CAPS entreprise, that allows the integration of heterogeneous hardware accelerators in a non-intrusive manner while preserving legacy codes.

Cosmological Reionisation Powered by Multi-GPUs by Dominique Aubert (Universite de Strasbourg and Romain Teyssieer(CEA) took the simulated distribution of gas and stars in the early Universe and modelled the propagation of ionising radiation and its effect on the gas. This modeling will help to understand the radio observations in a near future and the impact of this first stellar light on the formation of galaxies. Their code explicitly solves a set of conservative equations on a fixed grid in a similar manner to hydrodynamics and it follows the evolution of a fluid made of photons. However due to typical velocities close to the speed of light, the stringent CFL condition implies that a very large number of timesteps must be computed, making the code intrinsically slow. However, they ported it to a GPU architecture using CUDA which accelerating their code by a factor close to 80. Furthermore, by using an MPI layer, they also expanded it to a multi-GPU version. CUDATON is currently running on 128 GPUs installed on the new CCRT calculator of the French atomic agency (CEA). The code is able to perform 60 000 timesteps on a 10243 grid in 2.5 hours (elapsed). For comparison, the largest calculation made so far on the same topic involved a 4003 grid and required 11 000 cores to be run. Such a boost in the performance demonstrates the relevance of multi-gpu calculations for computational cosmology. It also opens bright perspectives for a systematic exploration of the impact of the physical ingredients on high resolution simulations since these calculations are extremely fast to complete.

Efficient Use of Hybrid Computing Clusters for Nanosciences by Luigi Genovese (ESFR, Grenoble), Matthieu Ospici (BULL, UJF/LIG, CEA, Grenoble), Jean Francois Mehaut (UJF/INRIA, Grenoble), and Thierry Deutsch (CEA, Grenoble) included their study of the programming and the utilization of hybrid clusters in the field of computational physics. These massively parallel computers are composed of a fast network (Infiniband) connecting classical nodes with multicore Intel processors and accelerators. In our case, the accelerators used are GPUs from NVIDIA. They first analyzed some ways to use with efficiency CPUs cores and GPUs together in a code (BiGDFT, [http://inac.cea.fr/L_Sim/BigDFT\](http://inac.cea.fr/L_Sim/BigDFT/)) without hotspot routines. Starting from this analysis, they have designed a new library: S_GPU, used to share GPUs between the CPU cores of a node. The implementation and the usage of S_GPU was described. They then evaluated and compared performances between S_GPU and others approaches to share GPUs with CPUs. This performance evaluation was based on BigDFT, an ab-initio simulation software designed to take advantage of massively hybrid parallel clusters as the Titane cluster (CCRT). Their experiments was performed on both one hybrid node as well as on a large number of nodes of their hybrid cluster.

Accelerating a Depth Imaging Seismic Application on GPUs: Status and Perspectives by Henri Calandra(TOTAL, Pau) described how the extraordinary challenge that the oil and gas industry must face for hydrocarbon exploration requires the development of leading edge technologies to recover an accurate representation of the subsurface. Seismic modeling and Reverse Time Migration (RTM) based on the full wave equation

discretization, are tools of major importance since they give an accurate representation of complex wave propagation areas. Unfortunately, they are highly compute intensive. He first presented the challenges in O&G and the need in terms of computing power for solving seismic depth imaging problems. He then showed how GPUs can be part of the solution and the the solutions developed at TOTAL.

Debugging for GPUs with DDT, David Lecomber (Allinea Ltd, Bristol, UK) described how evelopers are experimenting with CUDA, OpenCL and others to port (or rewrite) their code to take advantage of this technology, but are discovering there is more to programming than writing code. Finding bugs and optimizing performance are essential tasks - particularly so with a new complex model of program execution. He reveiwed the state of play - exploring what is possible now, and what is being done by Allinea and others to improve the lot of GPU developers who need to debug or optimize their codes.

2. Academic track

The above talks were followed by a more academic track the following day whose corresponding papers along with the introductory papers by Elster and her students, are included in proceedings. The secoond day GPU papers include two papers looking at GPU solvers:

- **Porus Rock Simulations and Lattice Boltzmann on GPUs** by Erik Ola Aksnes and Anne C. Elster (both NTNU, Norway) which looked at using the Lattice Boltzman Method on large 3D datasets on GPUs for fluid similatons, and
- **An efficient multi-algorithms sparse linear solver for GPUs** by Stéphane Vialle; Thomas Jost and Sylvain Constassot-Vivier (all from Supé lec Campus de Metz, Franc) which discussed how to implement a sparse solver on GPUs.

The EuroGPU 2009 workshop/minisymposium was rounded off by two GPU modeling presentations:

- **Abstraction of Programming Models Across Multi-Core and GPGPU Architectures** by Ian Grimstead and David R. Walker (Cardiff University, UK) , and
- **Modeling Communication on Modern GPU Systems** by Anne C. Elster, Daniele G. Spampinato and Thorvald Natvig (all from NTNU, Norway).

The last two papers should provide tools for those who want to get a good feel for the performance potential CPU and multi-core CPU versus GPU and multi-GPU systems.

The overview presentation and presentation on throughout computing on GPU resulted, along with the four academic presentations, in the following papers. These publications together with the on-line PDF files of the presentations from the industrial track should provide a great resource for those who want to take a closer look at how one can harness the great computational power of modern GPUs.