ST project 2012: “GPU cluster computing”

Exploiting massive parallelism using the 7680 CUDA core compute cluster

Last update: 25th of September, 2012
Contents

- Project overview
- Project goals
- Timeline
- People involved
- Application
  - Application: Sift object recognition
- Infrastructure
Project overview

• **Analysis** of a given application

• **GPU programming:**
  • Mapping of kernels to a single GPU
  • Extension to multiple GPUs
  • Extension to GPUs in multiple systems
  • Extension to both CPUs and GPUs in the cluster

• **Enable load-balancing and dynamic scheduling** in the cluster

• **Evaluation of GPU and cluster computing**
Project goals

- Learn the details of an HPC application

- Program for:
  - GPU accelerators (CUDA)
  - Multi-core processors and multiple systems (OpenMP, MPI, SSE)

- Experience and evaluate:
  - Multi-core and accelerator programming
  - Heterogeneous programming
  - Cluster computing
  - HPC and supercomputing
Project (sub)goals

- **Work in a team**
  - Distribute and synchronize tasks accordingly
  - Multi-disciplinary, intercultural groups

- **Porting** sequential code to the compute cluster
  - Identify critical and parallelizable kernels
  - Handle memory copies between processors/systems
  - Use a restricted subset of C for kernel functions

- **Parallelize** a given application
  - Orchestrate communication and synchronization
  - Exploit task- and data-level parallelism
Project deliverables

At the end of the project, you are expected to provide the following deliverables:

- The GPU-cluster ported source code
- A group presentation
- A group paper
- A one-page HTML webpage per group
Timeline (1/2)

Schedule
The project is divided in three parts:

• [25 sep - 28 sep] Introduction, lecture and tutorial week
• [02 okt - 12 okt] Application: SIFT object recognition

The lectures will be held in MetaForum MF 6.132. Lecture slides are available at the bottom of the resources page, and tutorial slides and details are available through the tutorials page.

Introduction, lecture and tutorial week

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Event</th>
<th>Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tue 25 Sep</td>
<td>10:30 - 11:00</td>
<td>Project introduction</td>
<td>Maurice &amp; Gert-Jan</td>
</tr>
<tr>
<td></td>
<td>11:00 - 12:00</td>
<td>Tutorial #1a: OpenMP introduction (1/2)</td>
<td>Maurice &amp; Gert-Jan</td>
</tr>
<tr>
<td></td>
<td>13:00 - 14:30</td>
<td>Lecture #1: Processor architectures</td>
<td>Henk Corporaal</td>
</tr>
<tr>
<td></td>
<td>14:30 - 15:30</td>
<td>Lecture #2: Parallel architectures</td>
<td>Henk Corporaal</td>
</tr>
<tr>
<td></td>
<td>15:30 - 17:00</td>
<td>Tutorial #1a: OpenMP introduction (2/2)</td>
<td>Maurice &amp; Gert-Jan</td>
</tr>
<tr>
<td>Wed 26 Sep</td>
<td>09:00 - 12:30</td>
<td>Tutorial #1b: SSE introduction</td>
<td>Maurice &amp; Gert-Jan</td>
</tr>
<tr>
<td></td>
<td>13:30 - 15:00</td>
<td>Lecture #3: GPU architectures</td>
<td>Zhenyu Ye</td>
</tr>
<tr>
<td></td>
<td>15:30 - 17:00</td>
<td>Tutorial #1c: OpenMP &amp; SSE</td>
<td>Maurice &amp; Gert-Jan</td>
</tr>
<tr>
<td>Thu 27 Sep</td>
<td>09:30 - 10:30</td>
<td>Tutorial #2: CUDA introduction</td>
<td>Gert-Jan &amp; Maurice</td>
</tr>
<tr>
<td></td>
<td>10:30 - 12:00</td>
<td>Tutorial #3: CUDA memory usage and strategies (1/2)</td>
<td>Gert-Jan &amp; Maurice</td>
</tr>
<tr>
<td></td>
<td>13:30 - 15:00</td>
<td>Tutorial #3: CUDA memory usage and strategies (2/2)</td>
<td>Gert-Jan &amp; Maurice</td>
</tr>
<tr>
<td></td>
<td>15:00 - 17:00</td>
<td>Tutorial #4: CUDA profiling, tools and libraries</td>
<td>Gert-Jan &amp; Maurice</td>
</tr>
</tbody>
</table>
# Timeline (2/2)

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Event</th>
<th>Instructor(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fri 28 Sep</td>
<td>09:00 - 10:00</td>
<td>Tutorial #5: CUDA advanced concepts (1/3)</td>
<td>Gert-Jan &amp; Maurice</td>
</tr>
<tr>
<td></td>
<td>10:00 - 12:30</td>
<td>Tutorial #5: CUDA advanced concepts (2/3)</td>
<td>Gert-Jan &amp; Maurice</td>
</tr>
<tr>
<td></td>
<td>13:30 - 17:00</td>
<td>Tutorial #5: CUDA advanced concepts (3/3)</td>
<td>Gert-Jan &amp; Maurice</td>
</tr>
<tr>
<td>Tue 02 Okt</td>
<td>09:30 - 10:00</td>
<td>Kick-off SIFT project (MF 12)</td>
<td>Maurice &amp; Gert-Jan</td>
</tr>
<tr>
<td></td>
<td>10:30</td>
<td>Start of the project (SIFT application)</td>
<td></td>
</tr>
<tr>
<td>Fri 05 Okt</td>
<td>13:30 - 14:30</td>
<td>Lecture #4: Supercomputers</td>
<td>Gert-Jan &amp; Maurice</td>
</tr>
<tr>
<td></td>
<td>14:30 - 15:30</td>
<td>Tutorial #6: CUDA and OpenMP (multi-GPU)</td>
<td>Gert-Jan &amp; Maurice</td>
</tr>
<tr>
<td></td>
<td>15:30 - 16:30</td>
<td>Tutorial #7: CUDA and MPI (multi-machine)</td>
<td>Gert-Jan &amp; Maurice</td>
</tr>
<tr>
<td></td>
<td>16:30 - 17:00</td>
<td>Tutorial #8: Combining CUDA, OpenMP and MPI</td>
<td>Gert-Jan &amp; Maurice</td>
</tr>
<tr>
<td>Fri 12 Okt</td>
<td>15:00</td>
<td>End of the project</td>
<td></td>
</tr>
<tr>
<td></td>
<td>15:00 - 17:00</td>
<td>Group presentations (MF 12)</td>
<td></td>
</tr>
</tbody>
</table>
People involved (1/2)

- **Coordination:**
  - Henk Corporaal

- **GPU programming:**
  - Gert-Jan van den Braak, Maurice Peemen

- **Lectures**
  - Henk Corporaal, Zhenyu Ye, Gert-Jan van den Braak
  - Maurice Peemen

- **Application experts:**
  - Zhenyu Ye, Gert-Jan van den Braak, Maurice Peemen
People involved (2/2)

• Lectures and tutorials:
  • GPU architecture lectures (Zhenyu, Henk)
  • Cluster computing lectures (Gert-Jan, Henk)
  • CUDA programming tutorials (Gert-Jan, Maurice)
  • OpenMP, SSE programming tutorials (Maurice, Gert-Jan)
  • Application specific lectures (application experts)
Contents

• Project overview
• Project goals
• Timeline
• People involved
• Application Sift Object recognition
• Infrastructure
Applications

• What are our application requirements?
  • Large code source:
    – Enough to provide significant work
  • Expertise:
    – Application-experts are appointed
  • Execution time:
    – Should be significant (in the range of hours)
  • Suitability for GPU’s/clusters:
    – Must consist of (multiple) parallelizable algorithms
  • Use of this application:
    – The application needs to be relevant
Application SIFT object recognition
Infrastructure

- **Tooling:**
  - MPI, OpenMP and CUDA installed on all systems
- **Hardware:**
  - Compute cluster with four nodes and one host
  - Each node contains multiple GPUs and a quad-core CPU
  - Each GPU contains many smaller processing elements
  - Around 2 TFLOPS per GPU (32 TFLOPS for the cluster)
Infrastructure

- 4 desktop computers, 16 GPUs total
  - Max power consumption ~ 4.000W
  - GPU cores 7.680 CUDA cores
  - Peak GPU FLOPS ~ 22.400 GFLOPS
  - Accumulated GPU bandwidth ~ 2.400 GB/s
Infrastructure

- 4 desktop computers, 16 GPUs total