Mathematical computations with GPUs

Introduction

Alexey A. Romanenko
arom@ccfit.nsu.ru
Novosibirsk State University
How to..

* Process terabytes of seismic data on desktop PC.
* Estimate tsunami impact faster than wave reaches dry earth.
* Rasterize 3D scenes in real-time.
Topics of the course

* architecture of a GPU and its difference from general-purpose CPU;
* general notions related to CUDA;
* contents and functionality of GPU-optimized libraries;
* area of applicability of GPUs for solving scientific and applied problems;
* general principles of optimizing programs for GPUs;
* methods of debugging and profiling programs running on GPUs;
* tools and instruments for developing, debugging and profiling programs running on GPUs.
Computer power

«K computer» 2011
68 544 x 8-core SPARC64 VIIIfx.
8,162 Pflops

«Titan» 2012
299 008 Opteron Cores
18 688 k20 NVIDIA GPU
20+ Pflops

Performance

Time

280 Tflops
212,992 CPUs
Growth of performance

* Growth of clock rate
* Number of cores/CPUs
* CPU architecture complexity
  * Number of registers
  * Pipe-line
  * Digital capacity (4bit CPU, 16, 32, 64)
  * etc
Computer graphics
Computer graphics

* Small vectors
* Small matrixes
* Filters/post-processing
* Calculation of projections
* etc.

* Huge number of identical small operations.
Performance of video cards

Theoretical GFLOP/s

- NVIDIA GPU Single Precision
- NVIDIA GPU Double Precision
- Intel CPU Double Precision
- Intel CPU Single Precision

- GeForce GTX TITAN
- GeForce GTX 680
- GeForce GTX 580
- GeForce GTX 480
- GeForce GTX 280
- GeForce 8800 GTX
- GeForce 7800 GTX
- GeForce 6800 Ultra
- Woodcrest
- Tesla C1060
- Harpertown
- Tesla C2050
- Tesla M2090
- Sandy Bridge
- Intel Core i7-3930K (Ferrymont)
- Intel Core i9-10850K (Rocket Lake)
- Intel Core i9-11900K (Rocket Lake)
- Intel Core i9-11900KF (Rocket Lake)
- Intel Core i9-12900K (Alder Lake)
- Intel Core i9-12900KF (Alder Lake)

Timeline:
- Apr-01
- Sep-02
- Jan-04
- May-05
- Oct-06
- Feb-08
- Jul-09
- Nov-10
- Apr-12
- Aug-13
- Dec-14
* Kepler: 28nm, 1.4 Tflops DP
* Maxwell: 22nm, 4 Tflops DP
Memory bandwidth

Graph showing the theoretical GB/s for different GPU models over the years from 2003 to 2012.
Heterogeneous systems

3 of 5 leaders in top500 list!
Heterogeneous vs. Homogeneous computing systems (06.2012)

**Tianhe-1A (2 place)**
- Xeon X5670 6C 2.93 GHz, NVIDIA 2050
- 4.7 Pflops (peak)
- 4040 kW

**Jaguar (3 place)**
- Opteron 6-core 2.6 GHz
- 2.3 Pflops (peak)
- 6950 kW
Toward exaflop

Direct scaling...

* **Tianhe-1A**
  * 4.7 Pflops (peak) * 213 = 1001.1 Pflops
  * 4040 kW * 213 = 860 MW

* **Jaguar**
  * 2.3 Pflops (peak) * 435 = 1000.5 Pflops
  * 6950 kW * 435 = 3023 MW
Sayano–Shushenskaya Dam

<table>
<thead>
<tr>
<th>Category</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turbines</td>
<td>10 × 640 MW (initial)</td>
</tr>
<tr>
<td></td>
<td>6 × 640 MW (current)</td>
</tr>
<tr>
<td>Installed capacity</td>
<td>5,120 MW (current)</td>
</tr>
<tr>
<td>Maximum capacity</td>
<td>6,400 MW</td>
</tr>
<tr>
<td>Annual generation</td>
<td>23.5 TW</td>
</tr>
</tbody>
</table>
Green500 (November 2013)
  * [www.green500.org/lists](http://www.green500.org/lists)
  * List of most power effective supercomputer
  * Ten of top have NVIDIA GPUs. Two – 6 month ago.

Performance / Watt
  * 4,5 – now
  * 0,3 – November, 2007

Reaching exascale means boosting speeds by 50-100 times, while keeping power relatively static.
### Application speedup

<table>
<thead>
<tr>
<th>Example Applications</th>
<th>URL</th>
<th>Application Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seismic Database</td>
<td><a href="http://www.headwave.com">http://www.headwave.com</a></td>
<td>66x to 100x</td>
</tr>
<tr>
<td>Mobile Phone Antenna Simulation</td>
<td><a href="http://www.acceleware.com">http://www.acceleware.com</a></td>
<td>45x</td>
</tr>
<tr>
<td>Molecular Dynamics</td>
<td><a href="http://www.ks.uiuc.edu/Research/vmd">http://www.ks.uiuc.edu/Research/vmd</a></td>
<td>21x to 100x</td>
</tr>
<tr>
<td>Neuron Simulation</td>
<td><a href="http://www.evolvedmachines.com">http://www.evolvedmachines.com</a></td>
<td>100x</td>
</tr>
<tr>
<td>MRI processing</td>
<td><a href="http://bic-test.beckman.uiuc.edu">http://bic-test.beckman.uiuc.edu</a></td>
<td>245x to 415x</td>
</tr>
<tr>
<td>Atmospheric Cloud Simulation</td>
<td><a href="http://www.cs.clemson.edu/~jesteel/clouds.html">http://www.cs.clemson.edu/~jesteel/clouds.html</a></td>
<td>50x</td>
</tr>
</tbody>
</table>
Approaches to GPU programming

Application

- Optimized libraries
- Compiler directives
- Programming languages (C/C++/FORTRAN)

Fast development
Maximum performance
CUDA Roadmap

* CUDA 1.0 – 2007
* CUDA 2.0 – 2008
* CUDA 3.0 – 2009
* CUDA 4.0 – 2011
* CUDA 5.0 – 2012
* CUDA 5.5 – 2013
* CUDA 6.0 – 2014

2014 – 7 years!
CUDA vs. OpenCL

**CUDA**
- NVIDIA GPU (Cray, HP, IBM, T-Platforms, NextIO...)
- Close to peak performance
- Functionality
- Comfort for developers (debugger, performance analyzer, etc.)
- Support
- Teaching materials, libraries

**OpenCL**
- Architecture is not fixed, universality
- Performance is not high priority
CUDA vs. OpenCL performance

CUDA applications have up to 30% higher performance than OpenCL application.

Useful links

* http://www.nvidia.com/object/cuda_home.html
* http://www.gputechconf.com/page/home.html
* http://docs.nvidia.com/