Real-Time Parallel Computing Using GPUs

Anne C. Elster & her students
Dept. of Computer & Info. Science
NTNU (Norwegian Univ. of Sci.&Tech.)
PARA 2008 & 2010

Para 2008 Proceedings hopefully in by end of year

Most extended abstracts now on web

PARA 2010 will be held June 6-9 in REYKJAVIK, ICELAND
ParCo-related Activities & Events:

**EU COST Action IC0805: Open European Network for High Performance Computing on Complex Environments (2009-2014)**

Meeting in Lisboa, mid-October. See: [www.complexhpc.org](http://www.complexhpc.org)

Talk to Elster re. WG on Numerical Algorithms
HPC History: Personal perspective

- 1980's: Concurrent and Parallel Pascal
- 1986: Intel iPSC Hypercube
  - CMI (Bergen) and Cornell (Cray at NTNU)
- 1987: Cluster of 4 IBM 3090s
- 1988-91: Intel hypercubes
  - Some on BBN
- 1991-94: KSR (MPI1 & 2)
- 1995-2005: SGI systems (some IBM SP)
- 2001-current: Clusters
- 2006:
  - IBM Supercomputer @ NTNU (Njord, 7+ TFLOPS, proprietary switch)
  - GPU programming (Cg)
- 2008:
  - Quadcore Supercomputer at UiTø (Stallo)
  - HPC-LAB at IDI/NTNU opens with
    - several NVIDIA donation
    - Several quad-core machines (1-2 donated by Schlumberger)
- 2009: More NVIDIA donations:
  - NVIDIA Tesla s1070 and
  - two Quadro FX 5800 cards (Jan ’09)
The Wal-Mart Effect
(PARA02)

• Wal-Mart – bigger than Sears, K-mart and JC Penney’s combined
  ➔ predicted to influence $40 billion of IT investments (MIT Review)
  ➔ has much more impact than Microsoft and Cisco could ever hope for…

  – Not driven by latest technology, but by business model
  – bad news for HPC?

  – Game market --> HPC market ➔ Future high-performance chips and systems --> NVIDIA Tesla!
"COT"-based SUPERCOMPUTER HARDWARE TRENDS:

• Intel iPSC (mid-1980’s)
  – The first iPSC had no separate communication processor ...
  – Specialized OS
  – 2-128 nodes

• Today’s PC clusters
  – Fast Ethernet or better (more expensive interconnect)
  – Linux OS
  – 32-bit cheapest, but many 64-bit cluster vendors 😊
  – Top500 supercomputers

Today’s GPU farms entering Top500 list..
Clustis 1 ca. 2003

SGI Altix

SGI Griddur/Embla

HPC Hardware Trends at NTNU/IDI

www.ntnu.no

PPAM 2009 Elster www.idi.ntnu.no/~elster/hpc-lab
NTNU’s Supercomputer (“Njord”) from IBM

- Also runs Norway’s operational weather forecasts (met.no)
- NOK 30 millions for system
- NOK 20 million for infrastructure
  - incl. new machine room, back-up power generator & batteries ++

Jørn Amundsen, Roar Skålin (it-sjef met.no) og Bjørn Lindi ved Njord

Comparisons:
- Our 1986-Cray cost NOK 130 million in today’s currency!
- NVIDIA s1070 Tesla w/ 960 cores costs ca NOK 100 000! (4TFlops)
HPC Hardware Trends at IDI

NVIDIA Tesla card

Unpacking NVIDIA s1070 and Quadro FX 5800 cards
GPUs: Graphical Processor Units

HISTORY:
• Late 70’s/ Early´80’s: Grafic drawing calculations on CPUs
• Xerox Alto computer: first special *bit block transfer* instruct
• Comodore Amiga: first mass-market video accelerator able to draw fills shapes & animations in HW. Graphics sub-system w/ several chips, incl. Dedicated to *bit blk xfer*
• Early 90’s: 2D accelleration
• Ca. 1995: VIDEO GAMES! --> 3D GPUs
GPU History continued:

• 1995-1998:
  – 3D rasterization
    (converting simple 3D geometric primitives
    (e.g. lines, triangles, rectangles) to 2D screen pixels)
  – Texture mapping
    (mapping 2D texture image to planar 3D surface)

• 1999-2000: 3D translation, rotation & scaling

• Towards 2000: GPUs more configurable

• 2001-2007: programmable
  (ability to change individual pixels)

• 2008 and beyond: more programmable
  (NVIDIA CUDA, OpenCL ...)

www.ntnu.no
The Nvidia Tesla Architecture

SIMD!
General programming on GPUs

- Rendering = executing
- GPU textures = CPU arrays
- Fragment shader programs = inner loops
- Rendering to texture memory = feedback
- Vertex coordinates = computational range
- Texture coordinates = Computational domain

- Now have NVIDIA´s CUDA library! (BLAS & FFT)
Limitations

• Branching usually not a good idea

• Random memory access problematic

• GPU cache is different from CPU cache
  – Optimized for 2D locality

Floating point precision

Session 1: GPU Basics
- Introduction (Strzodka) [PDF](#)
- Why GPUs? (Strzodka) [PDF](#)
- Prog. Environments and Ready-to-use Libraries (Goedeke) [PDF](#)
- GPU Architecture (Strzodka) [PDF](#)

Session 2: Introduction to OpenCL
- Introduction to OpenCL (Behr) [PDF](#)
- Hands-on Examples

Session 3: OpenCL Basics
- OpenCL API (Behr) [PDF](#)
- OpenCL Language (Behr) [PDF](#)
- Hands-on Examples

Session 4: Scientific Computing on GPUs
- Aspects of Scientific Computing on GPUs (Strzodka) [PDF](#)
- Case Study: GPU Cluster Computing for FEM (Goedeke) [PDF](#)

Session 5: Advanced OpenCL
- OpenCL Architecture and Optimization on AMD GPUs (Behr) [PDF](#)
- Hands-on Examples
- AMD OpenCL GPU Demo
HPC-Lab at IDI
OpenCL

Goals:

- leverage all computational resources in the system.
- unleash computational power to everyone;
- to be integrated with ordinary commercial applications. Example: next gen. games.

- For any data-parallel algorithm.
- Uses GPU, CPU, or both as compute device. Or any combination of GPUs and CPUs. Designed for heterogeneous parallel data computation.
- Simple and clean API.
- A Khronos group standard. Royalty free.
- Future: Support for multiple devices on multiple platforms. Implementations on the way from AMD, NVIDIA and Apple.
- Platform + device independence in focus. Must be in order to become successful.
OpenGL Overview

Your program

Compute Kernel

Data Stream

Data Stream

OpenCL
Platform layer, runtime and compiler

The runtime compiles the kernel, optimizes for the target device(s).

Compute device(s)

CPU  CPU  CPU  CPU  GPU  GPU
Recent Activities & Events:

IEEE International Parallel & Distributed Processing Symposium

MTAAP'09 (Friday, May 29)
Workshop on Multithreaded Architectures and Applications

Two presentations from HPC-Lab:

- Jan Chr. Meyer and Anne C. Elster: Super-Fast Adaptable Bit-Reversal on Multithreaded Architectures

- Daniele Spampinato and Anne C. Elster: Simplex-Based Linear Optimization on Multithreaded Architectures (Linear Programming on GPU!)
EuroGPU´09 Organizers

Anne C. Elster, NTNU, Norway

and

Stephane Requena, Genci, Paris, France

with

Guillaume Colin de Verdière, CEA
# EuroGPU at ParCo 2009:

## TUESDAY Sept. 1, 2009

<table>
<thead>
<tr>
<th>Time</th>
<th>Title/Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:30-12:00</td>
<td><strong>Intro. – GPU Computing</strong>, Anne C. Elster, Norwegian University of Science and Technology (NTNU), Trondheim, Norway</td>
</tr>
<tr>
<td>12:00-12:30</td>
<td><strong>Throughput Computing on Future GPUs</strong>&lt;br&gt;Rune Johan HOVLAND and Anne C. ELSTER, NTNU, Norway</td>
</tr>
<tr>
<td>12:30-14:00</td>
<td>LUNCH</td>
</tr>
<tr>
<td>14:00-14:30</td>
<td>I1: <strong>OpenCL a new standard for GPU programming</strong>&lt;br&gt;Francois BODIN -- Caps Entreprise, Rennes, France</td>
</tr>
<tr>
<td>14:30-15:00</td>
<td>I2: <strong>Heterogeneous Multicore Parallel Programming</strong>&lt;br&gt;Stéphane BIHAN** Caps Entreprise, Rennes, France</td>
</tr>
<tr>
<td>15:00-15:30</td>
<td>I3: <strong>Cosmological reionisation powered by multi-GPUs</strong>&lt;br&gt;Dominique AUBERT\textsuperscript{a}, Romain TEYSSIER\textsuperscript{b}&lt;br&gt;\textsuperscript{a} Université de Strasbourg, France \textsuperscript{b} CEA, France</td>
</tr>
<tr>
<td>15:30-16:00</td>
<td>Coffee/Tea Break &amp; Exhibitions</td>
</tr>
<tr>
<td>16:00-16:30</td>
<td>I3: <strong>Efficient use of hybrid computing clusters for nanosciences</strong>&lt;br&gt;Lugi GENOVESE\textsuperscript{a}, Matthieu OSPICI\textsuperscript{b}&lt;br&gt;Jean Francois MÉHAUT\textsuperscript{c}, Thierry DEUTSCH\textsuperscript{d}&lt;br&gt;\textsuperscript{a} ESFR, Grenoble, \textsuperscript{b} BULL, UJF/LIG, CEA, Grenoble&lt;br&gt;\textsuperscript{c} UJF/INRIA, Grenoble, \textsuperscript{d} CEA, Grenoble, France</td>
</tr>
<tr>
<td>16:30-17:00</td>
<td>I4: <strong>Accelerating depth imaging seismic application on GPUs, status and perspectives</strong>, Henri CALANDRA, TOTAL, Pa u</td>
</tr>
<tr>
<td>17:00-17:30</td>
<td>I5: <strong>Debugging for GPUs with DDT</strong>&lt;br&gt;David LECOMBER Allinea Ltd, Bristol, UK</td>
</tr>
<tr>
<td>19:00-20:00</td>
<td>Reception: City Hall</td>
</tr>
</tbody>
</table>
## WEDNESDAY Sept. 2, 2009  Euro GPU 2009 – DAY 2

<table>
<thead>
<tr>
<th>Time</th>
<th>Title/Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:00-10:30</td>
<td><strong>Porus Rock Simulations and Lattice Boltzmann on GPUs</strong>&lt;br&gt;<strong>Erik Ola AKSNES and Anne C. ELSTER, NTNU, Norway</strong></td>
</tr>
<tr>
<td>10:30-11:00</td>
<td><strong>An efficient multi-algorithms sparse linear solver for GPUs</strong>&lt;br&gt;<strong>Stéphane VIALLE; Thomas JOST and Sylvain CONSTASSOT-VIVIER, Supélec Campus de Metz, France</strong></td>
</tr>
<tr>
<td>11:00-11:30</td>
<td>Coffee/Tea Break &amp; Exhibitions</td>
</tr>
<tr>
<td>11:30-12:00</td>
<td><strong>Abstraction of Programming Models Across Multi-Core and GPGPU Architectures</strong>, <strong>Ian GRIMSTEAD and David R. WALKER, Cardiff University, UK</strong></td>
</tr>
<tr>
<td>12:00-12:30</td>
<td><strong>Modeling Communication on Modern GPU Systems</strong>, <strong>Anne C. ELSTER, Thorvald NATVIG, and Daniele G. SPAMPINATO, NTNU, Norway</strong></td>
</tr>
<tr>
<td>12:30-14:00</td>
<td>LUNCH</td>
</tr>
<tr>
<td>14:00-15:00</td>
<td><strong>PANEL DISCUSSION – ParCo 2009</strong></td>
</tr>
<tr>
<td>15:00-15:30</td>
<td>Coffee/Tea Break &amp; Exhibitions</td>
</tr>
<tr>
<td>15:30-16:30</td>
<td><strong>PANEL DISCUSSION on GPU Computing</strong>&lt;br&gt;Including informal presentation by <strong>Rune J ensen NTNU/CERN on Compiler Issues and Challenges</strong>&lt;br&gt;PANEL:&lt;br&gt;- Anne C. Elster (Organizer) – GPU scientific computing/academia&lt;br&gt;- Guillaume Coline Vedérié, CEA, France – GPU Applications&lt;br&gt;- Tim Lanfear, Nvidia, UK – GPU HW vendor&lt;br&gt;- Stéphane Vialle, Supelec Metz, France – GPU financial computing/academia</td>
</tr>
<tr>
<td></td>
<td>Excursion &amp; Conference Dinner</td>
</tr>
</tbody>
</table>

[Event website link](www.idi.ntnu.no/~elster/hpc-lab)
Rune Jensen: Optimizing BLAS

• Beat ATLAS, an auto-tunable BLAS library routine

• ATLAS
  – Self-tuning at install
  – Compiles lots of code versions
  – Compares their speed
  – Find patterns
  – Makes the best code for your CPU/memory/mainboard
  – Cores handwritten in assembly

– Now moving to multi-core. How about GPUs...
Courses Taught by Dr. Elster:

TDT4200 Parallel Computing
(Parallel programming with MPI, threads and NVIDIA CUDA)

TDT24 Parallel environments & Numerical Computing
- 2-day IBM CellBE Course (Fall 2007)
- GPU & Thread programming

TDT 4205 Compilers

DTD 8117 Grid and Heterogeneous Computing
HPC -LAB Sponsors/Collaborators

- NVIDIA

- CERN (birthplace of Internet; EU’s largest GRID project; Norwegian CTO (Sverre Jarp) :-)  
  - 7 Master students from my group have had summer jobs there  
  - 4 took their Master’s thesis there  
  - 1 still a staff member there, 2 will join this summer

- GE Healthcare  
  Our GPU wavelet algorithm now in their high-end cardiac ultrasound scanner!

- SCALI (Commercial MPI implementer)

- Schlumberger (formerly Voxel Vision)

- StatoilHydro

- several other departments at NTNU including Petroleum, Physics and Chemistry
The IDI HPC-Lab focuses on research related to novel GPU and multi-core architectures

- GPGPU for HPC
- Parallel and Distributed Algorithms
- Performance Evaluation and Benchmarking
- Parallelization of Seismic and Image Related Applications on GPUs and Multi-Cores
- Adaptive and Auto-Tuneable Algorithms and Implementations

Collaborators / Supporters: ARM, CERN, NVIDIA, StatoilHydro, Schlumberger, GE-Healthcare, and others
HPC Research Group - Spring 2008

GROUP MEMBERS:
- Assoc. Prof. Anne C. Elster
- Adjunct Assoc. Prof-Jørn Amundsen
- Henrik Nagel (PostDoc/now at NTNU HPC-Ctr)

- 2 PhD students
  - Jan Christian Meyer
  - Thorvald Natvig

- Recent Master students (grad. dates)
  - Øystein Borgen(June´06) - Schlumberger
  - Ingar Saltvik (June´06) - Fast
  - Nils Magnus Larsgård (Aug´06) - IBM
  - Erik Axel Nielsen (May´07) - consultant
  - Idar Borlaug (June´07) - StatoilHydro cons
  - Knut Imar Hagen (June´07) - --” --
  - Leif Christian Larsen (June´07) - Roxar
  - Jerôme Dubois (Feb ´08) - back in France
  - Andreas Bach (su ´08) - Uninett
  - Atle Rudshaug (su´08) - Numerical Rocks
  - Robin Eidissen (Feb. ´09) - IDI
  - Rune Jensen (May´09) - CERN
High-Performance Computing Group

Alumni

Master Student Alumni

Andreas Bach (2008)
Atle Rudshaug (2008)
Idar Borlaug (2007)
Thibault Collet (2007)
Knut Imar Hagen (2007)

Nils Magnus Larsgård (2007)
Erik Axel Nielsen (2007)
Christian Larsen (2007)
Øystein Borg (2006)
Ingvar Saltvik (2006)
HPC-LAB  Spring/Summer 2009
Master projects

- 1 worked on real-time snow simulations
- 1 works on “beating Atlas”
- 2 work with IO Center on flow through porous media on GPU
- 1 work with Schlumberger on line finding algorithms on GPU
- 1 works on GPU-CPU system configurations
- 1 work with Statoil-Hydro on check-pt restart of large applications
- 1 on sound processing on GPU
- 1 on LP on GPU
- 1 on special transforms for GPU
Smoke & Snow Particle Simulations

Started out as Real-time smoke simulation on dual core laptop (Torbjørn Vik, 2003)

Crude snow simulation on Multicore (Ingar Saltvik, 2007)

Snow simulation simulating several million snowflakes as particles, wide field interactions++ Using compute-power of the GPU (Robin Eidissen, 2009)

Paralelized Snow & Smoke Simulations
SNOW SIMULATION DEMO!

Robin Eidissen
(Teaching Assistant)
Åsmund Eldhuset: Line finding algorithms (in collaborations with Schlumberger)
DCT Compression
See also Daniel Haugen´s NOTUR 2009 Poster:

“Strategies for Handling Large Amounts of Data from Storage to GPUs”
(also in collaborations with Tore Fevang, Schlumberger)
Some of the many other applications

• Medicine:
  – ultrasound imaging
  – imaging of vessels before surgery

• Chemistry
  – molecular analysis and simulation
  – CFD

• Physics
  – particle simulations

• Marine technology
• Mathematical methods
• Computer algorithms, benchmarking ...
• ...

www.ntnu.no
Real-time Image enhancement for Porous Rocks (w/ Henrik Hesland)

- In collaboration with Dept. Of Petroleum Engineering
Global vs. Variable vs. Variable regional thresholds
a) <Update 3D> pressed, and volume loaded.
a) <Update 3D> pressed, and volume loaded.

b) Volume rotated.
Simulations of Fluid Flow through Porous Rocks using GPUs

Eirik Ola Aksnes
(supervisor: Anne C. Elster)

In collaboration with:
• Numerical Rocks &
• NTNU Chemistry Dept.

Use Lattice Boltzmann Method
Benchmarks: Fontainebleau

<table>
<thead>
<tr>
<th>Implementation</th>
<th>Average MLUPS</th>
<th>Maximum MLUPS</th>
<th>Total Time</th>
<th>Number Of Iterations</th>
<th>Permeability Obtained</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU 32</td>
<td>1.03</td>
<td>1.04</td>
<td>2152 s</td>
<td>445</td>
<td>1247.80 mD</td>
</tr>
<tr>
<td>GPU 32</td>
<td>58.81</td>
<td>59.15</td>
<td>38.0 s</td>
<td>445</td>
<td>1247.81 mD</td>
</tr>
<tr>
<td>CPU 64</td>
<td>0.94</td>
<td>0.94</td>
<td>2375.4 s</td>
<td>445</td>
<td>1247.80 mD</td>
</tr>
</tbody>
</table>
LBM Global memory usage

Our GPU implementation support lattice sizes up to $368^3$, which fit into the 4 GB mem. of the NVIDIA Quadro FX 5800 card.
Summary of LBM/Porous Simulation:

• To support large lattices and to get high performance: Swap-instead-of-copy-approach.

• The configuration of grids and thread blocks of kernels were properly configured.

• Register and shared memory usage of the kernels were minimized. Structure-of-arrays / coalescing.

• To get matching result from CPU and GPU with single floating-point precision: Round-off errors in the simulation model were reduced.
“Large Seismic Application on GPU”

-- collaboration with StatoilHydro
(to be presented at a later conference)

Owe Johansen
“Parallel Techniques for Estimation and Correction of Aberration on Medical Ultrasound Images”

-- Collaboration w/ NTNU Med. Tech. (NOTUR 2009 Poster)
HPC-Lab supports research and studies of novel GPU and multi-core architectures

- Parallel and Distributed Algorithms
- **Performance Evaluation and Benchmarking**
- Parallelization of Seismic and Image Related Applications on GPUs and Multi-Cores
- Adaptive and Auto-Tuneable Algorithms and Implementations

**Project/Master Thesis Topics**

- GPGPU for HPC
- Compiler Techniques for Parallel Linear Algebra
- Memory Latency Impact GPU-CPU Configurations
- Other projects in parallel and distributed computing, suggest something that interests you!

**Collaborators / Supporters**

ARM, CERN, NVIDIA, StatoilHydro, Schlumberger, GE-Healthcare, and others
“Communication Challenges on Multi-GPU Systems”
(NOTUR 2009 Poster)

Daniele G. Spampinato
“Modelling Overlapping Communication and Computation”

(NOTUR 2009 Poster)

Jan Christian Meyer
PhD Student
Zotac Nvidia ION

Specifications:
Nvidia ION GPU,
Graphics Engine Clock: 450MHz
Shader Click: 1100MHz
16 stream processors
MS DirectX 10
Intel Atom (integrated) up to 533 MHz FSB
DDR2 667/800
Up to 4GB RAM
1 mini PCIExpress
“High Data Volumes and Streaming on Future GPU Systems”

-- Collaboration w/ Dr. Magnus Lie Hetland (IDI)
And Dr. Øystein Thorbjørnsen (Fast/MicroSoft)

On-board GPUs

Rune Hovland
Master of Tech at NTNU June 2008.
Now consultant and Sirius IT, Oslo, Norway
Throughput Computing

- Information Retrieval (IR), including searches, focuses on throughput and large data volumes.
- Goal for IR: process large amounts of data efficiently, something current GPUs are not good at.
- Here are some modelling and suggestions on how to fix this
Application Related Topics

• Search engines
  – Architecture
  – Inverted index (datastructure optimized for term look-ups e.g. looking for frequency of term
    • Document-level inverted index
    • Word-level
    • Block-level)

Compression (application dependent)
  - variable byte encoding
  - Golomb encoding (bit-wise)
  - Pfor Delta (avoiding branches and optimizes pipeline
Search Engine Architecture

Diagram showing the architecture of a search engine:
- **Crawler** retrieves documents from the document collection.
- **Indexer** processes the documents to create term indexes.
- **Query Engine** receives query terms and returns query results.
- **Document Processor** analyzes the documents to extract terms.
- **Frontend** presents the query results to the user.
Memory Gap
(Patterson & Hennesey)
ION vs 9300m vs GTX 280

![Graph showing performance comparison between Malloc, GTX 280, CudaMallocHost, GTX 280, Malloc, 9300m, CudaMallocHost, 9300m, Malloc, ION, and CudaMallocHost, ION. The x-axis represents data size in KB, and the y-axis represents transfer speed in GB/s.](image-url)
Hide datacopy between host & GPU while accessing disk:

Sequential

Parallel

Time

Read from hard-drive
Copy to memory on GPU

Norwegian University of Science and Technology
But, memory mapped files best ...
Data access figures

- Registers
- Shared Memory
- Prefetching to Shared Memory
- Texture Memory

Data sizes (bytes)

<table>
<thead>
<tr>
<th>Data Size</th>
<th>Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Texture Memory</td>
<td>15.6</td>
</tr>
<tr>
<td>Prefetching to Shared Memory</td>
<td>116.1</td>
</tr>
<tr>
<td>Shared Memory</td>
<td>18.6</td>
</tr>
<tr>
<td>Registers</td>
<td>17.9</td>
</tr>
</tbody>
</table>
Expanding Memory hierarchy

• Nvidia Tesla architecture
• Allows GPU to have:
  – Its own dedicated memory
  – Use part of host memory
Any data used by GPU must be copied to device
Any results copied back (or displayed)

Would like: Direct access to host memory!

-> Change memory hierarchy
Alternative: Combine Dedicated & Shared Memory

- Allow GPU to use both dedicated and local memory, but divided into two levels in memory hierarchy
- Would require CPU can “DMA” to host memory used as device memory
- All data GPU will use, stored here
Zero-copy in CUDA

- To be introduced in CUDA 2.2
- Allows user to page-locked memory from GPU
- Only paged-locked memory is limiting since a scarce resource
- User may still need to extra copy ops if there is extensive memory usage on host (CPU)
Fall 2009 Master Projects

Ahmed Aquari
Schlumberger project on GPU

Aleksander Gjermundsen
LBM/solvers - Snow Simulation

Gaojie He
GPU-enhanced parallel games

Øystein Krogh
Snow Simulations - terrain interactions

Holger Ludvigsen
GPU Ray tracing using OptiX

+ 3 visualization students
+ 1-2 || arch/m multicore students

---

Runar Refsnæs (Math)
Gagandeep Singh (Math)
Roald Fernandez (Cybernetics)
Peter Sveistrup (Cybernetics)
Dzienkuje! /Thank you!
(jen-koo-ye)

Contact Info:

elster@idi.ntnu.no

http://www.idi.ntnu.no/~elster/hpc-lab
GPU-related Activities & Events:

EU COST Action IC0805: Open European Network for High Performance Computing on Complex Environments (2009-2014)

ParCo2009

Lyon, France, Sep 1-4, 2009
MS on GPU Computing /EuroGPU2009
HPC-LAB at SC´08

HPC-LAB at SC´08

www.idi.ntnu.no/~elster/hpc-lab
NTNU Gløshaugen
(formerly Norwegian Institute of Technology)
“Dynamic Optimization of MPI Communication”

(NOTUR 2009 Poster)
HPC Group contin.

At SC´07 in Reno, NV, USA
GPU Apps
(from NVIDIA’s web pages)

CUDA applications actively in use today by these researchers and organizations include:
Oil and gas

- **Acceleware**: Kirchoff Time Migration library
- **ffA**: 3D Seismic processing software
- **Headwave**: Prestack data processing
- **Mercury Computer systems**: 3D data visualization
- **SeismicCity**: 3D seismic imaging for prestack depth migration
- **SMT**: Kingdom – Seismic Processing
Computational Chemistry and Molecular Dynamics:

- GROMACS molecular dynamics
- HOOMD molecular dynamics
- NAMD molecular dynamics
- VMD visualization of molecular dynamics
Bio-Informatics and Life Sciences:

- GPU HMMER: CUDA version of HMMER
- LISSOM: Human neocortex modeling
- MUMmerGPU: High-throughput DNA sequencing
Financial Computing and Options Pricing:

- Aqumin: 3D Visualization of market data
- Exegy: Risk Analysis
- Hanweck: options pricing
- SciComp: derivatives pricing
Mathematical Computing

- Jacket CUDA plugin for MATLAB from Accelereyes
- LabVIEW from National Instruments
GeoSciences:

- Tsunami simulation – Tokyo Institute of Technology
- Weather Research and Forecast (WRF) model
- Geographical Information Systems - Manifold
Medical Imaging, CT, MRI:

- AxeRecon CT reconstruction library from Acceleware
- SnapCT tomographic reconstruction software from Digisens
Electrodynamics and Electromagnetics

- CST Microwave Studio
- FDTD solver from Acceleware
Electronic Design Automation

- ADS SPICE simulator from Agilent EESof
- OmegaSim GX SPICE simulator from Nascentric
- Sentaraus TCAD from Synopsys

Blocksize: 8
Why LOTs?

- Attempts to solve blocking problems w/ DCTs
  --> Better objectively & subjectively compression
## Feig-Linzer DCT vs Fast LOT

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Multiples</th>
<th>Adds/ Subtractions</th>
<th>Multiply-adds</th>
<th>FLOPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>FL mult-add DCT</td>
<td>64</td>
<td>256</td>
<td>160</td>
<td>640</td>
</tr>
<tr>
<td>Fast LOT w/ FIL mult-add DCT</td>
<td>288</td>
<td>512</td>
<td>256</td>
<td>1312</td>
</tr>
</tbody>
</table>

But LOT very vectorizable!
Compression of Image Data on Clusters using GPUs and Quad-Core CPUs

Leif Christian Larsen, IDI-NTNU
Anne Cathrine Elster, IDI-NTNU (main supervisor)
Tore Fævang, Schlumberger Ltd (co-supervisor)
leifcl@idi.ntnu.no
elster@idi.ntnu.no
tfevang@slb.com

• In client-server visualization applications using clusters for computations, image compression is used to reduce communication time between cluster nodes and from cluster nodes to the client, thereby improving application responsiveness.

• Transform coding methods are employed for compressing image data. Transform computation may be offloaded to the graphics processing unit (GPU), which is well suited for such computations. Offloading computations to the GPU releases CPU time for other tasks, and in certain cases offers significantly improved performance over using multi-core CPUs.

Discrete Cosine Transform – Wallclock time

Discrete Cosine Transform – CPU time

• GPU transform implementations compute hundreds of transform values simultaneously by using SIMD vector instructions and multiple GPU vertex and pixel processors in parallel.

• The GPU is particularly well suited for transforms involving heavy computations and large amounts of data, since data upload/download-initialization time becomes less significant as arithmetic complexity increases.

Acknowledgement. This project is done in collaboration with Schlumberger Limited, Trondheim, which provides hardware equipment and other resources.
NTNU Collaborations with CERN:

• Two Norwegian students in summer 2003:
  1 Master student from Elster’s NTNU group (Hisdal)
  + 1 from Ousada’s Univ. of Oslo’s Physics group
Hisdal went on to do his MS thesis at CERN
(with CERN staff and Elster)

• Six MS students from NTNU in summer 2004-2005
  Three of these went on to do their MS Theses at
  CERN (with CERN staff and Elster)
Master Students

Dr. Anne C. Elster
Lab Director

Thorvald Natvig
PhD Student

Jan Christian Meyer
PhD Student

Robin Eidissen
(Teaching Assistant)

Rune E. Jensen

Olav Fagerlund

Eirik O. Aksnes

Daniel Haugen

Henrik Hesland

Åsmund Herikstad

Owe Johansen

Rune Hovland

Daniele G. Spampinato

www.ntnu.no