Parallel Computing: What has changed lately?

David B. Kirk
Future Science and Engineering Breakthroughs Hinge on Computing

- Computational Geoscience
- Computational Chemistry
- Computational Medicine
- Computational Modeling
- Computational Physics
- Computational Biology
- Computational Finance
- Image Processing
Faster is not “just Faster”

- **2-3X faster is “just faster”**
  - Do a little more, wait a little less
  - Doesn’t change how you work

- **5-10x faster is “significant”**
  - Worth upgrading
  - Worth re-writing (parts of) the application

- **100x+ faster is “fundamentally different”**
  - Worth considering a new platform
  - Worth re-architecting the application
  - Makes new applications possible
  - Drives “time to discovery” and creates fundamental changes in Science
The GPU is a New Computation Engine

Relative Floating Point Performance

Era of Shaders

Fully Programmable

G80
CPU
Powerful Multi-core Control Processor

- Operating system
- Database
- Productivity
- Temporal compression
- Recursive algorithms

GPU
Powerful Massively Parallel Computation Processor

- Oil and gas seismic
- Financial risk modeling
- Medical Imaging
- Finite element computing
- Genetic pattern match
Data to Design

Acceleware EM Field simulation technology for the GPU

- 3D Finite-Difference and Finite-Element (FDTD)
- Modeling of:
  - Cell phone irradiation
  - MRI Design / Modeling
  - Printed Circuit Boards
  - Radar Cross Section (Military)

Pacemaker with Transmit Antenna

Performance

<table>
<thead>
<tr>
<th></th>
<th>CPU 3.2 GHz Core 2 Duo</th>
<th>1 GPU</th>
<th>2 GPUs</th>
<th>4 GPUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
<td>1X</td>
<td>11X</td>
<td>22X</td>
<td>45X</td>
</tr>
</tbody>
</table>

Supercomputing 2007 © NVIDIA Corporation 2007
Terabyte Data to Drilling Decision

- Visualize Terabytes of data
- Interactive data processing and analysis

HEADWAVE

Prestack for Interpreters
Headwave offers routine access to terabyte+ prestack data while working in a familiar interpretation and modeling environment. Read about our Petrel plugin.

Highlights
- Headwave brings instant access, visualization and interpretation of terabyte+ prestack surveys in Petrel Workflow Tools® 2007
- Download the brochure (2.5MB)
- Our innovative architecture - Headwave pioneers true real time answer and comparison of terabyte datasets using GPUs for acceleration.
Parallel GPUs with Multithreading:
705 GFLOPS /w 3 GPUs

- One host thread is created for each CUDA GPU
- Threads are spawned and attach to their GPU based on their host thread ID
  - First CUDA call binds that thread’s CUDA context to that GPU for life
  - Handling error conditions within child threads is dependent on the thread library and, makes dealing with any CUDA errors somewhat tricky, left as an exercise to the reader.
- Map slices are computed cyclically by the GPUs
- Want to avoid false sharing on the host memory system
  - map slices are usually much bigger than the host memory page size, so this is usually not a problem for this application
- Performance of 3 GPUs is stunning!
- Power: 3 GPU test box consumes 700 watts running flat out
Evolved Machines

- Simulate the brain circuit
- Sensory computing: vision, olfactory
- 130X Speed up
Matlab: Language of Science

15X with MATLAB CPU+GPU


Pseudo-spectral simulation of 2D Isotropic turbulence

http://www.amath.washington.edu/courses/571-winter-2006/matlab/FS_2Dturb.m
Other Links

**Astrophysics**
Astrophysical simulations based on smoothed particle hydrodynamics: Fourier Volume Rendering
Andrew Corrigan and John Wallin: Computational and Data Sciences, George Mason University
http://cds.gmu.edu/~acorriga/pubs/meshless_fvr

**Astrophysics**
Astrophysical N-body simulation: The Chamomile Scheme
Tsuyoshi Hamada and Toshiaki Itakata: Computational Astrophysics Lab, RIKEN
http://progrape.jp/cs/

**Financial Simulation**
Computational Finance: Swaption volatility
Level 3 Finance

**Financial Simulation**
Quantitative Risk Analysis and Algorithmic Trading Systems
Hanweck Associates
http://www.hanweckassoc.com/home.html

**Medical Imaging**
National Library of Medicine Insight Segmentation and Registration Toolkit (ITK)
Won-Ki Jeong: Scientific Computing & Imaging Institute, University of Utah
http://www.itk.org/

**Physical Simulation**
Simulation Open Framework Architecture for real-time simulation with an emphasis on medical simulation
http://www-evasion.imag.fr/%7EFrancois.Faure/Sofa/web/home

**Video Capture**
3D Surface Image Capture and “4D Capture” of Stereo Video Time Sequencing
Dimensional Imaging
http://www.dli3d.com/

**GIS**
Geographic Information System (GIS) and Mapping products
Manifold
http://www.manifold.net/

**Bioscience**
Computational biology string matching: CMATCH
Michael C. Schatz and Cole Trapnell: Center for Bioinformatics & Computational Biology
University of Maryland
http://www.cseb.umd.edu/software/cmatch/

**Gene Sequence Analysis**
Genomic Data Sequence Analysis: SWBoost (Smith-Waterman Boost)
Genboost
http://www.genboost.com/swboost.php
Graphics Programming Model

Graphics Application

Vertex Program

Rasterization

Fragment Program

Display
Streaming GPGPU Programming

OpenGL Program to Add A and B

Start by creating a quad

“Programs” created with raster operation

Vertex Program

Read textures as input to OpenGL shader program

Fragment Program

Write answer to texture memory as a “color”

CPU Reads Texture Memory for Results

All this just to do A + B
Example Fluid Algorithm

CPU

GPGPU

Multiple passes through video memory

Parallel execution on-chip

Single thread out of cache

Data/Computation

Program/Control
New GPU Computing Model

- Dedicated computing mode
- Thread programs use ‘C’
- On-chip shared memory
- General load/store

Thread ID

Thread Program
Written in ‘C’

Parallel Data Cache

- Registers
- Constants
- optional Texture

Global Memory
The Future of Computing is Parallel

- CPU clock rate growth is slowing, future speed growth will be from parallelism
- GeForce-8 Series is a massively parallel computing platform
  - 12,288 concurrent threads, hardware managed
  - 128 x4 Thread Processor cores at 1.35 GHz == 518 GFLOPS peak
- GPU Computing features enable C on Graphics Processing Unit
CUDA Software Development Kit

CUDA Optimized Libraries: math.h, FFT, BLAS, ...

Integrated CPU + GPU C Source Code

NVIDIA C Compiler

NVIDIA Assembly for Computing (PTX)

CPU Host Code

CUDA Driver

Standard C Compiler

Debugger Profiler

GPU

CPU
CUDA: C on the GPU

- A simple, explicit programming language solution
- Extend only where necessary

```c
__global__ void KernelFunc(...);
__shared__ int SharedVar;

KernelFunc<<< 500, 128 >>>(...);
```

- Explicit GPU memory allocation
  - cudaMalloc(), cudaMemcpy(), cudaMemcpy2D(), ...

- Memory copy from host to device, etc.
C-Code Example to Add Arrays

CPU C program

```c
void add_matrix_cpu
(float *a, float *b, float *c, int N)
{
    int i, j, index;
    for (i=0;i<N;i++) {
        for (j=0;j<N;j++) {
            index =i+j*N;
            c[index]=a[index]+b[index];
        }
    }
}
```

```c
void main()
{
    ..... 
    add_matrix(a,b,c,N);
}
```

CUDA C program

```c
__global__ void add_matrix_gpu
(float *a, float *b, float *c, int N)
{
    int i=blockIdx.x*blockDim.x+threadIdx.x;
    int j=blockIdx.y*blockDim.y+threadIdx.y;
    int index =i+j*N;
    if( i <N && j <N) c[index]=a[index]+b[index];
}
```

```c
void main()
{
    dim3 dimBlock (blocksize,blocksize);
    dim3 dimGrid (N/dimBlock.x,N/dimBlock.y);
    add_matrix_gpu<<<dimGrid,dimBlock>>>(a,b,c,N);
}
```
Compiling CUDA

C/C++ CUDA Application

NVCC

PTX Code

CPU Code

Virtual

PTX to Target Translator

GPU

Target code

Target

GPU
CUDA Stable Fluids Demo

CUDA port of:
Come visit the class!

UIUC ECE498AL – Programming Massively Parallel Processors
(http://courses.ece.uiuc.edu/ece498/al/)

- David Kirk (NVIDIA) and Wen-mei Hwu (UIUC) co-instructors
- CUDA programming, GPU computing, lab exercises, and projects
- Lecture slides and voice recordings
Implications and Opportunities

Massively parallel computing allows
- Drastic reduction in “time to discovery”
- New, 3rd paradigm for research: computational experimentation
- The “democratization of supercomputing”
  - $2,000/Teraflop SPFP in personal computers today
  - $5,000,000/Petaflops DPFP in clusters in two years
  - HW cost will no longer be the main barrier for big science
- This is once-in-a-career opportunity for many!

Call to Action
- Research in Parallel Programming models and Parallel Architecture
- Teach massively parallel programming to CS/ECE students, scientists and other engineers.
Questions?