Introduction to CUDA
GPU Performance History

- GPUs are massively multithreaded many-core chips
  - Hundreds of cores, thousands of concurrent threads
  - Huge economies of scale
  - Still on aggressive performance growth
CUDA

CUDA is industry-standard C
- Write a program for one thread
- Instantiate it on many parallel threads
- Familiar programming model and language

CUDA is a scalable parallel programming model
- Program runs on any number of processors without recompiling
GPU Sizes Require CUDA Scalability

- 32 SP Cores
- 128 SP Cores
- 240 SP Cores

© NVIDIA Corporation 2008
CUDA runs on NVIDIA GPUs...
Over 100 Million CUDA GPUs Deployed

GeForce®
Entertainment

Quadro®
Design & Creation

Tesla™
High-Performance Computing

© NVIDIA Corporation 2008
Pervasive CUDA Parallel Computing

- CUDA brings data-parallel computing to the masses
  - Over 100 M CUDA-capable GPUs deployed since Nov 2006

- Wide developer acceptance
  - Download CUDA from [www.nvidia.com/cuda](http://www.nvidia.com/cuda)
  - Over 50K CUDA developer downloads
  - A GPU “developer kit” costs ~$100 for several hundreds GFLOPS

- Data-parallel supercomputers are everywhere!
  - CUDA makes this power readily accessible
  - Enables rapid innovations in data-parallel computing

- Parallel computing rides the commodity technology wave
CUDA Zone: www.nvidia.com/cuda
Introducing Tesla T10P Processor

1.4 billion transistors
1 Teraflop of processing power
240 SP processing cores
30 DP processing cores with
IEEE-754 double precision

…NVIDIA’s 2nd Generation CUDA Processor

© NVIDIA Corporation 2008
CUDA Computing with Tesla T10

- 240 SP processors at 1.45 GHz: 1 TFLOPS peak
- 30 DP processors at 1.44Ghz: 86 GFLOPS peak
- 128 threads per processor: 30,720 threads total
<table>
<thead>
<tr>
<th></th>
<th>NVIDIA GPU</th>
<th>SSE2</th>
<th>Cell SPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>IEEE 754</td>
<td>IEEE 754</td>
<td>IEEE 754</td>
</tr>
<tr>
<td>Rounding modes for FADD and FMUL</td>
<td>All 4 IEEE, round to nearest, zero, inf, -inf</td>
<td>All 4 IEEE, round to nearest, zero, inf, -inf</td>
<td>Round to zero/truncate only</td>
</tr>
<tr>
<td>Denormal handling</td>
<td>Full speed</td>
<td>Supported, costs 1000’s of cycles</td>
<td>Flush to zero</td>
</tr>
<tr>
<td>NaN support</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Overflow and Infinity support</td>
<td>Yes</td>
<td>Yes</td>
<td>No infinity, clamps to max norm</td>
</tr>
<tr>
<td>Flags</td>
<td>No</td>
<td>Yes</td>
<td>Some</td>
</tr>
<tr>
<td>FMA</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Square root</td>
<td>Software with low-latency FMA-based convergence</td>
<td>Hardware</td>
<td>Software only</td>
</tr>
<tr>
<td>Division</td>
<td>Software with low-latency FMA-based convergence</td>
<td>Hardware</td>
<td>Software only</td>
</tr>
<tr>
<td>Reciprocal estimate accuracy</td>
<td>24 bit</td>
<td>12 bit</td>
<td>12 bit</td>
</tr>
<tr>
<td>Reciprocal sqrt estimate accuracy</td>
<td>23 bit</td>
<td>12 bit</td>
<td>12 bit</td>
</tr>
<tr>
<td>log2(x) and 2^x estimates accuracy</td>
<td>23 bit</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
Applications
Folding@home Performance Comparison

<table>
<thead>
<tr>
<th>OS Type</th>
<th>Current TFLOPS*</th>
<th>Active CPUs</th>
<th>Total CPUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Windows</td>
<td>208</td>
<td>218169</td>
<td>2172051</td>
</tr>
<tr>
<td>Mac OS X/PowerPC</td>
<td>7</td>
<td>8563</td>
<td>120066</td>
</tr>
<tr>
<td>Mac OS X/Intel</td>
<td>23</td>
<td>7332</td>
<td>62880</td>
</tr>
<tr>
<td>Linux</td>
<td>60</td>
<td>35036</td>
<td>330960</td>
</tr>
<tr>
<td>ATI GPU</td>
<td>447</td>
<td>4065</td>
<td>9396</td>
</tr>
<tr>
<td>NVIDIA GPU</td>
<td><strong>1450</strong></td>
<td>13185</td>
<td>27428</td>
</tr>
<tr>
<td>PLAYSTATION®3</td>
<td>1078</td>
<td>38229</td>
<td>599812</td>
</tr>
<tr>
<td>Total</td>
<td>3273</td>
<td>324579</td>
<td>3322593</td>
</tr>
</tbody>
</table>

nano seconds of simulation per day

800
700
600
500
400
300
200
100
0

CPU  | PS3  | Radeon HD 3870 | Radeon HD 4850 | Tesla 8-series | Tesla 10-Series |
--- | --- | --------------- | --------------- | --------------- | --------------- |
4   | 100 | 170             | 377             | 423             | 740             |

F@H kernel based on GROMACS code

© NVIDIA Corporation 2008
Lattice Boltzmann

1000 iterations on a 256x128x128 domain
Cluster with 8 GPUs: 7.5 sec
Blue Gene/L 512 nodes: 21 sec

10000 iterations on irregular 1057x692x1446 domain with 4M fluid nodes

<table>
<thead>
<tr>
<th></th>
<th>Time (s)</th>
<th>MLUPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 C870</td>
<td>760</td>
<td>53</td>
</tr>
<tr>
<td>2 C1060</td>
<td>159</td>
<td>252</td>
</tr>
<tr>
<td>8 C1060</td>
<td>42</td>
<td>955</td>
</tr>
</tbody>
</table>

Blood flow pattern in a human coronary artery, Bernaschi et al.
Desktop GPU Supercomputer Beats Cluster

CalcUA
256 Nodes (512 cores)

FASTRA
8 GPUs in a Desktop

CUDA accelerated Linpack

Standard HPL code, with library that intercepts DGEMM and DTRSM calls and executes them simultaneously on the GPUs and CPU cores. Library is implemented with CUBLAS

Cluster with 8 nodes:
- Each node has 2 Intel Xeon E5462 (2.8Ghz), 16GB of memory and 2 Tesla GPUs (1.44Ghz clock).
- The nodes are connected with SDR Infiniband.

<table>
<thead>
<tr>
<th>T/V</th>
<th>N</th>
<th>NB</th>
<th>P</th>
<th>Q</th>
<th>Time</th>
<th>Gflops</th>
</tr>
</thead>
<tbody>
<tr>
<td>WR11R2L2</td>
<td>118144</td>
<td>960</td>
<td>4</td>
<td>4</td>
<td>874.26</td>
<td>1.258e+03</td>
</tr>
</tbody>
</table>

\[ \|\|Ax-b\|\|_\infty/(\epsilon(\|A\|_\infty*\|x\|_\infty +\|b\|_\infty)*N)\| = 0.0031157 \ldots \text{PASSED} \]
Accelerating MATLAB®

Pseudo-spectral simulation of 2D Isotropic turbulence

Use MEX files to call CUDA from MATLAB, 17x speed-up

1024x1024 mesh, 400 RK4 steps, Windows XP, Core2 Duo 2.4Ghz vs GeForce 8800GTX

Applications in several fields

Interactive visualization of volumetric white matter connectivity

Ionic placement for molecular dynamics simulation on GPU

Transcoding HD video stream to H.264

Simulation in Matlab using .mex file CUDA function

Astrophysics N-body simulation

Financial simulation of LIBOR model with swaptions

GLAME@lab: An M-script API for linear Algebra operations on GPU

Ultrasound medical imaging for cancer diagnostics

Highly optimized object oriented molecular dynamics

Cmatch exact string matching to find similar proteins and gene sequences