2009: The GPU Computing Tipping Point

Jen-Hsun Huang, CEO
Someday, our graphics chips will have 1 TeraFLOPS of computing power, will be used for playing games to discovering cures for cancer to streaming video to millions of people connected on the Internet.

…..Right!
NVIDIA Businesses

Consumer Graphics

Professional Workstation Graphics

High Performance Computing

Mobile & Embedded Computing
Revolutionizing Computer Graphics

3D Accelerations
Fixed Pipelines

Programmable Shading
Pipelines of Processors

Computational Visualization
Massive Array of Processors
Nehalem CPU:
- 3 Ghz
- 4 cores
- 4 way SIMD
- 2 FLOPS/cycle
- \(= 96 \text{ GFLOPS}\)

Programmable Shading

- 2,300,000 pixels/frame
- \(\times 3\) depth complexity
- \(\times 100\) shader inst./component
- \(\times 1.5\) FLOPS/inst.
- \(\times 4\) components/pixel
- \(\times 60\) frames/sec
- \(\times 2\) stereoscopic

\(= 500\) shader GFLOPS (approx. 10% of graphics ops)
Computer Graphics – A Study in Parallelism

CPU

CPU + 1 year
“The Brick Wall”

“Power Wall + Memory Wall + ILP Wall = Brick Wall”

The Landscape of Parallel Computing Research: A View from Berkeley
David Patterson et al.

Source: Hennessy & Patterson, CAAQA, 4th Edition
Co-Processing

The Right Processor for the Right Tasks
Many processors – eventually thousands
Latency tolerant - execute 1000’s of threads
General load/store
On-chip shared-memory
CUDA programs scales across any size GPU

240 SP Cores
Computer Graphics – A Study in Parallelism

Where \( P = 99\% \), \( S = 100 \times \)

\[
S = \frac{6.4 \text{ TOPS}}{0.5 \times 96 \text{ GFLOPS}} = 100
\]
Computer Graphics – A Study in Parallelism

CPU

CPU + 1 year

CPU + GPU

\[ \frac{1}{(1 - P) + \frac{P}{S}} \]

Where \( P = 99\% \), \( S = 100x \)

Overlapped Execution
### Computer Graphics – A Study in Parallelism

<table>
<thead>
<tr>
<th>System Configuration</th>
<th>Fallout 3 1920x1200; 4x AA</th>
<th>Far Cry 2 1920x1200; 4x AA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Core i5</strong> + GeForce GTX 275</td>
<td>69.1 FPS</td>
<td>49.6 FPS</td>
</tr>
<tr>
<td><strong>Core i7</strong> + GeForce GTX 275</td>
<td>69.8 FPS</td>
<td>50.7 FPS</td>
</tr>
</tbody>
</table>
The Next Big Thing – Physics

Simulate Amazing Worlds

3D Accelerations
Fixed Pipelines

Programmable Shading
Pipelines of Processors

Computational Visualization
Massive Array of Processors

GOPS

10,000

1,000

100

0.2

1,000
ILM – Siggraph 2009
Directable, high resolution simulation of fire on the GPU

“the GPU gave us unbelievable speed-ups over the typical CPU. We built a GPU farm that could handle these massive simulations. **What would take a day to run on a CPU, we were able to simulate in 40 minutes.** The graphics processor is ideal for handling millions of instructions in split-seconds.“

Tim Alexander and Robert Weaver, ILM
Post July 1, 2009
Co-Processing

Ideal for Ray Tracing

- User Input
- JIT Compile
- Disk I/O
- Animation
- Acceleration Build
- Traversal
- Intersection
- Shading
- Camera Model
- Tone Mapping
Co-Processing
Ideal for Physics Processing

- Network Synchronization
- Ragdoll physics
- Ray Casting
- AI
- Collision Detection
- Deformation
- Smoothed Particle Hydrodynamics (SPH)
- Particles
Co-Processing

Ideal for Molecular Dynamics

- Steering
- Time Stepping
- Disk I/O

- Pairlist calculation
- Pairlist update
- Non-bonded force calculation

- Fluorescence microphotolysis
- Direct Coulomb Summation
- Cutoff potential summation
# Huge Speed-Ups Across Many Fields

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Field</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-Electron Repulsion Integral</td>
<td>Quantum Chemistry</td>
<td>130X</td>
</tr>
<tr>
<td>Lattice Boltzmann</td>
<td>CFD</td>
<td>123X</td>
</tr>
<tr>
<td>Euler Solver</td>
<td>CFD</td>
<td>16X</td>
</tr>
<tr>
<td>GROMACS</td>
<td>Molecular Dynamics</td>
<td>137X</td>
</tr>
<tr>
<td>Lattice QCD</td>
<td>Physics</td>
<td>30X</td>
</tr>
<tr>
<td>Multifrontal Solver</td>
<td>FEA</td>
<td>20X</td>
</tr>
<tr>
<td>nbody</td>
<td>Astrophysics</td>
<td>100X</td>
</tr>
<tr>
<td>Simultaneous Iterative Reconstruction Technique</td>
<td>Computed Tomography</td>
<td>32X</td>
</tr>
</tbody>
</table>
Theoretical Speed-Up

- 1core + gpu
- 4core
- 1core
## Oil & Gas: Seismic Processing

- **Equal Performance:** 1
- **31x Less Space:** 2000 CPU Servers
- **20x Lower Cost:** ~$8M
- **27x Lower Power:** 1200 kWatts

- **32 Tesla S1070s:** ~$400K
- **45 kWatts**
Co-Processing
The Right Processor for the Right Tasks

2015 Projection

CPU-Alone $1.2^6$ 3X

CPU+GPU $50 \times 1.5^6$ 570X
Universal Translator
Augmented Reality
GPU Computing has reached “the tipping point”