GPU Computing
Opportunities for
HPC in Industry

Stan Posey
HPC Applications and Industry Development
NVIDIA, Santa Clara, CA, USA
Why GPU Computing?  High Performance

Peak Performance
Gflops

Tesla T10
Tesla G80
Harpertown 3 GHz
Nehalem 3 GHz

NVIDIA GPU

Peak Memory Bandwidth
GB/sec

Tesla T10
Tesla G80
Harpertown 3 GHz
Nehalem 3 GHz

NVIDIA GPU
Objective of GPU Computing: HPC Efficiency

Operations Should be Processed on the Device Best Suited for HPC Efficiency

[Efficiency in Performance, Parallel Scaling, Power Consumption, Space, Human Effort, etc.]

CPU + GPU Co-Processing
Heterogeneous Computing
## GPU Computing for Desktops and Servers

**Tesla S1070 1U System**

- **GPUs:** 4 Tesla GPUs
- **Single Precision Performance:** 4.14 Teraflops
- **Double Precision Performance:** 346 Gigaflops
- **Memory:** 16 GB (4 GB / GPU)

**Tesla C1060 Computing Board**

- **GPUs:** 1 Tesla GPU
- **Single Precision Performance:** 933 Gigaflops
- **Double Precision Performance:** 78 Gigaflops
- **Memory:** 4 GB
NVIDIA Accomplishments in Recent 5 Years

2004
- Began strategic investments in GPU as an HPC microprocessor

2006
- G80 first GPU with built-in compute features, 128 core multi-threaded, scalable architecture
- CUDA SDK Beta

2007
- Tesla 8-series introduction based on G80, 128 cores – gen-1
- CUDA SDK 1.0, 1.1

2008
- Tesla 10-series introduction based on GT 200, 240 cores – gen-2
- CUDA SDK 2.0

2009 (2010 release)
- Tesla third generation, code named “Fermi” 512 cores – gen-3
- CUDA SDK 3.0
NVIDIA GPU Progression for HPC Market

Research Adoption

- Tesla gen-1; 2007
  - GPU achievements demonstrated at conferences; in journals

Research Expansion | Industry Adoption

- Tesla gen-2; 2008
  - Broad research expansion to all disciplines of science
  - Moderate ISV development and product introduction
  - In-house industry software: seismic; finance; aerospace
  - Gov’t funding of GPU developments for target industries

Industry Expansion

- Tesla gen-3; 2010 (Fermi)
  - Broad ISV development move ongoing in preparation for Fermi
  - Benefits to pharmaceuticals, manufacturing, energy, etc.
Abundance of HPC Research Achievements

Example: Gene Sequencing and Assembly in Bioinformatics
## Select Industry HPC Customers and ISVs

<table>
<thead>
<tr>
<th>Life Sciences &amp; Medical Equipment</th>
<th>Oil &amp; Gas</th>
<th>EDA</th>
<th>Finance</th>
<th>CAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Planck</td>
<td>GE Healthcare</td>
<td>Hess</td>
<td>Synopsys</td>
<td>AccelEyes</td>
</tr>
<tr>
<td>FDA</td>
<td>Siemens</td>
<td>TOTAL</td>
<td>Nascentric</td>
<td>MathWorks</td>
</tr>
<tr>
<td>Robarts Research</td>
<td>Techniscan</td>
<td>CGG/Veritas</td>
<td>Gauda</td>
<td>Wolfram</td>
</tr>
<tr>
<td>Medtronic</td>
<td>Boston Scientific</td>
<td>Chevron</td>
<td>CST</td>
<td>National Instruments</td>
</tr>
<tr>
<td>AGC</td>
<td>Eli Lilly</td>
<td>Headwave</td>
<td>Agilent</td>
<td>Ansys</td>
</tr>
<tr>
<td>Evolved machines</td>
<td>Silicon Informatics</td>
<td>Acceleware</td>
<td>Symcor</td>
<td>Access</td>
</tr>
<tr>
<td>Smith-Waterman DNA sequencing</td>
<td>Stockholm Research</td>
<td>Seismic City</td>
<td>Level 3</td>
<td>Analytics</td>
</tr>
<tr>
<td>AutoDock</td>
<td>Harvard</td>
<td>P-Wave</td>
<td>SciComp</td>
<td>Tech-x</td>
</tr>
<tr>
<td>NAMD/VMD</td>
<td>Delaware</td>
<td>Seismic</td>
<td>Hanweck</td>
<td>RIKEN</td>
</tr>
<tr>
<td>Folding@Home</td>
<td>Pittsburg</td>
<td>Imaging</td>
<td>Quant</td>
<td>SOFA</td>
</tr>
<tr>
<td>Howard Hughes Medical</td>
<td>ETH Zurich</td>
<td>Mercury</td>
<td>Catalyst</td>
<td>Renault</td>
</tr>
<tr>
<td>CRIBI Genomics</td>
<td>Institute Atomic Physics</td>
<td>ffA</td>
<td>RogueWave</td>
<td>Boeing</td>
</tr>
</tbody>
</table>

**CUDA**
Tesla Results for Risk Analysis in Finance

Exegy Ticker Plant for Accelerated Risk Evaluation

- Valuation at Risk (VaR)
- 1024 equities, $2^{20}$ simulations
- 180x Faster with CUDA
- 45 mins → 12 secs

Tesla Results for Seismic in Oil & Gas

**CPU Rack**: 64 CPU nodes (DS = 512 CPU cores)
**GPU Rack**: 8 Tesla S1070s (32 GPUs) +16 CPU nodes

Source: TOTAL, ISC 2009

**8 Intel Xeon 5460 cores @ 3.16 GHz**
**4 Tesla GPUs (1 Tesla S1070)**

Source: Chevron, SEG 2008
## Tesla Results for Seismic in Oil & Gas

<table>
<thead>
<tr>
<th>Successful Customers</th>
<th>GPU vs CPU Improvements</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Performance / Watt</td>
</tr>
<tr>
<td></td>
<td>18x - 27x</td>
</tr>
<tr>
<td></td>
<td>12x - 17x</td>
</tr>
<tr>
<td></td>
<td>Performance / Space</td>
</tr>
<tr>
<td></td>
<td>20x - 31x</td>
</tr>
<tr>
<td></td>
<td>15x - 20x</td>
</tr>
<tr>
<td></td>
<td>Performance / Cost</td>
</tr>
<tr>
<td></td>
<td>15x - 20x</td>
</tr>
<tr>
<td></td>
<td>10x - 12x</td>
</tr>
</tbody>
</table>

### Oil & Gas ISVs

- Hess
- Total
- Chevron
- Petrobras
- Acceleware
- FSS
- Gessmage
- Geostreamer
- Headwave
- Seismicity
- Stone Ridge Technology
# HPC Focus Areas for Research and Industry

## NVIDIA HPC Focus Area

<table>
<thead>
<tr>
<th>Research HPC</th>
<th>Sample Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPM [Computational Physics and Mathematics]</td>
<td>ALEGRA; VPIC; Neutronics</td>
</tr>
<tr>
<td>CCB [Computational Biology and Chemistry]</td>
<td>GAMESS; AMBER; BLAST</td>
</tr>
<tr>
<td>CWO [Climate Weather Ocean modeling]</td>
<td>WRF; HYCOM; MOM; HOMME</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Industry HPC</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>EEP [Energy Exploration and Production]</td>
<td>GeoDepth; Eclipse; ProMAX</td>
</tr>
<tr>
<td>CAE [Computer Aided Engineering]</td>
<td>FLUENT; STAR-CD; Abaqus</td>
</tr>
<tr>
<td>EDA [Electronic Design Automation]</td>
<td>Proteus; OPCpro; Hercules</td>
</tr>
<tr>
<td>CFE [Computational Finance and Economics]</td>
<td>Monte Carlo Simulations</td>
</tr>
<tr>
<td>DIR [Digital Imaging and Rendering]</td>
<td>RenderMan, etc</td>
</tr>
</tbody>
</table>
Drivers for Research HPC Objectives

National Interests of Global Advancement and Improved Quality of Life

Competition
- Economic Growth
- Technology Export
- Education Development

Global Advancement
- Defense
- Energy
- Technology Transfer to Industry

Quality of Life Improvements
- Health Sciences
- Environment / Climate
- Agriculture Science

Application Software Objectives: Efficient Compilers; Development Tools
Drivers for Industry HPC Objectives

Product Development for Regulated and Competitive Consumer Markets

Competition
- Reduced Lead Times
- Reduced Cost
- Improved Quality

Consumer Appeal
- Price-Performance
- Latest Features
- Comfort & Style

Government Regulations
- Product Safety
- Energy Efficient
- Environment Impact

Application Software Objectives: Efficient ISV and COTS software
Example: Ford Motor and CAE ISV Software

<table>
<thead>
<tr>
<th>Software Provider</th>
<th>Application Software</th>
<th>Sample Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIMULIA</td>
<td>Abaqus/Standard</td>
<td>Thermal Stress of Powertrain</td>
</tr>
<tr>
<td>MSC.Software</td>
<td>MSC.Nastran</td>
<td>Body NVH</td>
</tr>
<tr>
<td>Altair Engineering</td>
<td>RADIOSS</td>
<td>Crashworthiness &amp; Safety</td>
</tr>
<tr>
<td>LSTC</td>
<td>LS-DYNA</td>
<td></td>
</tr>
<tr>
<td>ANSYS Inc.</td>
<td>ANSYS FLUENT</td>
<td>CFD – External Aerodynamics</td>
</tr>
<tr>
<td>CD adapco Group</td>
<td>STAR-CD</td>
<td>CFD – Engine In-cylinder flow</td>
</tr>
</tbody>
</table>

NOTE: This list of applications is not meant to imply use of NVIDIA GPU computing.
GPU Computing Opportunities in Automotive

Improved Crashworthiness Simulation Through Optimization

- Manufacturing process and tolerances ignored
- Ideally Manufactured Model
- Crashworthiness Simulation
- Metal Forming Simulations
- Realistic Model of Formed Metal and Assembly
- Crashworthiness Simulation

Manufacturing process and tolerances modeled
GPU Computing Opportunities in Automotive

Improved Crashworthiness Simulation Through Optimization

Crashworthiness Example: combination of parameterized models of 10,000’s of jobs, and high-fidelity models of 100’s of jobs

Metal Forming Simulations

Unlikely performance

Most likely performance

Deterministic:

Non-Deterministic:
GPU Computing Opportunities in Automotive

Improved Crashworthiness Simulation Through Optimization

Crashworthiness Example: combination of parameterized models of 10,000's of jobs, and high fidelity models of 100's of jobs

Metal Forming Simulations

Unlikely performance

Most likely performance

Deterministic (Capability):
Large job across CPU+GPU cluster

Non-Deterministic (Capacity):
Small jobs, one per CPU+GPU

BIW of Formed Metal and Assembly

Full System
NVIDIA GPU Progression for HPC Market

- **Research Adoption**
  - Tesla gen-1; 2007
  - GPU achievements demonstrated at conferences; in journals

- **Research Expansion | Industry Adoption**
  - Tesla gen-2; 2008
  - Broad research expansion to all disciplines of science
  - Moderate ISV development and product introduction
  - In-house industry software: seismic; finance; aerospace
  - Gov’t funding of GPU developments for target industries

- **Industry Expansion**
  - Tesla gen-3; 2010 (Fermi)
  - Broad ISV development move ongoing in preparation for Fermi
  - Benefits to pharmaceuticals, manufacturing, energy, etc.
Introduction of the ‘Fermi’ Architecture

“The World’s 1st Complete GPU Computing Architecture”

- More than 2x the cores (512 total)
- 8x more peak double precision performance; > 30% single precision
- ECC for all levels of memory
- L1 (64K) and unified L2 (768K) cache memory hierarchy – 1st ever for GPU
- ~2x more memory bandwidth (Use of GDDR5 memory vs. GDDR3)
- Roadmap for 1 TB of main memory
- Concurrent Kernels, C++, FTN
Increasing GPU Performance Gap With CPUs

**Tesla 10-Series (Today)**

- Peak Memory Bandwidth
  - GB/sec
  - NVIDIA GPU
  - Tesla T10
  - Tesla G80
  - Nehalem 3 GHz
  - Harpertown 3 GHz

**Fermi (Q2/2010)**

- Peak Memory Bandwidth
  - GB/sec
  - NVIDIA GPU
  - Fermi
  - Tesla T10
  - Tesla G80
  - Harpertown 3 GHz
  - Nehalem 3 GHz
“Oak Ridge National Lab (ORNL) has already announced it will be using Fermi technology in an upcoming super that is ‘expected to be 10-times more powerful than today’s fastest supercomputer.’

Since ORNL’s Jaguar supercomputer, for all intents and purposes, holds that title, and is in the process of being upgraded to 2.3 Petaflops ...
Recent NVIDIA Developments in HPC

- Tesla Fermi Announcement and Information

- NVIDIA Collaboration with Microsoft in HPC

- NVIDIA Support Announced for OpenCL

- ORNL and NVIDIA to Collaborate on 20 PFLOPS
  - [http://www.hpcwire.com/offthewire/62919517.html](http://www.hpcwire.com/offthewire/62919517.html)
Thank you, Questions?

Stan Posey
HPC Applications and Industry Development
NVIDIA, Santa Clara, CA, USA