Agenda

• A Little about NVIDIA
• Update on GPUs in Computational Chemistry
• A Couple of Examples
• Upcoming Events
VISUALIZATION

QUADRO™

PARALLEL COMPUTING

TESLA™

PERSONAL COMPUTING

GeForce™, TEGRA™
NVIDIA and HPC Evolution of GPUs

- Public, based in Santa Clara, CA | ~$4B revenue | ~5,500 employees
- Founded in 1999 with primary business in semiconductor industry
  - Products for graphics in workstations, notebooks, mobile devices, etc.
  - Began R&D of GPUs for HPC in 2004, released first Tesla and CUDA in 2007
- Development of GPUs as a co-processing accelerator for x86 CPUs

HPC Evolution of GPUs

- **2004**: Began strategic investments in GPU as HPC co-processor
- **2006**: G80 first GPU with built-in compute features, 128 cores; CUDA SDK Beta
- **2007**: Tesla 8-series based on G80, 128 cores – CUDA 1.0, 1.1
- **2008**: Tesla 10-series based on GT 200, 240 cores – CUDA 2.0, 2.3
- **2009**: Tesla 20-series, code named “Fermi” up to 512 cores – CUDA SDK 3.0, 3.2
# Tesla Data Center & Workstation GPU Solutions

## Tesla M-series GPUs

<table>
<thead>
<tr>
<th></th>
<th>M2090</th>
<th>M2070</th>
<th>M2050</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cores</td>
<td>512</td>
<td>448</td>
<td>448</td>
</tr>
<tr>
<td>Memory</td>
<td>6 GB</td>
<td>6 GB</td>
<td>3 GB</td>
</tr>
<tr>
<td>Memory bandwidth (ECC off)</td>
<td>177.6 GB/s</td>
<td>150 GB/s</td>
<td>148.8 GB/s</td>
</tr>
<tr>
<td>Peak Perf Gflops</td>
<td>Single Precision</td>
<td>1331</td>
<td>1030</td>
</tr>
<tr>
<td></td>
<td>Double Precision</td>
<td>665</td>
<td>515</td>
</tr>
</tbody>
</table>

## Tesla C-series GPUs

<table>
<thead>
<tr>
<th></th>
<th>C2070</th>
<th>C2050</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cores</td>
<td>448</td>
<td>448</td>
</tr>
<tr>
<td>Memory</td>
<td>6 GB</td>
<td>3 GB</td>
</tr>
<tr>
<td>Memory bandwidth (ECC off)</td>
<td>148.8 GB/s</td>
<td>148.8 GB/s</td>
</tr>
<tr>
<td>Peak Perf Gflops</td>
<td>Single Precision</td>
<td>1030</td>
</tr>
<tr>
<td></td>
<td>Double Precision</td>
<td>515</td>
</tr>
</tbody>
</table>
GPUs are Disruptive

**Speed**
Days to minutes
Minutes to seconds

**Throughput**
Simulate more scenarios

**Insight**
Real time analysis
Add GPUs - Accelerate Computing

Speedup
Minimum Port, Big Speed-up

Application Code

- Only Critical Functions
  Parallelize using CUDA Programming Model

- Rest of Sequential CPU Code
We not only created the world's fastest computer, but also implemented a heterogeneous computing architecture incorporating CPU and GPU, this is a new innovation.

Premier Wen Jiabao
More Performance, Less Power

Performance per Megawatts Power

Fastest Top500 Systems

#2, Tienhe 1A
7168 Tesla GPUs
2.5 PFLOPS

#4, Nebulae
4650 Tesla GPUs
1.2 PFLOPS

#5, Tsubame 2.0
4224 Tesla GPUs
1.194 PFLOPS
Strategic Focus on Applications

- Senior-level relationship and market managers
- Dedicated technical resources
- More than 150 people devoted to libraries, tools, application porting and market development
- Worldwide focus
Reaching a Broad Range of Markets

- Scientific computing
- Creative pro
- Education / research
Why Computational Chemistry?
Usage of TeraGrid National Supercomputing Grid

2008 TeraGrid Usage By Discipline

- Molecular Biosciences: 29%
- Chemistry: 13%
- Materials Research: 6%
- Physics: 21%
- Astronomical Sciences: 12%
- Chemical, Thermal Systems: 4%
- Advanced Scientific Computing: 5%
- Earth Sciences: 5%
- All 15 Others: 3%

Half of the cycles
TESLA
Supercomputing

Molecular Dynamics/Mechanics
# Leading MD Applications

<table>
<thead>
<tr>
<th>Application</th>
<th>Features Supported</th>
<th>GPU Perf</th>
<th>Release Status</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMBER</td>
<td>PMEMD: Explicit &amp; Implicit Solvent</td>
<td>8X</td>
<td>V11 Released</td>
<td>Single and multi-GPUs. Expect 2x more performance in V11 patch release (shortly)</td>
</tr>
<tr>
<td>GROMACS</td>
<td>Implicit (5x), Explicit (2x) Solvent</td>
<td>2x-5x</td>
<td>Single GPU released, Version 4.5.4</td>
<td>Next release: 2H2011 Better Explicit, MPI</td>
</tr>
<tr>
<td>LAMMPS</td>
<td>Lennard-Jones, Gay-Berne</td>
<td>3.4-24x</td>
<td>Released</td>
<td>Single and multi-GPU.</td>
</tr>
<tr>
<td>NAMD</td>
<td>Non-bond force calculation</td>
<td>2x-7x</td>
<td>Released, v2.8</td>
<td>Single and multi-GPU.</td>
</tr>
</tbody>
</table>

GPU Perf compared against Multi-core x86 CPU socket. GPU Perf benchmarked on GPU supported features and may be a kernel to kernel perf comparison.
### Additional MD/MM Applications Ramping

<table>
<thead>
<tr>
<th>Application</th>
<th>Features Supported</th>
<th>GPU Perf</th>
<th>Release Status</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abalone</td>
<td>TBD, “Simulations”</td>
<td>4-29X</td>
<td>Released</td>
<td>Single GPU. Agile Molecule, Inc.</td>
</tr>
<tr>
<td></td>
<td>(on 1060 GPU)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACEMD</td>
<td>Written for use on GPUs</td>
<td>“µ-sec long trajectories on workstation”</td>
<td>Released</td>
<td>Production bio-molecular dynamics (MD) software specially optimized to run on single and multi-GPUs</td>
</tr>
<tr>
<td>DL_POLY</td>
<td>Two-body Forces, Link-cell Pairs, Ewald SPME forces, Shake VV</td>
<td>4x</td>
<td>V 4.0 Source only Results Published</td>
<td>Next release: 2H2011 Multi-GPU, multi-node supported</td>
</tr>
<tr>
<td>HOOMD-Blue</td>
<td>Written for use on GPUs</td>
<td>2X</td>
<td>Released, Version 0.9.2</td>
<td>Single and multi-GPU.</td>
</tr>
<tr>
<td></td>
<td>(32 CPU cores vs. 2 10XX GPUs)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

GPU Perf compared against Multi-core x86 CPU socket. GPU Perf benchmarked on GPU supported features and may be a kernel to kernel perf comparison.
<table>
<thead>
<tr>
<th>Related Applications</th>
<th>Features Supported</th>
<th>GPU Perf</th>
<th>Release Status</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core Hopping</td>
<td>GPU accelerated application</td>
<td>Up to 5000X</td>
<td>Released, Suite 2011</td>
<td>Single and multi-GPUs. Schrodinger, Inc.</td>
</tr>
<tr>
<td>FastROCS</td>
<td>Real-time shape similarity searching/comparison</td>
<td>800-3000X</td>
<td>Released</td>
<td>Single and multi-GPUs. Open Eyes Scientific Software</td>
</tr>
<tr>
<td>VMD</td>
<td>High quality rendering, large structures (100 million atoms), GPU acceleration for computationally demanding analysis and visualization tasks, multiple GPU support for very fast display of molecular orbitals arising in quantum chemistry calculations</td>
<td>100-125X or greater</td>
<td>Released, Version 1.9</td>
<td>Visualization from University of Illinois at Urbana-Champaign</td>
</tr>
</tbody>
</table>

GPU Perf compared against Multi-core x86 CPU socket. GPU Perf benchmarked on GPU supported features and may be a kernel to kernel perf comparison.
Quantum Chemistry
## Quantum Chemistry

<table>
<thead>
<tr>
<th>Application</th>
<th>Features Supported</th>
<th>GPU Perf</th>
<th>Release Status</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAMESS-US</td>
<td>Libqc with Rys Quadrature Algorithm, integral evaluation, closed shell Fock matrix construction</td>
<td>2.5X</td>
<td>Released</td>
<td>Single GPU supported in 10/1/10 release. Multi-GPU supported later 2011 release.</td>
</tr>
<tr>
<td>NWChem</td>
<td>Triples part of Reg-CCSD(T), CCSD &amp; EOMCCSD task schedulers</td>
<td>3-8X projected</td>
<td>Date TBA, in development</td>
<td>Development GPGPU benchmarks: <a href="http://www.nwchem-sw.org">www.nwchem-sw.org</a></td>
</tr>
<tr>
<td>QCHEM</td>
<td>Various features including RI-MP2</td>
<td>8-14X projected</td>
<td>Date TBA, In development</td>
<td>Significant porting already</td>
</tr>
<tr>
<td>TeraChem</td>
<td>&quot;Full GPU-based solution&quot;</td>
<td>44-650X vs. GAMESS CPU ver.</td>
<td>Version 1.45 released</td>
<td>Single and Multi-GPU. Completely redesigned to exploit massive GPU parallelism</td>
</tr>
</tbody>
</table>

GPU Perf compared against Multi-core x86 CPU socket. GPU Perf benchmarked on GPU supported features and may be a kernel to kernel perf comparison.
## Material Science

<table>
<thead>
<tr>
<th>Application</th>
<th>Features Supported</th>
<th>GPU Perf</th>
<th>Release Status</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abinit</td>
<td>BigDFT - 50% of the program (short convolutions)</td>
<td>6-30X</td>
<td>Released June 2009</td>
<td><a href="http://inac.cea.fr/L_Sim/BigDFT/news.html">http://inac.cea.fr/L_Sim/BigDFT/news.html</a></td>
</tr>
<tr>
<td>Quantum-Espresso/</td>
<td>PWscf package: linear algebra (matrix multiply), explicit computational kernels, 3D FFTs</td>
<td>TBD</td>
<td>Released May 5, 2011</td>
<td>Created by Irish Centre for High-End Computing</td>
</tr>
<tr>
<td>PWscf</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

GPU Perf compared against Multi-core x86 CPU socket. GPU Perf benchmarked on GPU supported features and may be a kernel to kernel perf comparison.
Make it easy to find: Tesla Bio WorkBench

Applications
- Molecular Dynamics & Quantum Chemistry
  - Amber 11
  - GROMACS
  - TeraChem
  - LAMMPS
  - aceMD
  - GAMESS
  - HOOMD-blue

- Bio-Informatics
  - HAMMER
  - Hex (Docking)
  - CUDA-BLASTP
  - CUDASW++
  - MUMmerGPU
  - CUDA-EC

Community
- Download, Documentation
- Technical papers
- Discussion Forums
- Benchmarks & Configurations

Platforms
- Tesla Personal Supercomputer
- Tesla GPU Clusters
A Couple of Examples
AMBER and NVIDIA Collaboration

**AMBER 11**
- Collaboration supported both explicit solvent Particle Mesh Ewald Molecular Dynamics simulations (NVE, NVT, NPT) and implicit solvent Generalized Born simulations - **CUDA support in AMBER 11 - Released April 2010**

**AMBER 11 plus patches**
- Collaboration on improvements that double (2X) performance - **Released August 2011**
See the Performance Difference!
Trp-Cage (304 atom) Protein Folding Simulation via GPU Accelerated AMBER 11

80 ns/day
4 core CPU

367 ns/day
4 core CPU + Tesla M2070 GPU

Up to 4X Performance Increase

Data courtesy of AMBER.org
Outstanding AMBER Results — Just add GPUs

- **Base node configuration:** Dual Xeon X5670s and Dual Tesla M2070 GPUs on each node

**JAC DFHR**
- Dihydrofolate reductase (DFHR)
- 23,558 atoms
- Joint AMBER CHARMM (JAC) benchmark
- Water modeled as an explicit TIP3P solvent

And Get Up to 3.5X Performance Increase
Adding GPUs Improves Performance on Small and Large Molecules

- **JAC DFHR (23,558 atoms)**
  - No GPU: 1.0
  - With GPU: 190% higher

- **Cellulose NPT (408,609 atoms)**
  - No GPU: 1.0
  - With GPU: 300% higher

**Base node configuration:**
Dual Xeon X5670s and Dual Tesla M2070 GPUs per node
Make Research More Productive with GPUs

Adding Two 2070 GPUs to a Node Yields a 3x Performance Increase

Base node configuration:
Dual Xeon X5670s and Dual Tesla M2070 GPUs per node

AMBER 11 on 2X E5670 CPUs + 2X Tesla C2070s (per node)
AMBER 11 on 2X E5670 CPUs (per node)
LAMMPS

Standard Molecular Dynamics code, using spatial clustering
http://lammps.sandia.gov/
Measuring speedup of GPU implementation of Lennard-Jones and Morse potential models.

Execution profile is concentrated in force calculations.
Good thread parallelism in particle neighbor-coalescing.
Program uses MPI plus GPU.

32 nodes without using GPU's: runtime = 02:49:15
32 nodes with 2 GPU's per node: runtime = 00:49:42
Speedup is 3.4x, which is more cost-effective than adding more nodes.

So far adding a third GPU per node did not yield additional speedup.
Looking at GPU utilization pattern to find more efficient load balance.
LAMMPS Gay-Berne Benchmark
GPGPU Comparison

M2050 Node
Westmere– SLES 10
Processor: Xeon 5650 (2.66 GHz)
GPU: Tesla M2050
Cache: 12MB/processor, shared
Node: 2-processor 12-core 3-GPU SL390s

M2090 Node
Westmere– SLES 10
Processor: Xeon 5650 (2.66 GHz)
GPU: Tesla M2090
Cache: 12MB/processor, shared
Node: 2-processor 12-core 8-GPU SL390s

Each process uses 1 Core & 1 GPU

Performance relative to 12 Core + 0 GPU

High is Best
**LAMMPS Lennard Jones Benchmark**

**Single-node Performance**

**Cluster:**
- **Westmere– SLES 10**
- **Processor:** Xeon 5650 (2.66 GHz)
- **GPU:** Tesla M2050
- **Cache:** 12MB/processor, shared
- **Node:** 2-processor 12-core 3-GPU SL390s

**Build type:**
- **Double Precision only CPU**
  - Each process uses 1 Core
- **Single Precision on GPU**
  - Each process uses 1 Core & 1 GPU
- **Single/Double Precision on GPU**
  - Each process uses 1 Core & 1 GPU
- **Double Precision on GPU**
  - Each process uses 1 Core & 1 GPU

**Performance relative to:**
- Serial on 1 Core + 0 GPU

**Tick mark at full nodes**
LAMMPS Single SL390s Parallel Performance Sorted by Arithmetic Type

Westmere– RH EL 5.5
Processor: Xeon 5670 (2.93 GHz)
GPU: Tesla M2050
Cache: 12MB/processor, shared
Node: 2-processor 12-core 3-GPU SL390s

Peak DP Gflops/sec 515
Peak SP Gflops/sec 1030

Relative Performance

15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0

High is Best

12 Xeon cores
12 Xeon cores
12 Xeon cores
3 GPUs
3 GPUs
3 GPUs

No GPU
All Single Precision
Double Precision
Accumulation
All Double Precision

Graph showing performance comparison between No GPU, All Single Precision, Double Precision Accumulation, and All Double Precision. The performance is sorted by arithmetic type with high is best.
Upcoming Events
GPU Technology Conference 2012
May 14-17 | San Jose, CA

The one event you can’t afford to miss

- Learn about leading-edge advances in GPU computing
- Explore the research as well as the commercial applications
- Discover advances in computational visualization
- Take a deep dive into parallel programming

Ways to participate

- Speak - share your work and gain exposure as a thought leader
- Register - learn from the experts and network with your peers
- Exhibit/Sponsor - promote your company as a key player in the GPU ecosystem

www.gputechconf.com
3 days
133 hours of technical content
60 startups
Over 1,400 attendees from 40 countries
91 Research Posters
Co-located with GTC 2012...

Accelerated High Performance Computing Symposium (AHPC)
Hosted by Los Alamos National Laboratory & NVIDIA

- Learn how accelerator technologies can be leveraged in innovative ways to advance the state-of-the-art for simulations on large-scale systems
- Identify hardware and software requirements that can meet the requirements of power, scalability and fault tolerance needed for the next generation of HPC
- Understand how legacy codes can be adapted to make use of modern computing architectures
- Provide feedback to the vendor community to aid in the adoption of accelerator technologies

“The growing success of GTC makes it a natural venue for co-hosting the Accelerated HPC Symposium. This event draws senior scientists from national research labs across the globe, and their interests in hardware and software development make for a perfect match with GTC.”

~Ben Bergen, Research Scientist, Los Alamos National Laboratory

Sign up for announcements at www.gputechconf.com
Acknowledgements

NVIDIA
Dr Scott Le Grand (Google)
LAMMPS Development Team
Dr Dave Mullally (HP)
Carl Ponder (NVIDIA)
Duncan Poole (NVIDIA)
Dr Ross Walker (SDSC)
Thank You