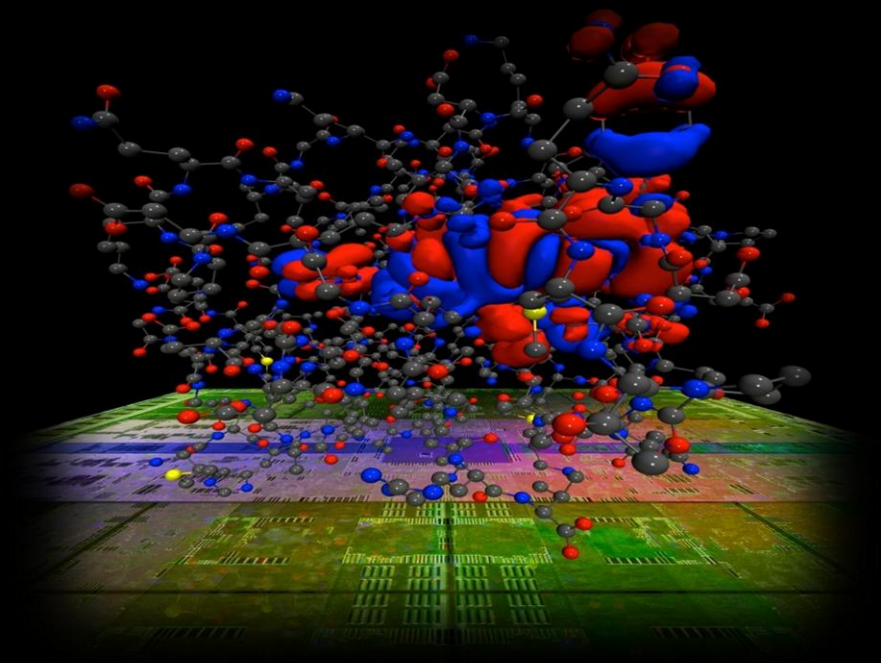


GPGPU Progress in Computational Chemistry

Mark Berger, Life and Material Sciences Alliances Manager
mberger@nvidia.com

Agenda

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

VISUALIZATION

QUADRO™

PARALLEL COMPUTING

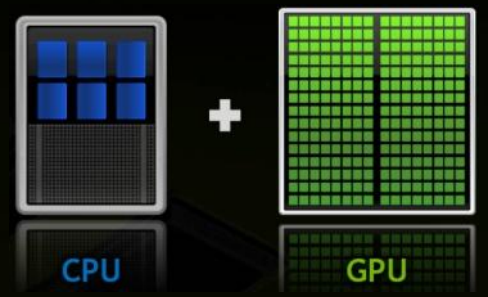
TESLA™

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

3 Years With
3 Generations

Tesla Data Center & Workstation GPU Solutions

Tesla M-series GPUs
M2090 | M2070 | M2050

Integrated CPU-GPU Servers & Blades

		M2090	M2070	M2050
Cores		512	448	448
Memory		6 GB	6 GB	3 GB
Memory bandwidth (ECC off)		177.6 GB/s	150 GB/s	148.8 GB/s
Peak Perf Gflops	Single Precision	1331	1030	1030
	Double Precision	665	515	515

Tesla C-series GPUs
C2070 | C2050

Workstations
2 to 4 Tesla GPUs

C2070	C2050
448	448
6 GB	3 GB
148.8 GB/s	148.8 GB/s
1030	1030
515	515

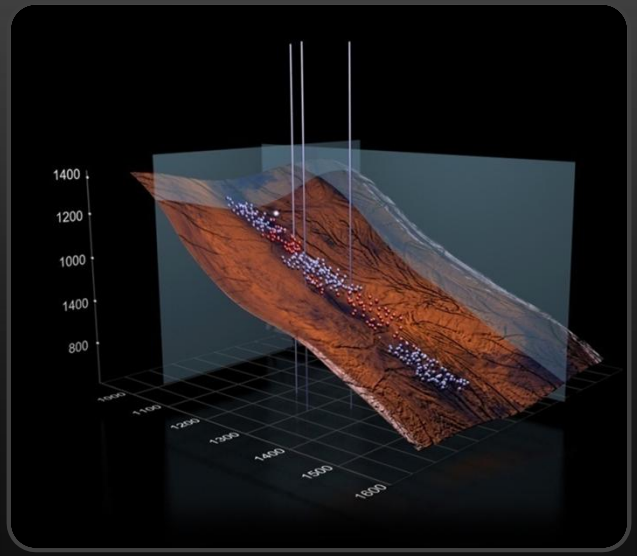
GPUs are Disruptive

Speed

Days to minutes
Minutes to seconds

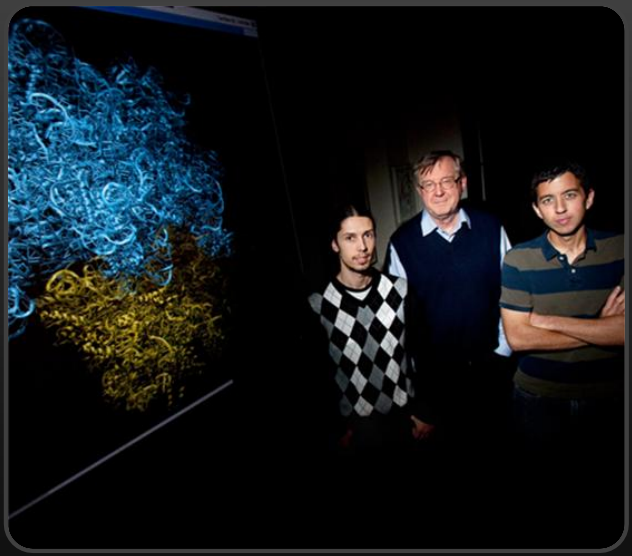
Throughput

Simulate more scenarios

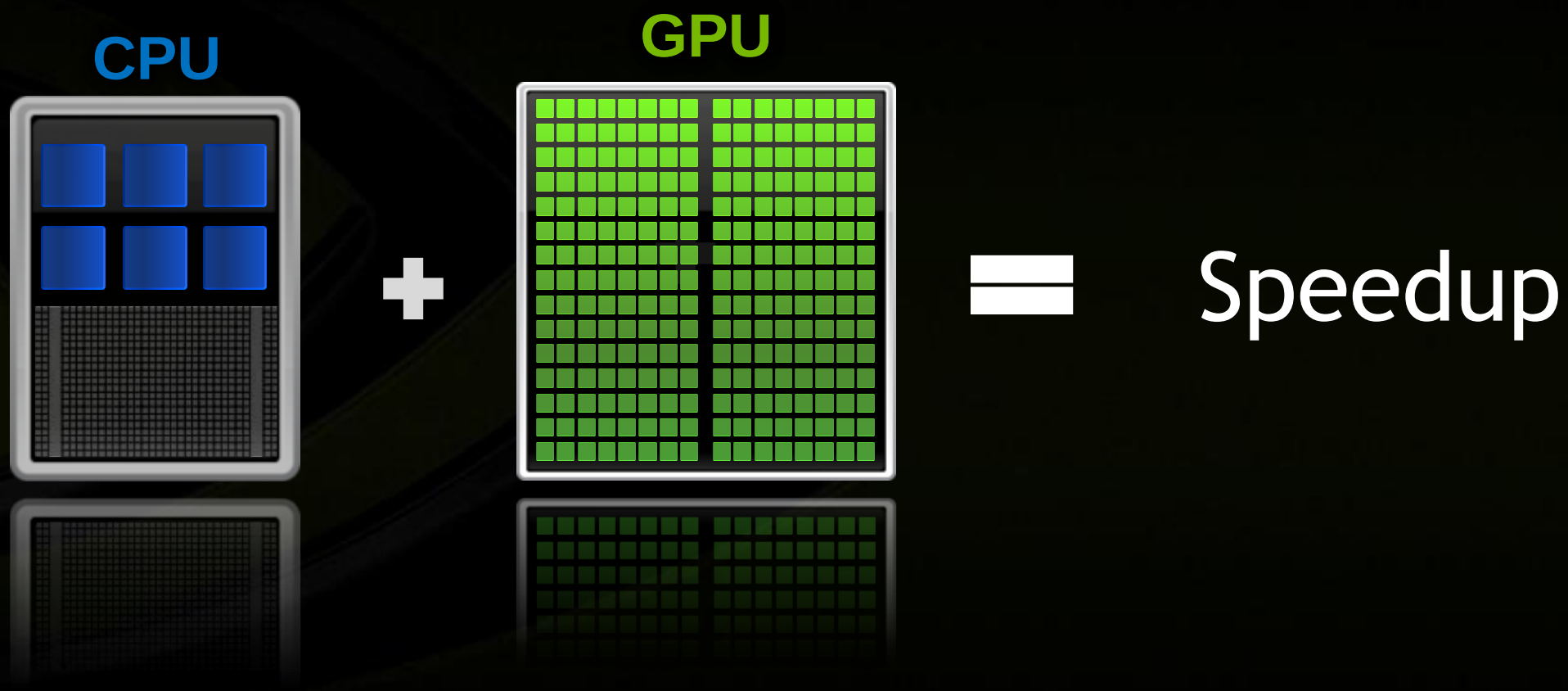


Insight

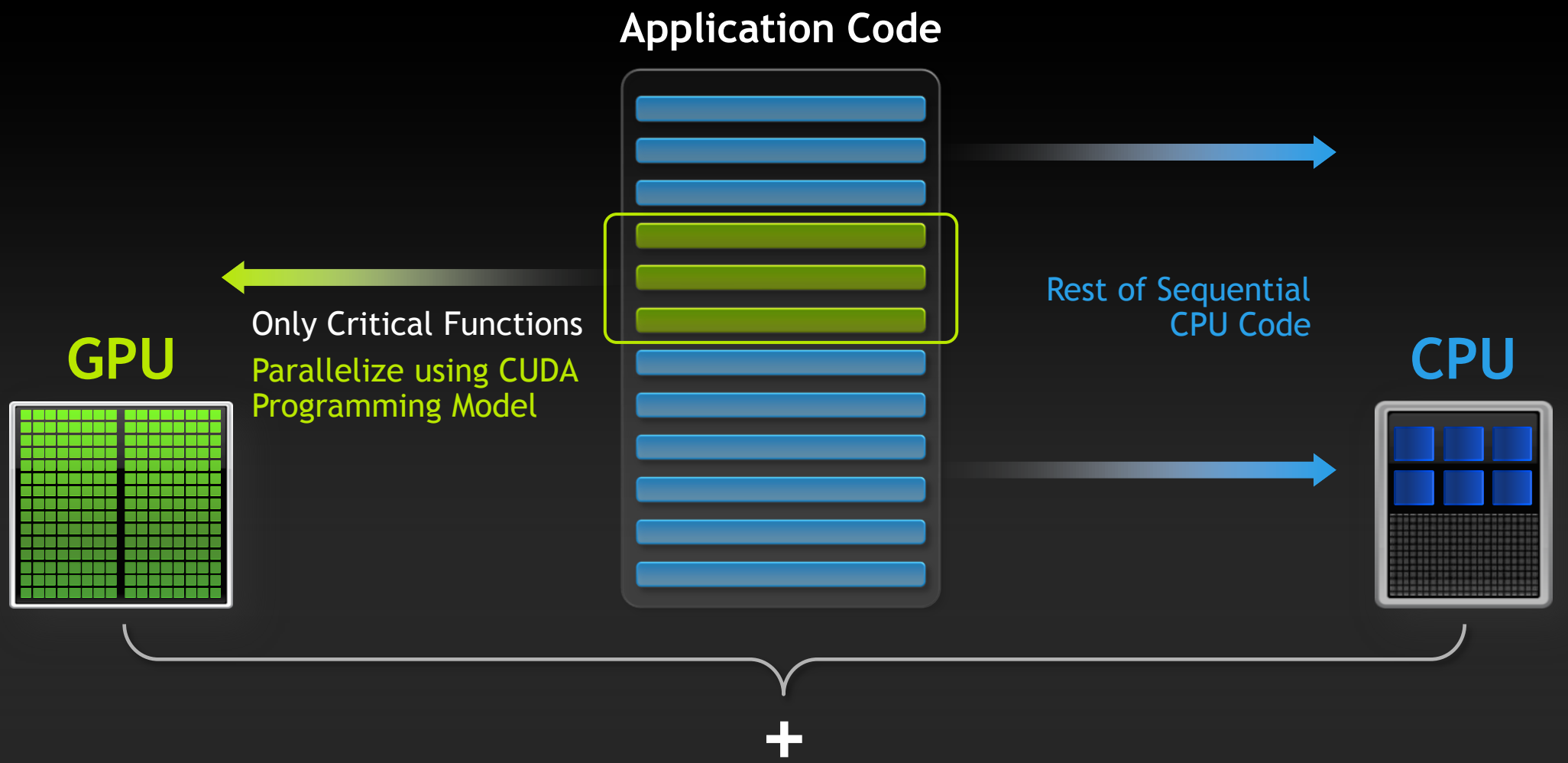
Real time analysis



Add GPUs - Accelerate Computing



Minimum Port, Big Speed-up



Tesla GPUs Power 3 of Top 5 Supercomputers

#2 : Tianhe-1A

7168 Tesla GPU's 2.5 PFLOPS

#4 : Nebulae

4650 Tesla GPU's 1.2 PFLOPS

#5 : Tsubame 2.0

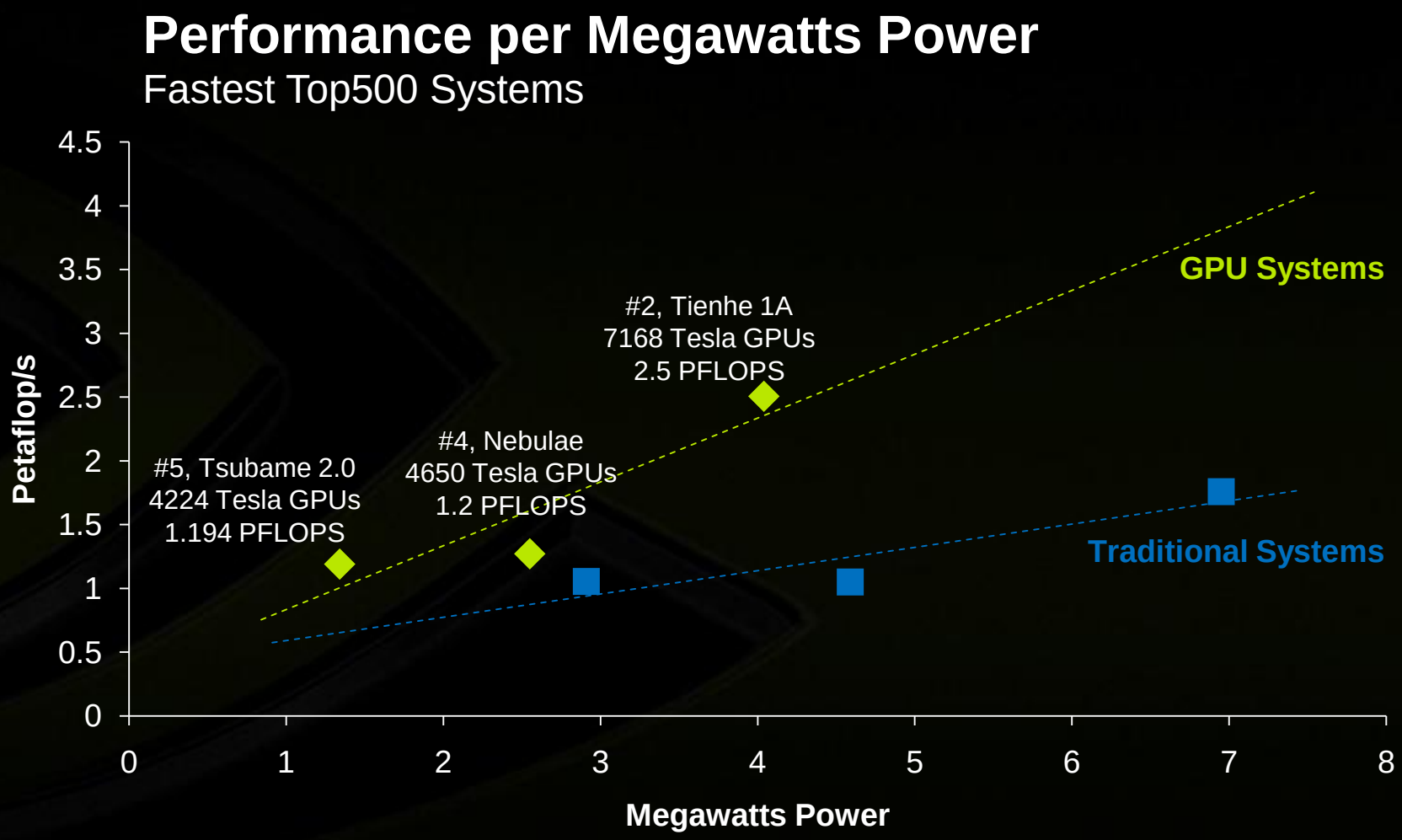
4224 Tesla GPU's 1.194 PFLOPS

"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

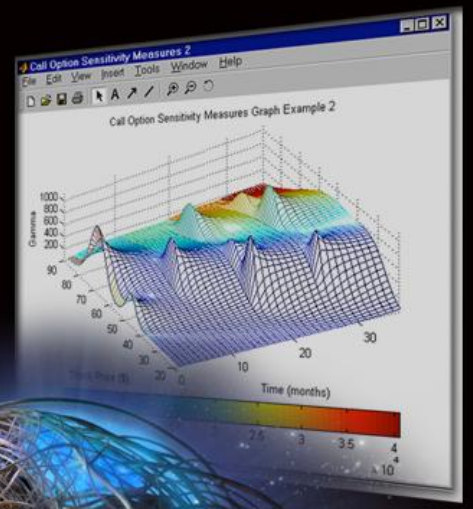
Public comments acknowledging Tianhe-1A

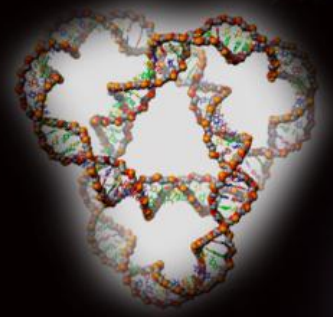
More Performance, Less Power

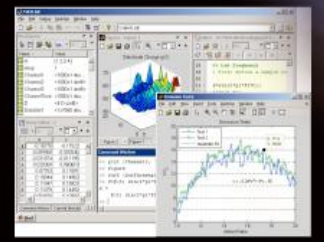


Strategic Focus on Applications

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

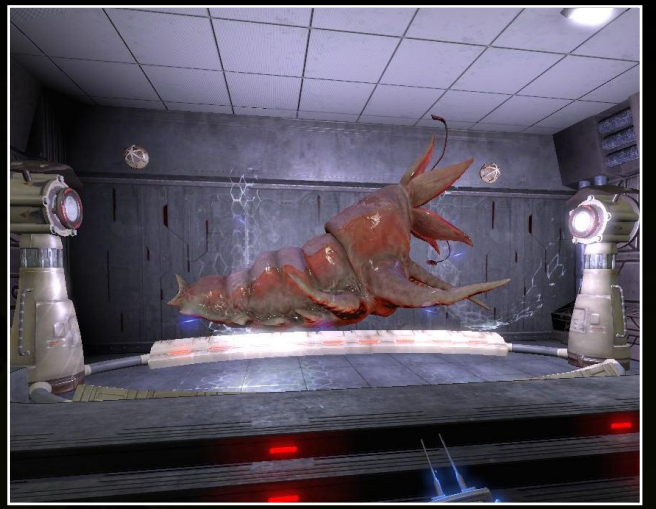




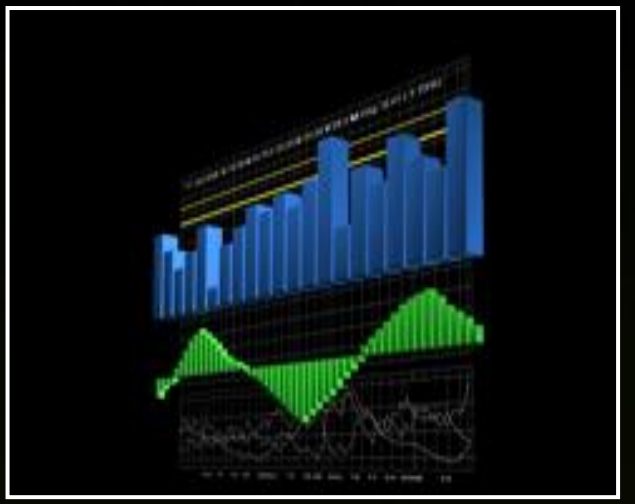


Reaching a Broad Range of Markets

Scientific computing



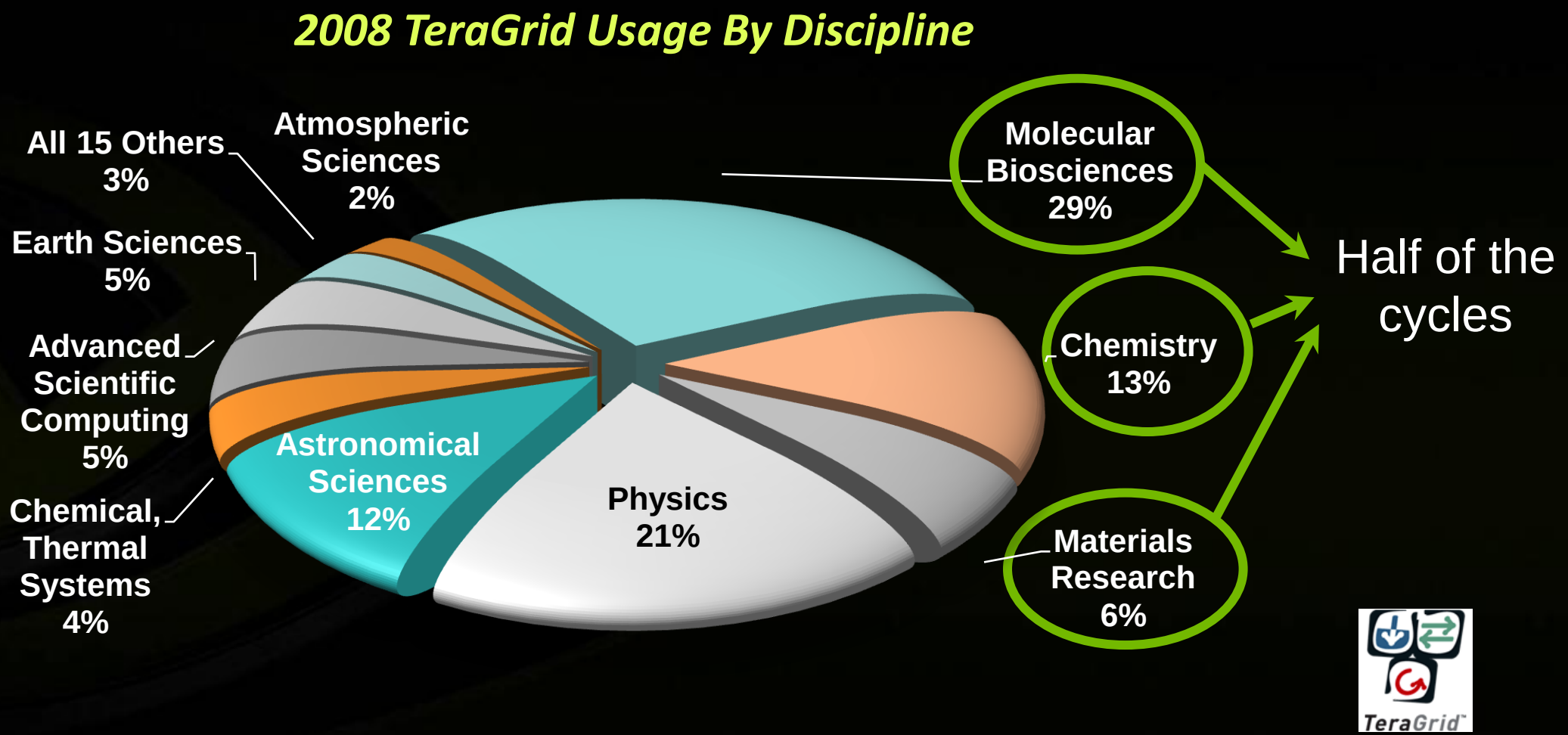
Creative pro



Education / research

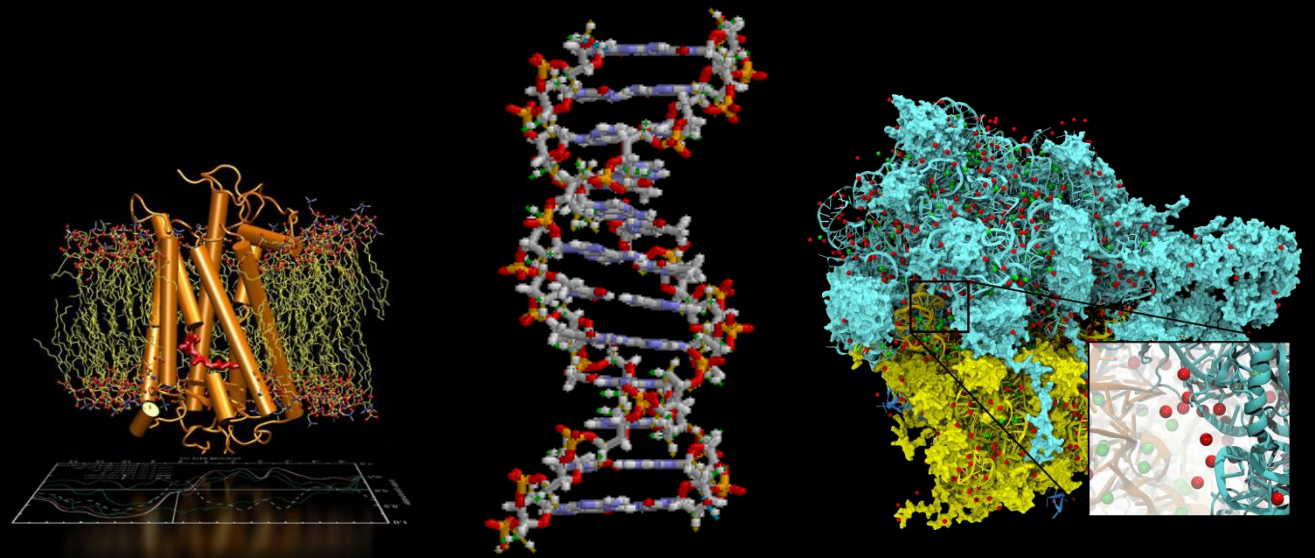
Why Computational Chemistry?

Usage of TeraGrid National Supercomputing Grid



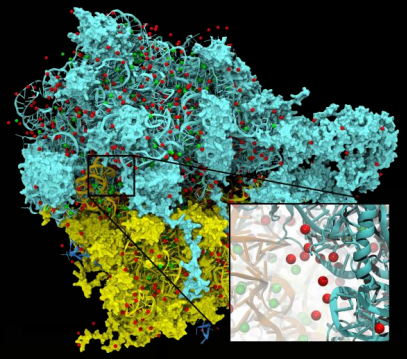
TESLA

Supercomputing



Molecular Dynamics/Mechanics

Leading MD Applications



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

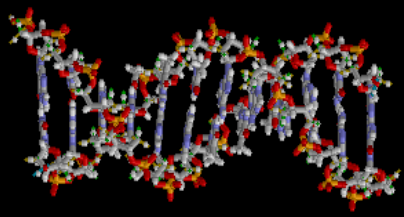
GPU Perf compared against Multi-core x86 CPU socket.
GPU Perf benchmarked on GPU supported features and may be a kernel to kernel perf comparison

Additional MD/MM Applications Ramping

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

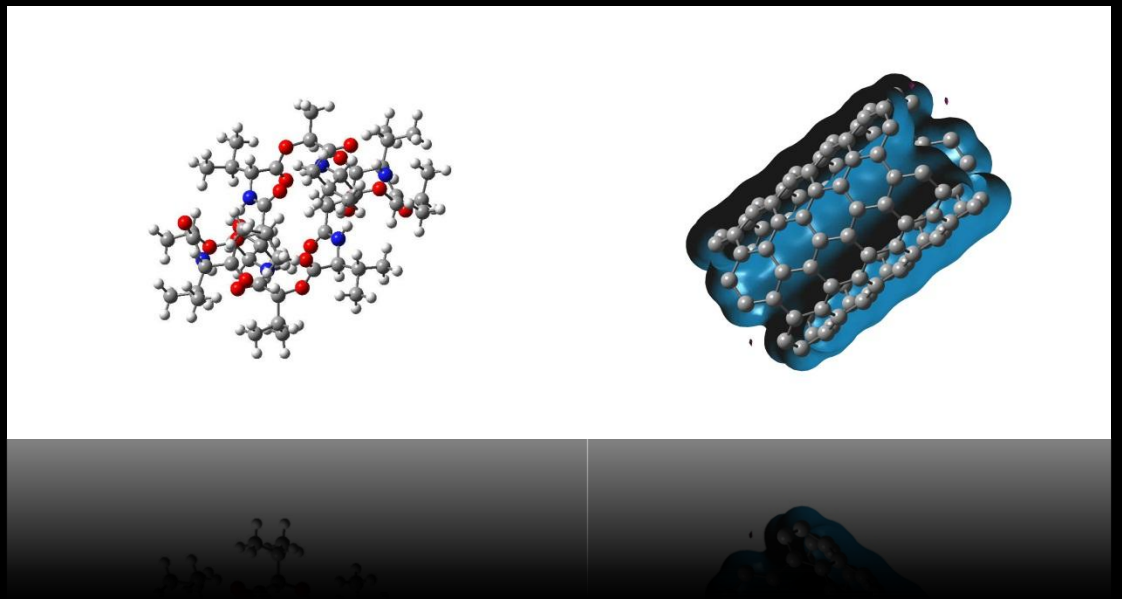
GPU Perf compared against Multi-core x86 CPU socket.
GPU Perf benchmarked on GPU supported features and may be a kernel to kernel perf comparison

Viz and "Docking" Applications



Related Applications	Features Supported	GPU Perf	Release Status	Notes
Core Hopping	GPU accelerated application	Up to 5000X	Released, Suite 2011	Single and multi-GPUs. Schrodinger, Inc.
FastROCS	Real-time shape similarity searching/comparison	800-3000X	Released	Single and multi-GPUs. Open Eyes Scientific Software
VMD	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	100-125X or greater	Released, Version 1.9	Visualization from University of Illinois at Urbana-Champaign

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

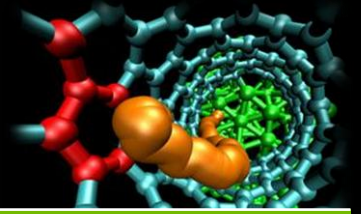


TESLA

Supercomputing

Quantum Chemistry

Quantum Chemistry

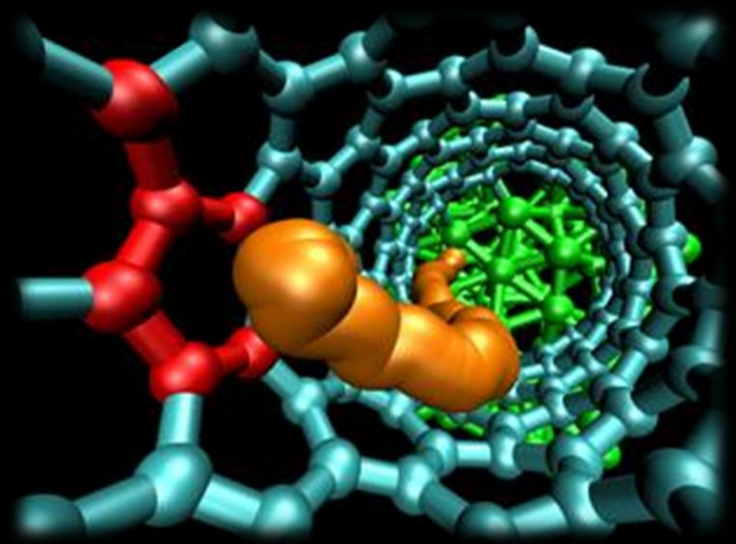


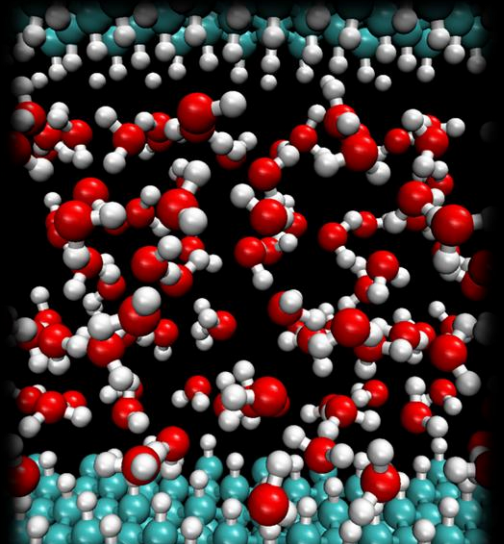
Application	Features Supported	GPU Perf	Release Status	Notes
GAMESS-US	Libqc with Rys Quadrature Algorithm, integral evaluation, closed shell Fock matrix construction	2.5X	Released	Single GPU supported in 10/1/10 release. Multi-GPU supported later 2011 release.
NWChem	Triples part of Reg-CCSD(T), CCSD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

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

TESLA

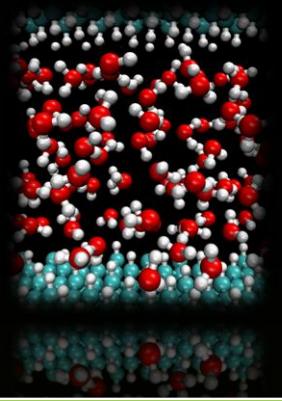
Supercomputing





Material Science

Material Science



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

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 FAST. FLEXIBLE. FREE. | TeraChem

NAMD Scalable Molecular Dynamics | LAMMPS | acemd GPUGRID.net

VMD Visual Molecular Dynamics | GAMESS | HOOMD-blue

Bio-Informatics

HMMER Scalable Informatics University at Buffalo | Hex (Docking)

CUDA-BLASTP | CUDASW++

MUMmerGPU | CUDA-EC

Community

Download, Documentation | Technical papers | Discussion Forums | Benchmarks & Configurations

Platforms

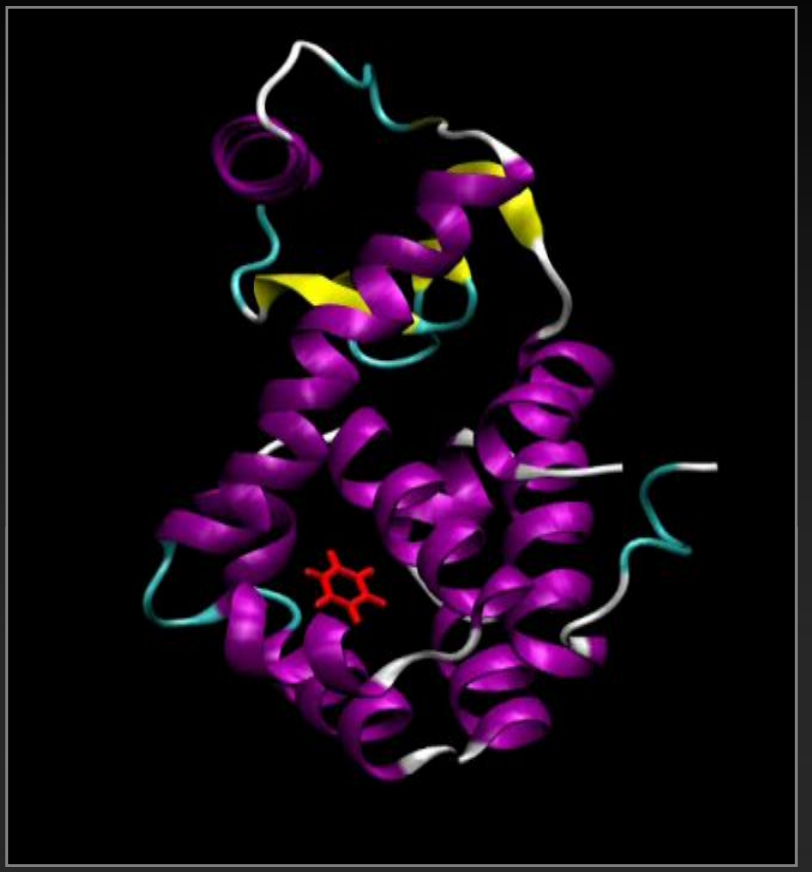
Tesla Personal Supercomputer

Tesla GPU Clusters

A Couple of Examples

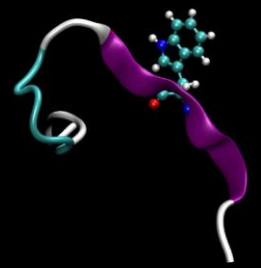
AMBER and NVIDIA Collaboration

- **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



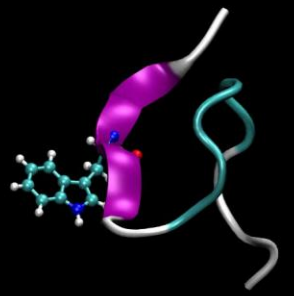
See the Performance Difference!

Trp-Cage (304 atom) Protein Folding Simulation via GPU Accelerated AMBER 11



80 ns/day

4 core CPU



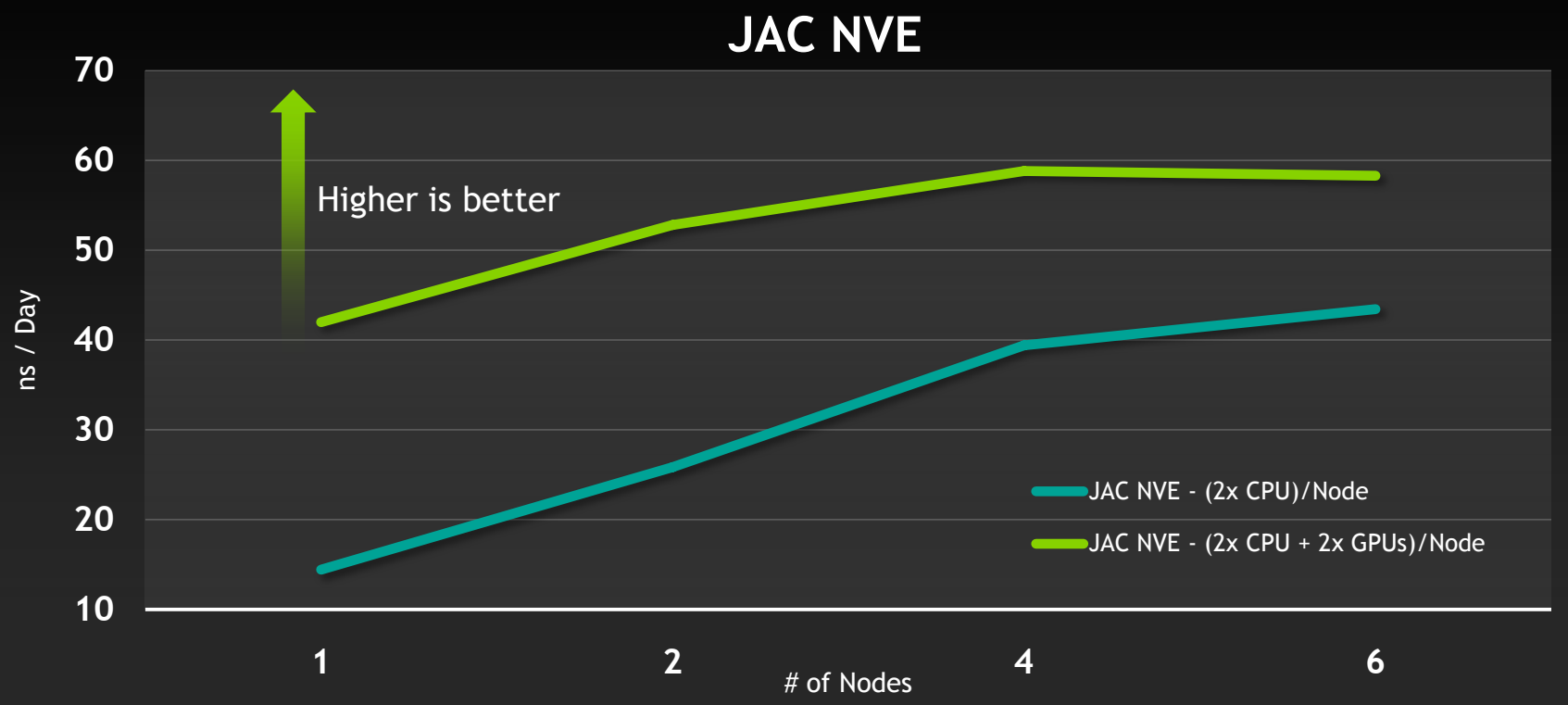
367 ns/day

4 core CPU + Tesla M2070 GPU

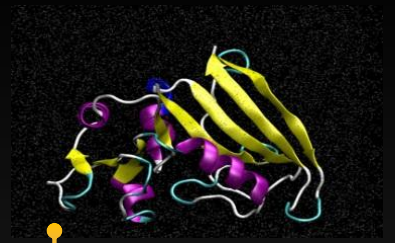
Up to 4X Performance Increase

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

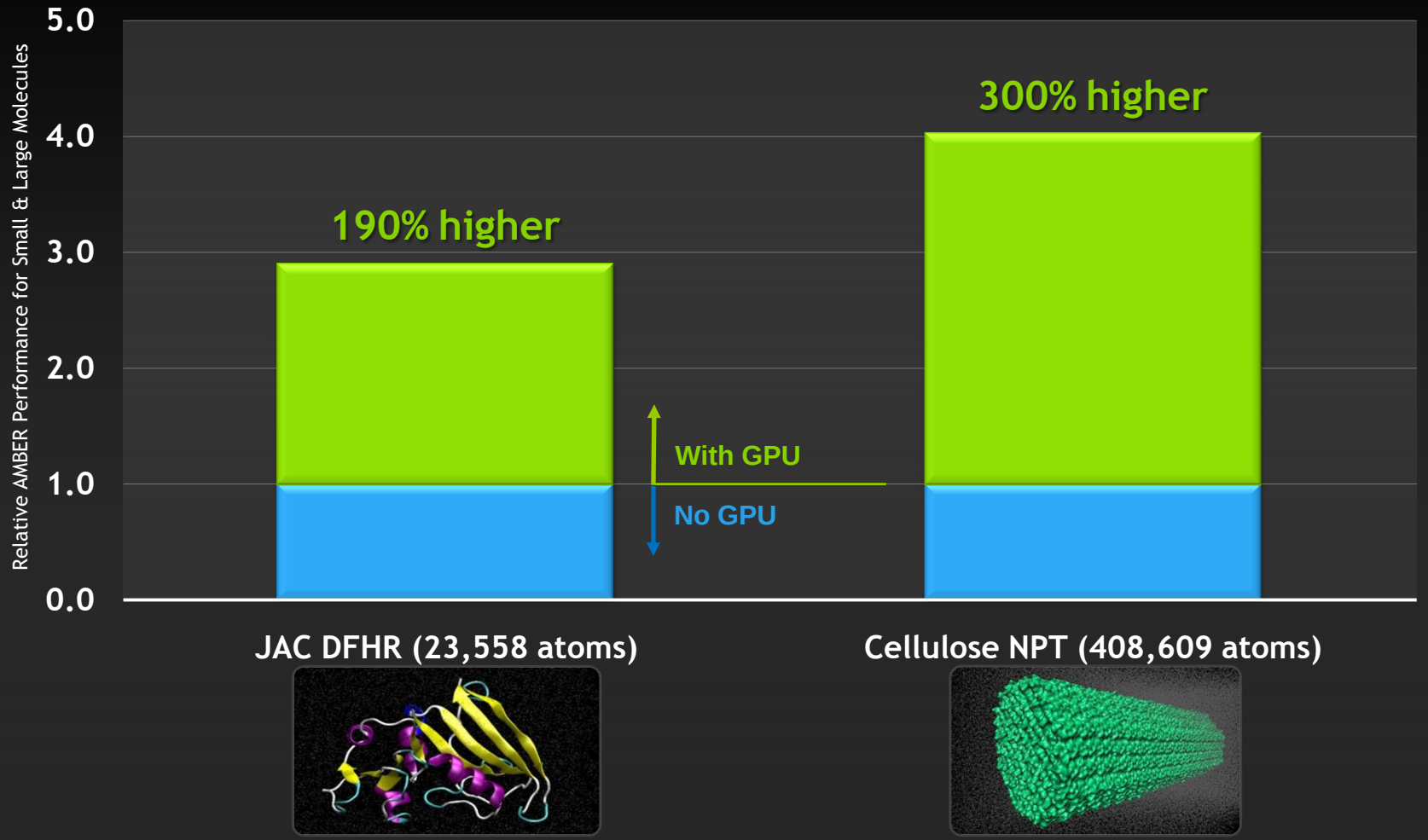


JAC DFHR

- 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

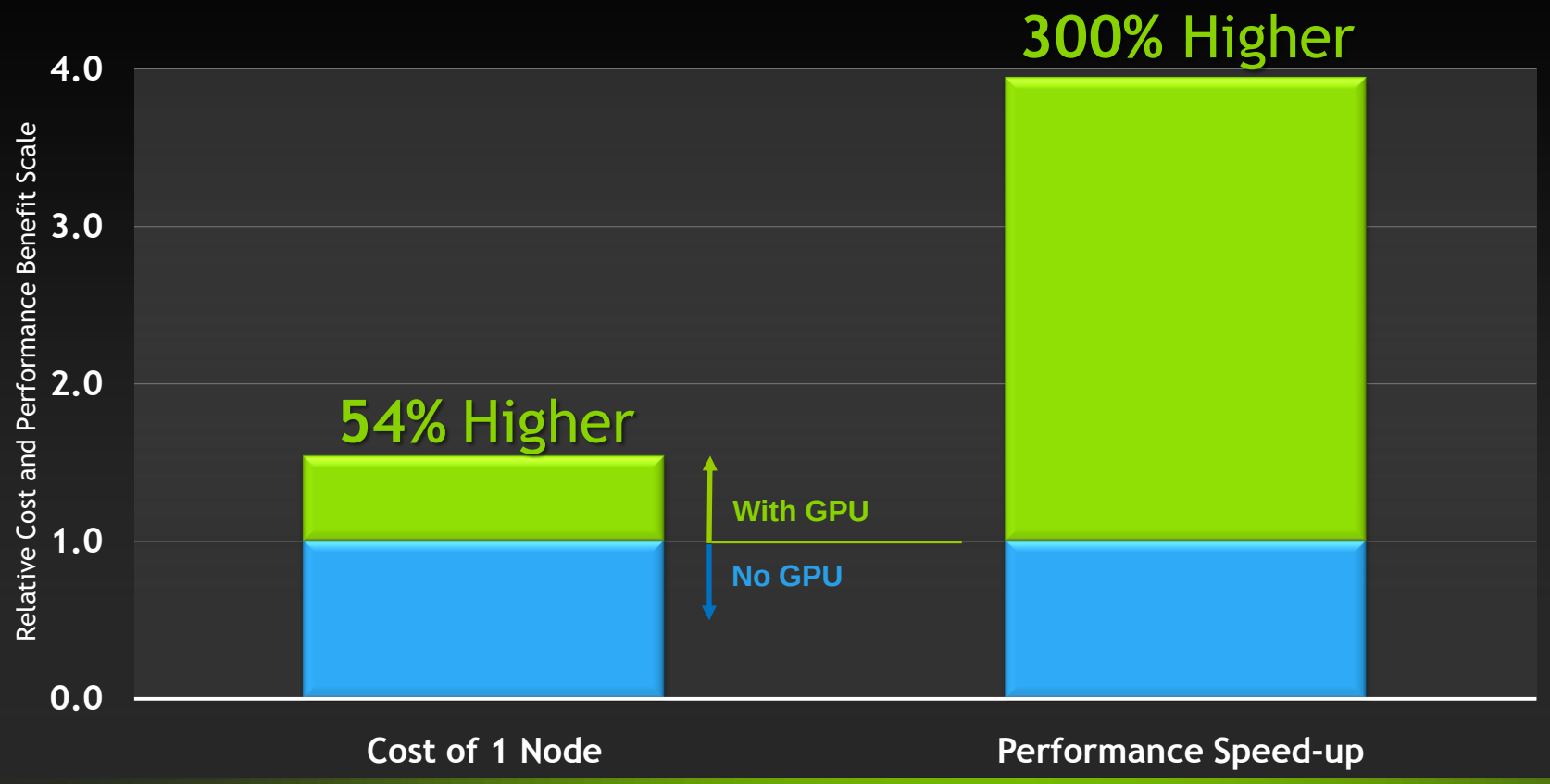
Adding GPUs Improves Performance on Small and Large Molecules



- (2x E5670 CPU + 2x Tesla C2070) per node
- 2x E5670 CPUs per node

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

Make Research More Productive with GPUs



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

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

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

LAMMPS

Carl Ponder, Ph.D.
NVIDIA Corporation

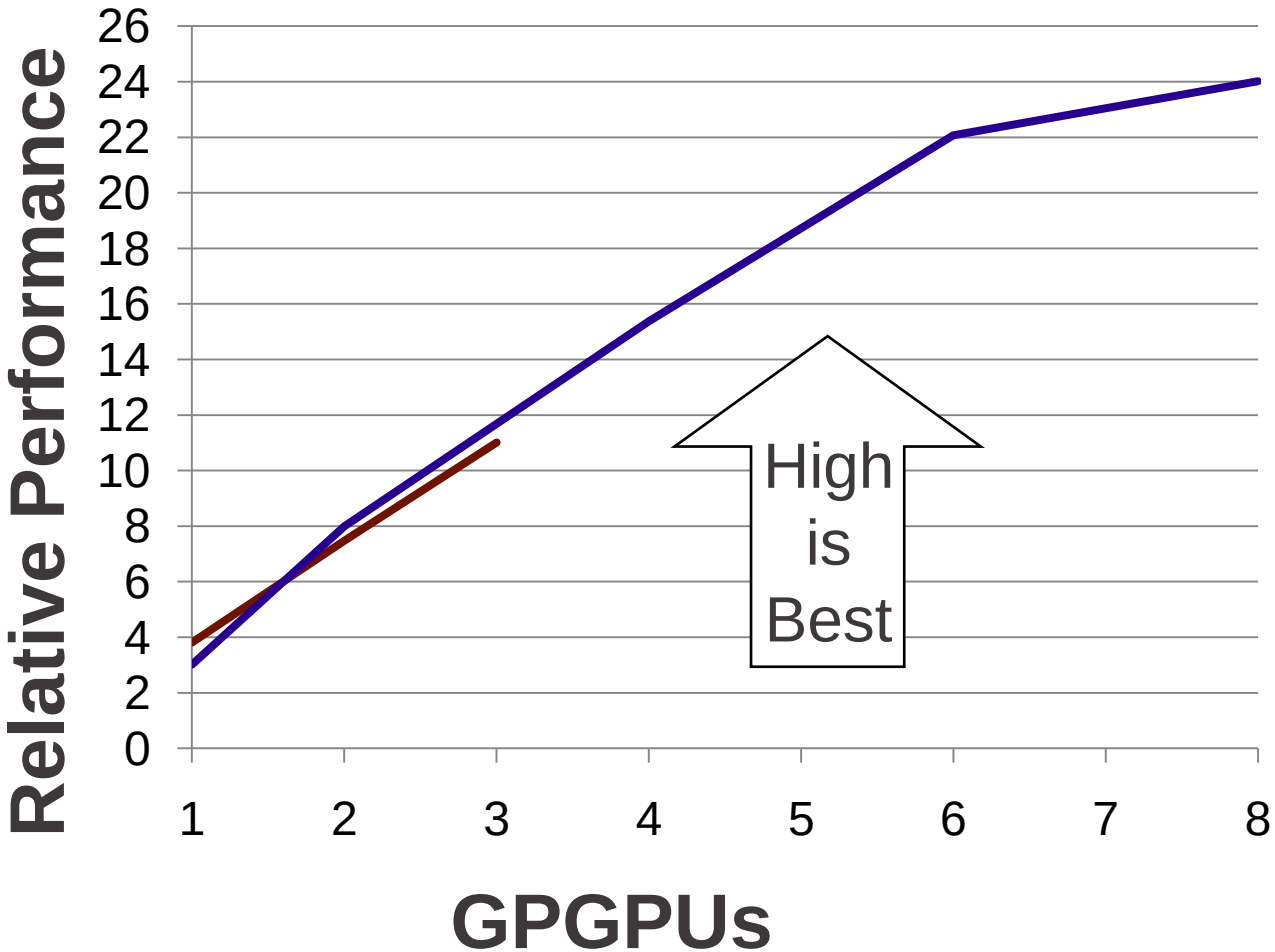
Standard Molecular Dynamics code, using spatial clustering
http://lammps.sandia.gov/
Measuring speedup of GPU implementation of Lennard-Jones and Morse potential models.

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

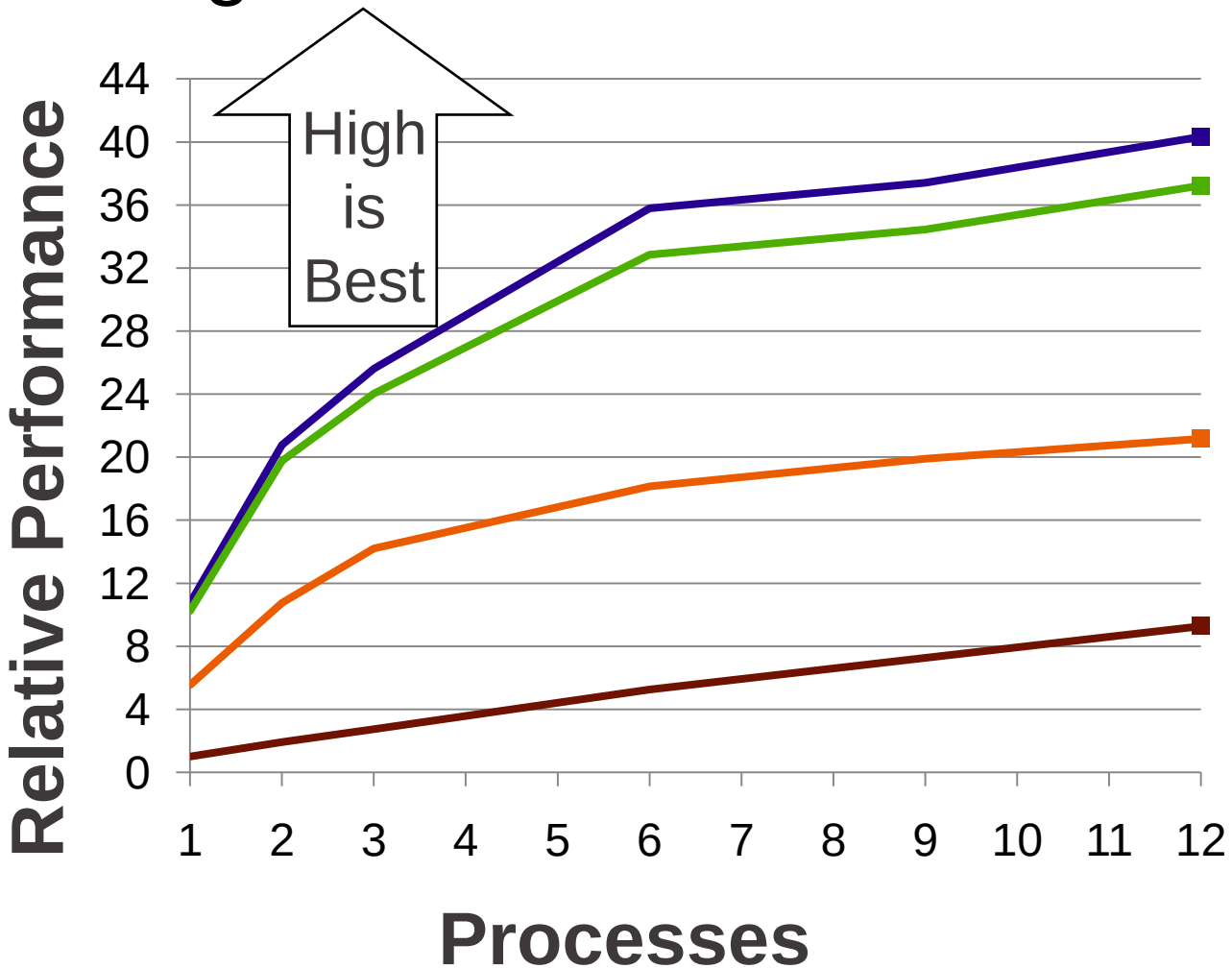
M2090 Node

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

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

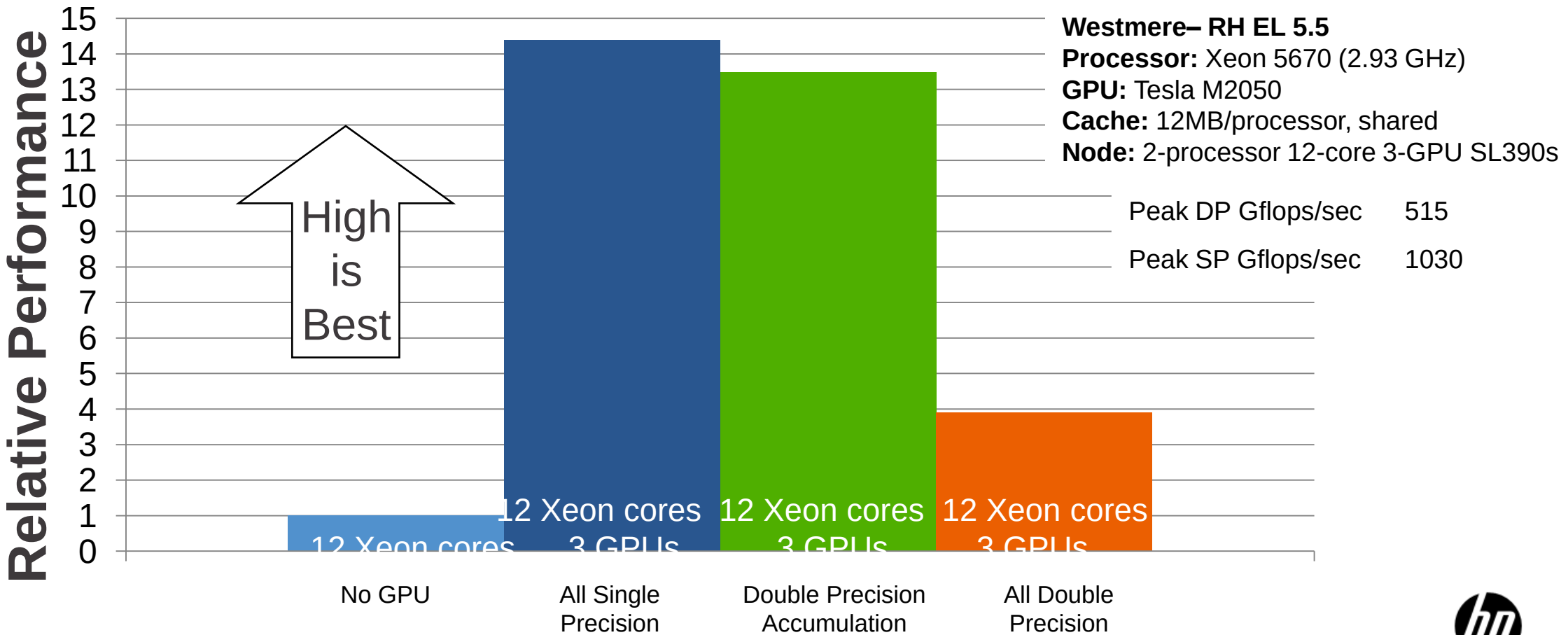
Build type:

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

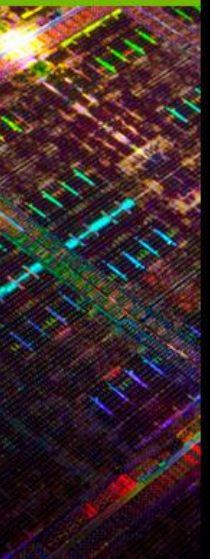
LAMMPS Single SL390s Parallel Performance Sorted by Arithmetic Type



Upcoming Events

GPU TECHNOLOGY CONFERENCE

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

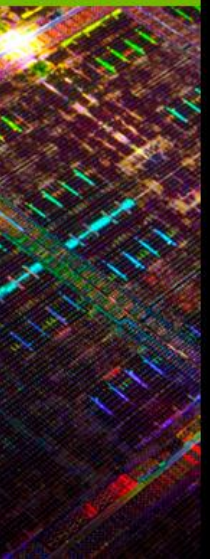
Ways to participate

- 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


GPU TECHNOLOGY CONFERENCE
3 days
133 hours of technical content
60 startups
Over 1,400 attendees from 40 countries
91 Research Posters

GPU TECHNOLOGY CONFERENCE



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
NVIDIA