Duncan Poole

GPU Progress in Life Sciences

© NVIDIA Feb26th 2010

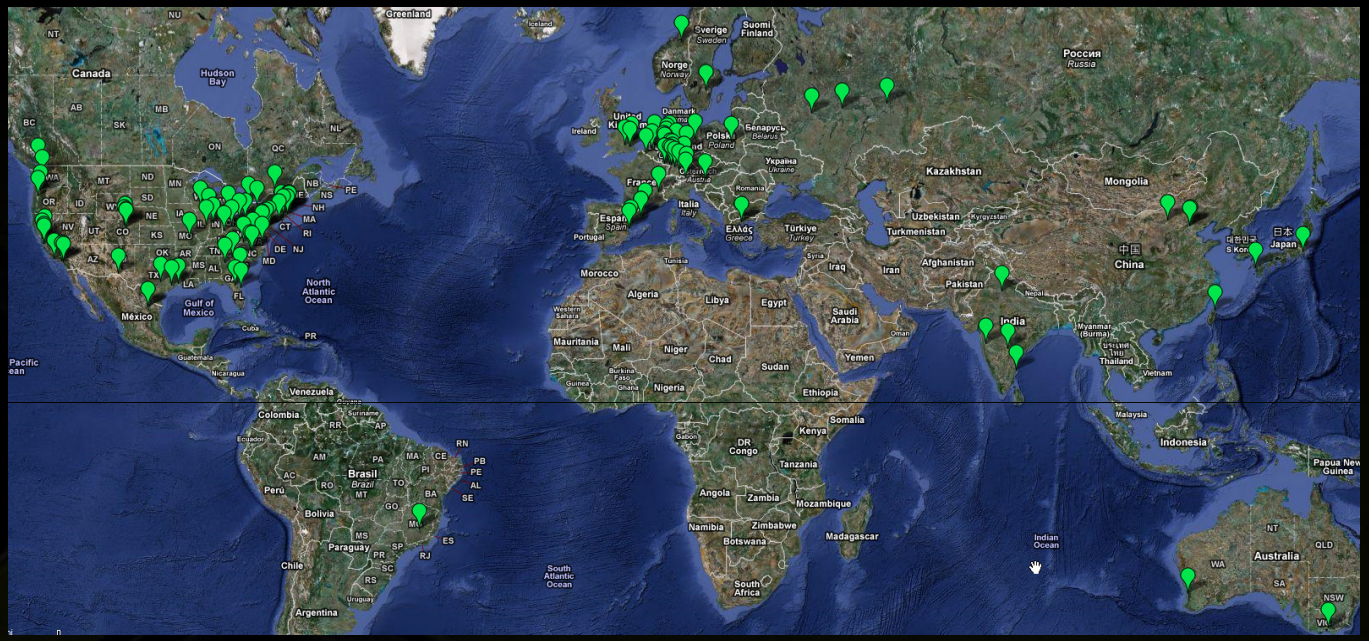
How researchers are engaging with NVIDIA

CUDA Zone

Developer Portal

Collaboration

Prof Partnership



1000+ Research Papers

200+ universities teaching CUDA

120 Million CUDA GPUs

60,000+ Active Developers

© NVIDIA Feb26th 2010

Fermi vs Tesla: The Computational GPU

Performance

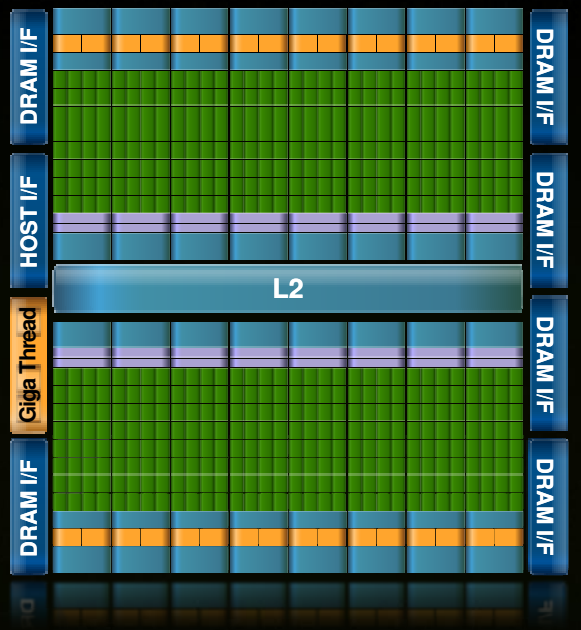
- 8x Peak Double Precision of CPUs
- IEEE 754-2008 SP & DP Floating Point

Flexibility

- Increased Shared Memory from 16 KB to 64 KB
- Added L1 and L2 Caches
- ECC on all Internal and External Memories
- Enable up to 1 TeraByte of GPU Memories
- High Speed GDDR5 Memory Interface

Usability

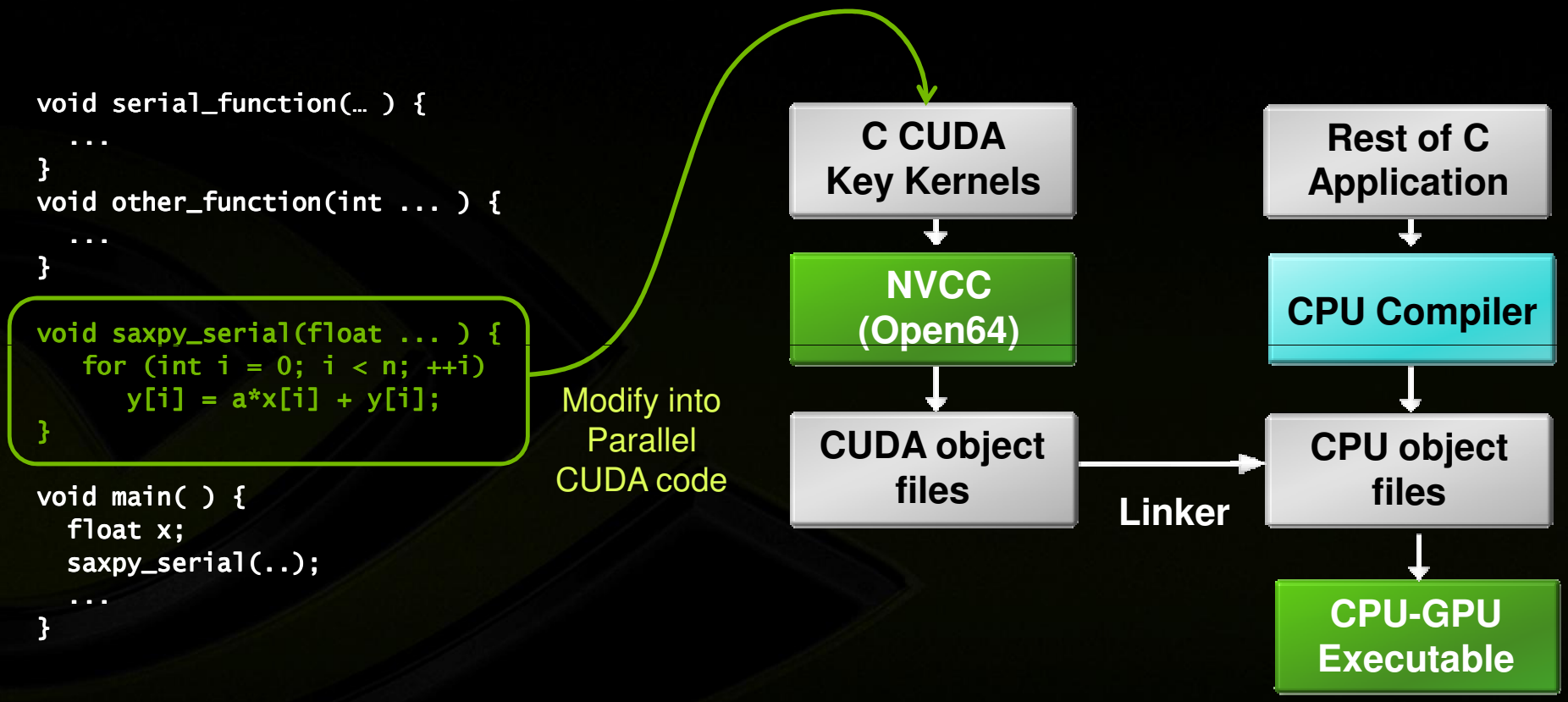
- 512 Cores vs 240 Cores
- Multiple Simultaneous Tasks on GPU
- 10x Faster Atomic Operations
- C++ Support
- System Calls, printf support



© NVIDIA Feb26th 2010

Disclaimer: Specifications subject to change

Compiling C for CUDA Applications



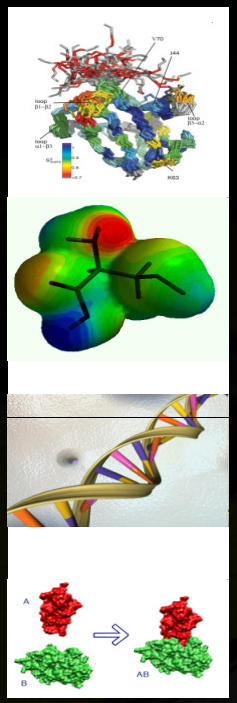
© NVIDIA Feb26th 2010

Compute Tools

	Windows	Linux	OSX
Languages	C for CUDA (NVCC) C++ for CUDA OpenCL (OCG) DirectCompute	C for CUDA (NVCC) C++ for CUDA OpenCL Fortran (PGI)	C for CUDA (NVCC) C++ for CUDA OpenCL
Libraries	Libraries	Libraries	Libraries
Debug & Profile	NEXUS Visual Profiler	cuda-gdb Visual Profiler	cuda-gdb Visual Profiler
		Allinea DDT TotalView	Apple Shark
Dev Environment	Visual Studio + NEXUS	Command-Line + cuda-gdb	Command-Line + cuda-gdb
			Apple XCode

© NVIDIA Feb26th 2010

GPU Benefits to Life Sciences Researcher



Reduced Time to Solution

Single runs currently take days, even on multiple nodes

Increase Throughput

Multiple runs at one time on a cluster

Larger Problem Sizes

Simulations will progress from 40-60K atoms to 500K+ atoms

Lowers System Costs (power costs)

Research workstations, departmental clusters, supercomputing centers

Need critical mass of life science applications:

Code User - “Do my key codes run on a GPU?”

“Does my code scale to multiple nodes?”

“Can I run longer, run larger, or multiple jobs?

© NVIDIA Feb26th 2010

Tesla Bio Workbench Applications

- AMBER (MD)
- ACEMD (MD)
- GROMACS (MD)
- GROMOS (MD)
- LAMMPS (MD)
- NAMD (MD)
- TeraChem (QC)
- VMD (Visualization MD & QC)

- Docking
  - GPU AutoDock
- Sequence analysis
  - CUDASW++ (SmithWaterman)
  - MUMmerGPU
  - GPU-HMMER
  - CUDA-MEME Motif Discovery
  - CUDA-BLAST

© NVIDIA Feb26th 2010

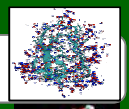
Introducing Tesla Bio WorkBench

NVIDIA

Applications

Amber 10 

GROMACS FAST. FLEXIBLE. FREE.

TeraChem 

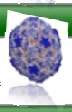
HMMER Scalable Informatics

NAMD Scalable Molecular Dynamics 

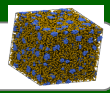
LAMMPS

ACEMD 

GPU-AutoDock

VMD Visual Molecular Dynamics 

GROMOS

HOOMD blue 

MUMmerGPU

Community

Download, Documentation

Technical papers

Discussion Forums

Benchmarks & Configurations

Platforms

Tesla Personal Supercomputer

Tesla GPU Clusters

© NVIDIA Feb26th 2010

Molecular Dynamics and Quantum Chemistry Applications

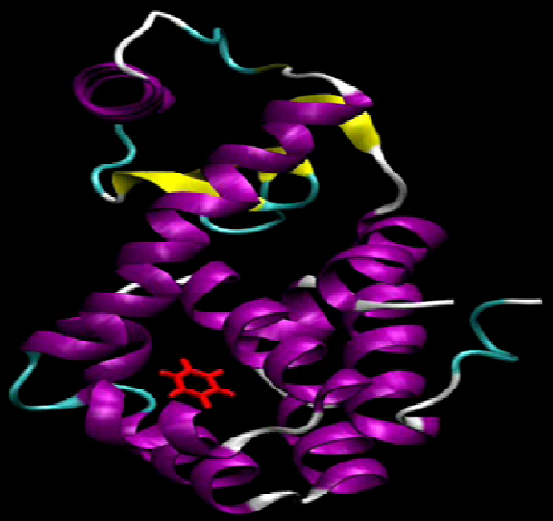
© NVIDIA Feb26th 2010

AMBER: Assisted Model Building with Energy Refinement

- Collaboration with NVIDIA to produce CUDA version of AMBER.
  - PMEMD Engine
  - Implicit Solvent GB (V1.0 complete)
  - Explicit Solvent PME (in progress)

- Focus on accuracy.
  - It MUST pass the AMBER regression tests.
  - Energy conservation MUST be comparable to double precision CPU code.

AMBER 10 CUDA Patch is located here – will be standard for AMBER 11



Source: "NVIDIA GPU Acceleration of Molecular Dynamics Simulations in AMBER" – Ross C. Walker

© NVIDIA Feb26th 2010

AMBER Molecular Dynamics

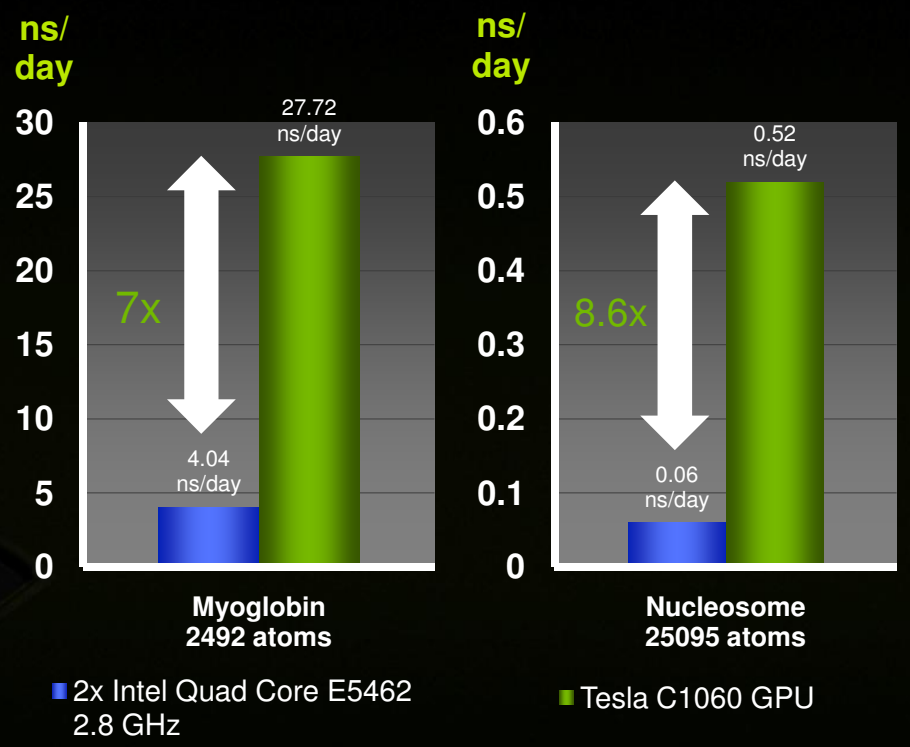
Patch Now
- **Generalized Born**

Q1 2010
- **PME: Particle Mesh Ewald**
- **Beta release**

Q2 2010
- **Multi-GPU + MPI support**
- **Beta 2 release**

Performance
- GB 65-120x single core
- PME 24 cores

More Info
http://www.nvidia.com/object/amber_on_tesla.html



Data courtesy of San Diego Supercomputing Center

© NVIDIA Feb26th 2010

AMBER: Assisted Model Building with Energy Refinement

© NVIDIA Feb26th 2010

GROMACS Molecular Dynamics

Beta now

- Particle Mesh Ewald (PME)
- Implicit solvent GB
- Arbitrary forms of non-bonded interactions

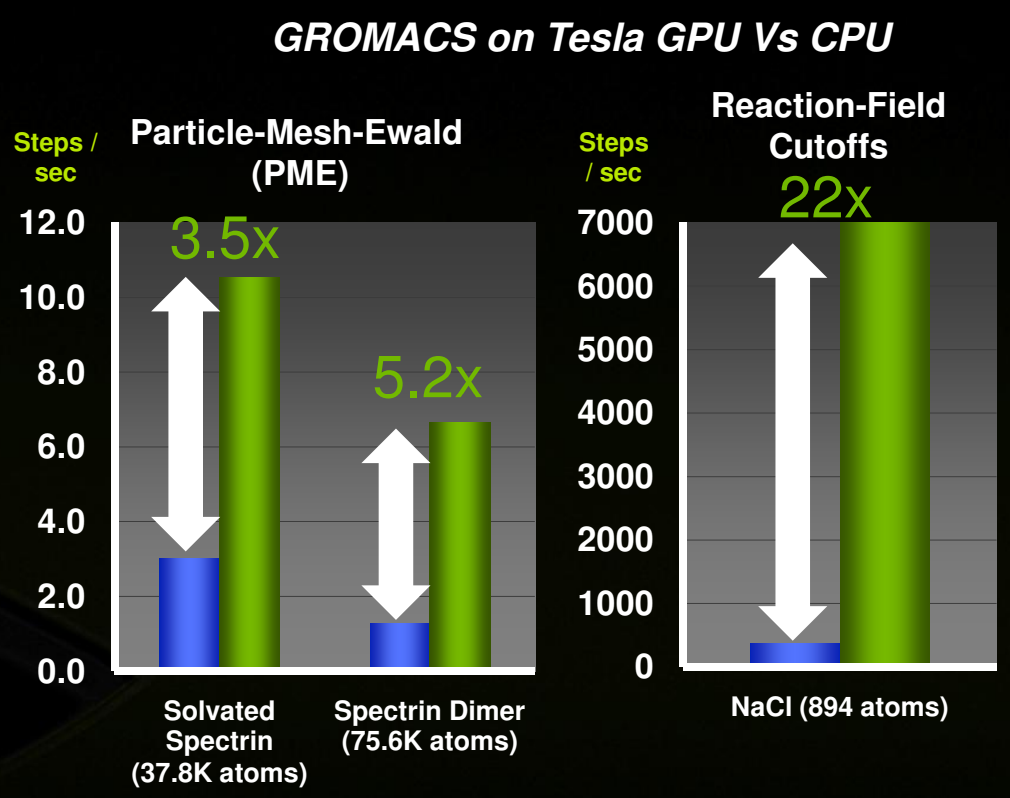
Q2 2010

- Multi-GPU + MPI support
- Beta 2 release

- PME results
  - 1 Tesla GPU 3.5x-4.7x faster than CPU

More Info

http://www.nvidia.com/object/gromacs_on_tesla.html



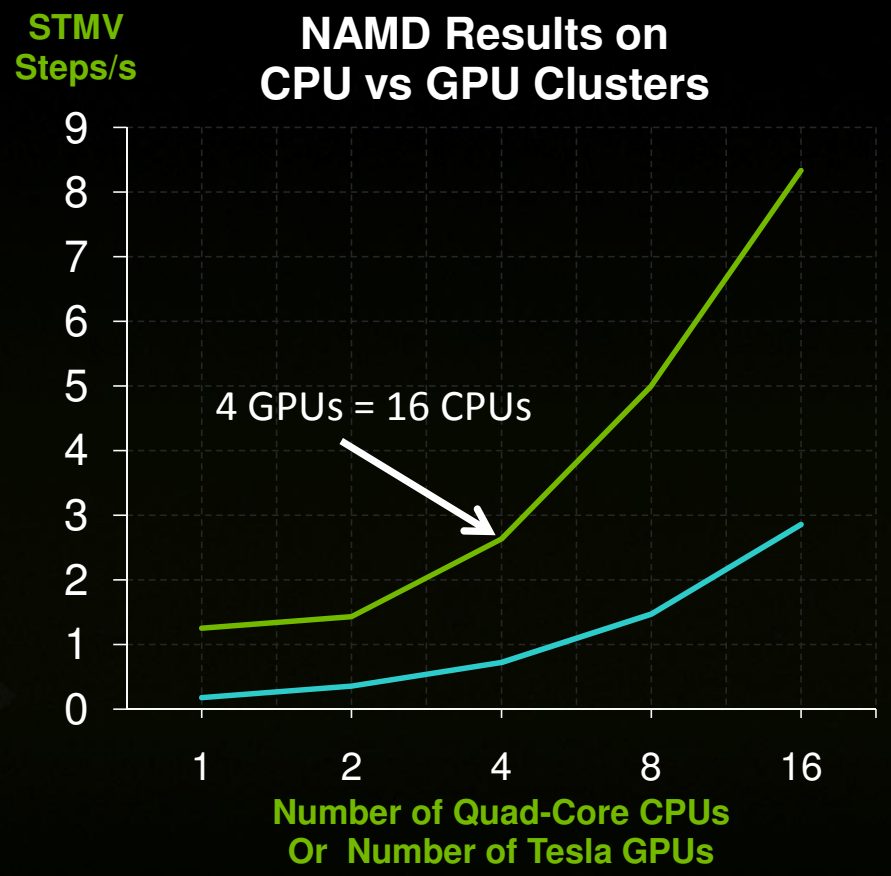
Data courtesy of Stockholm Center for Biomembrane Research

© NVIDIA Feb26th 2010

NAMD: Scaling Molecular Dynamics on a GPU Cluster

- Feature complete on CUDA : available in NAMD 2.7 Beta 2
  - Full electrostatics with PME
  - Multiple time-stepping
  - 1-4 Exclusions
- 4 GPU Tesla PSC outperforms 8 CPU servers
- Scales to a GPU cluster

More Info
http://www.nvidia.com/object/namd_on_tesla.html



Data courtesy of Theoretical and Computational Bio-physics Group, UIUC

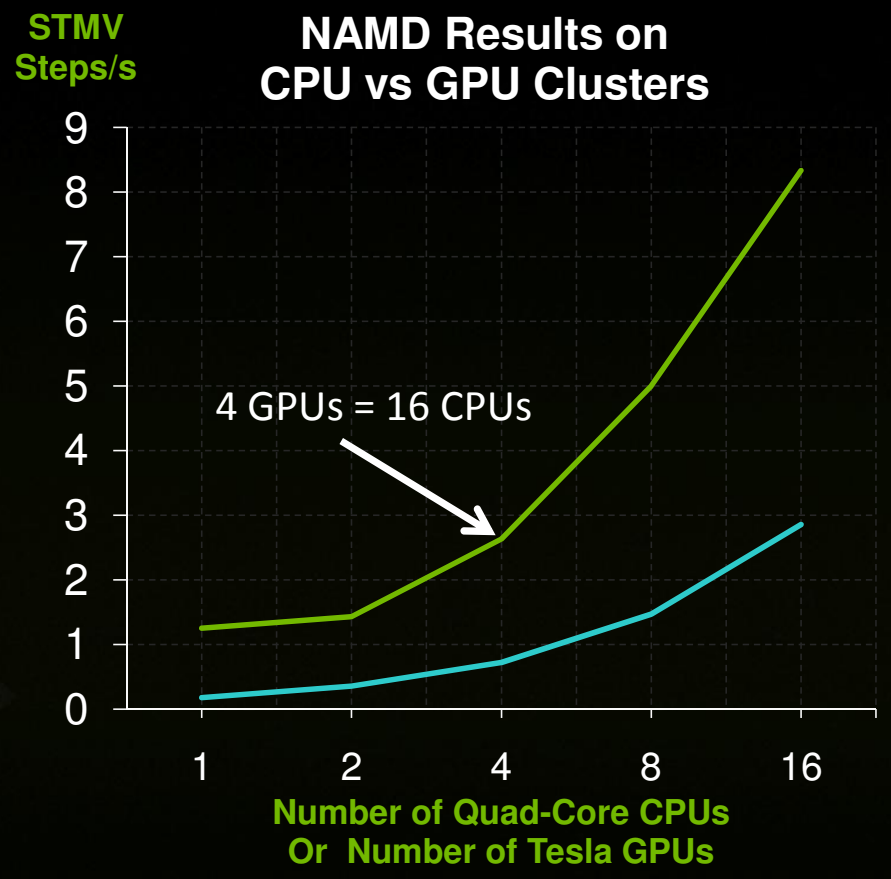
© NVIDIA Feb26th 2010

NAMD: Scaling Molecular Dynamics on a GPU Cluster

- **NAMD 2.7 Beta 2**
- **4 GPU Tesla PSC outperforms 8 CPU servers**

"We expect GPUs to maintain their current factor-of-10 advantage in peak performance relative to CPUs, while their obtained performance advantage for well-suited problems continues to grow."
– Phillips, Stone ACM Oct 2009.

Researchers also noted a 2.7x better performance/watt



Data courtesy of Theoretical and Computational Bio-physics Group, UIUC

More Info
http://www.nvidia.com/object/namd_on_tesla.html

© NVIDIA Feb26th 2010

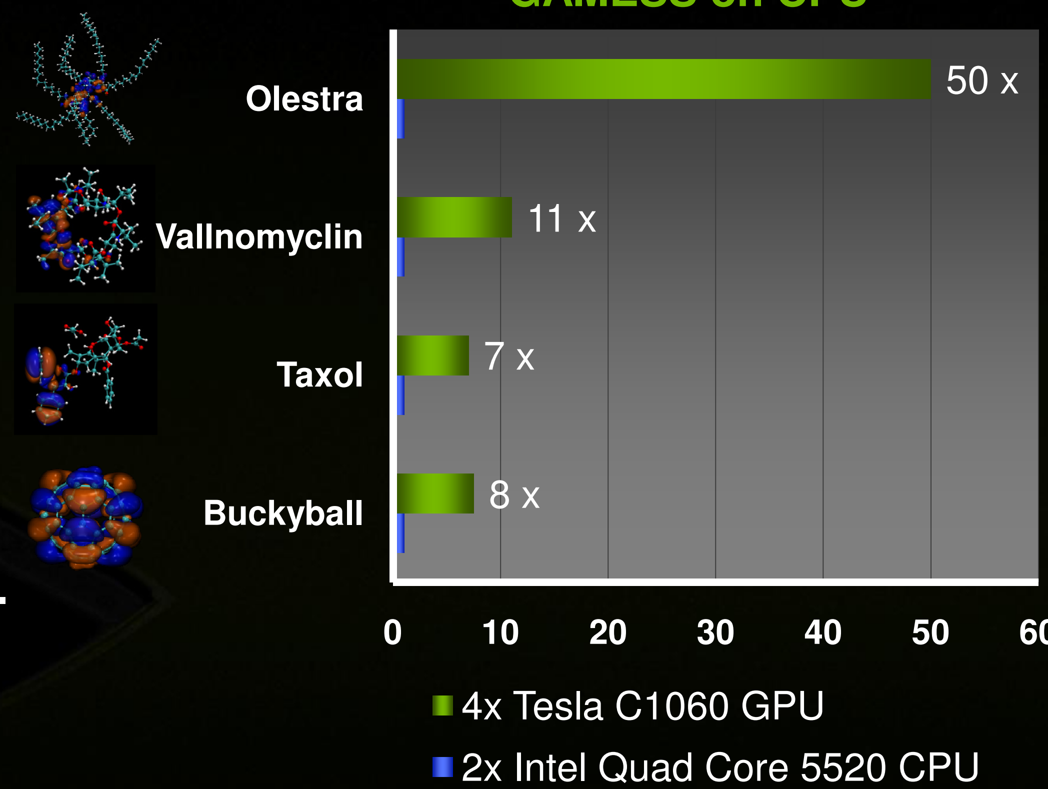
TeraChem: Quantum Chemistry Package for GPUs

Beta now
- HF, Kohn-Sham, DFT
- Multiple GPUs supported

Q1 2010
- Full release
- MPI support

- **First QC SW written ground-up for GPUs**
- **4 Tesla GPUs outperform 256 quad-core CPUs**

Speedup of TeraChem on GPU vs GAMESS on CPU

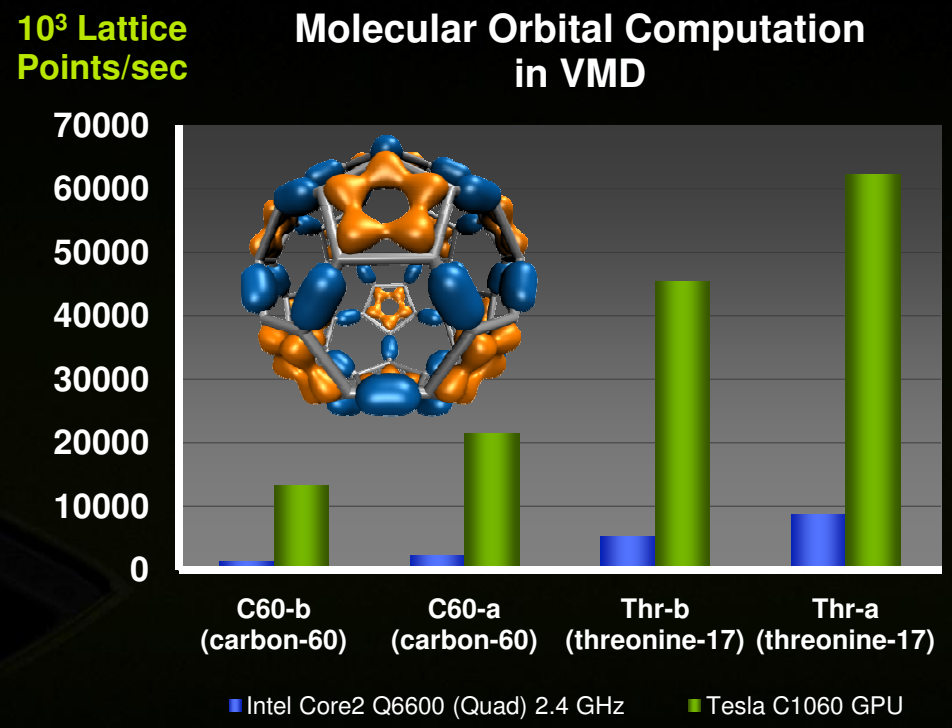


More Info
http://www.nvidia.com/object/terachem_on_tesla.html

© NVIDIA Feb26th 2010

VMD: Acceleration using CUDA GPUs

- Several CUDA applications in VMD 1.8.7
  - Molecular Orbital Display
  - Coulomb-based Ion Placement
  - Implicit Ligand Sampling
- Speedups : 20x - 100x
- Multiple GPU support enabled



Images and data courtesy of Beckman Institute for Advanced Science and Technology, UIUC

More Info

http://www.nvidia.com/object/vmd_on_tesla.html

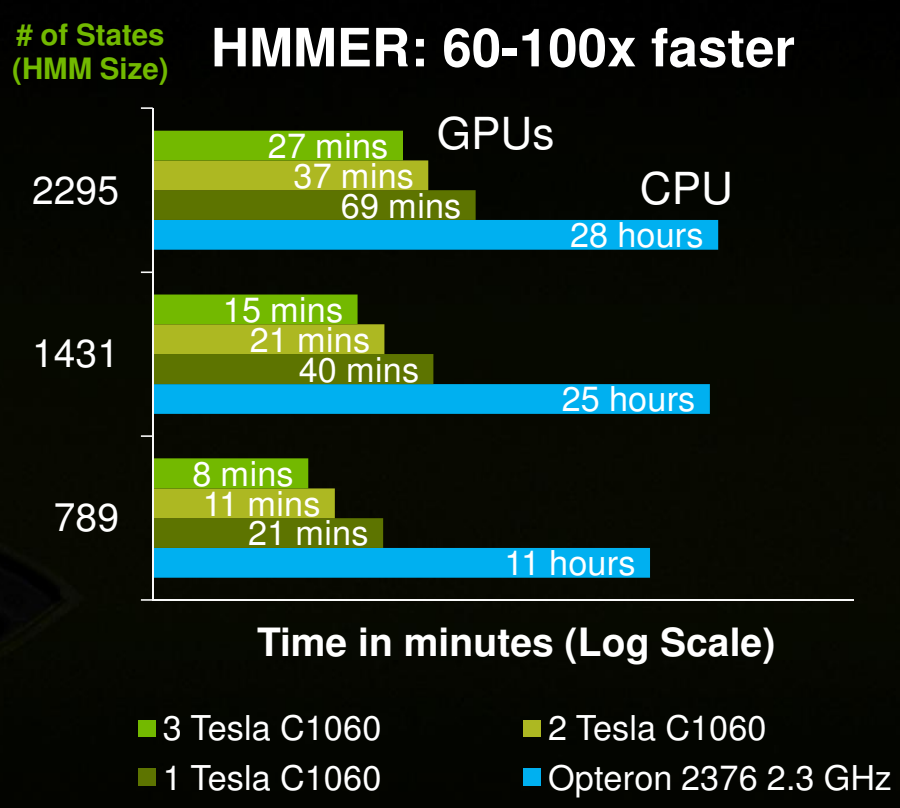
© NVIDIA Feb26th 2010

Bio-Informatics Applications

© NVIDIA Feb26th 2010

GPU-HMMER: Protein Sequence Alignment

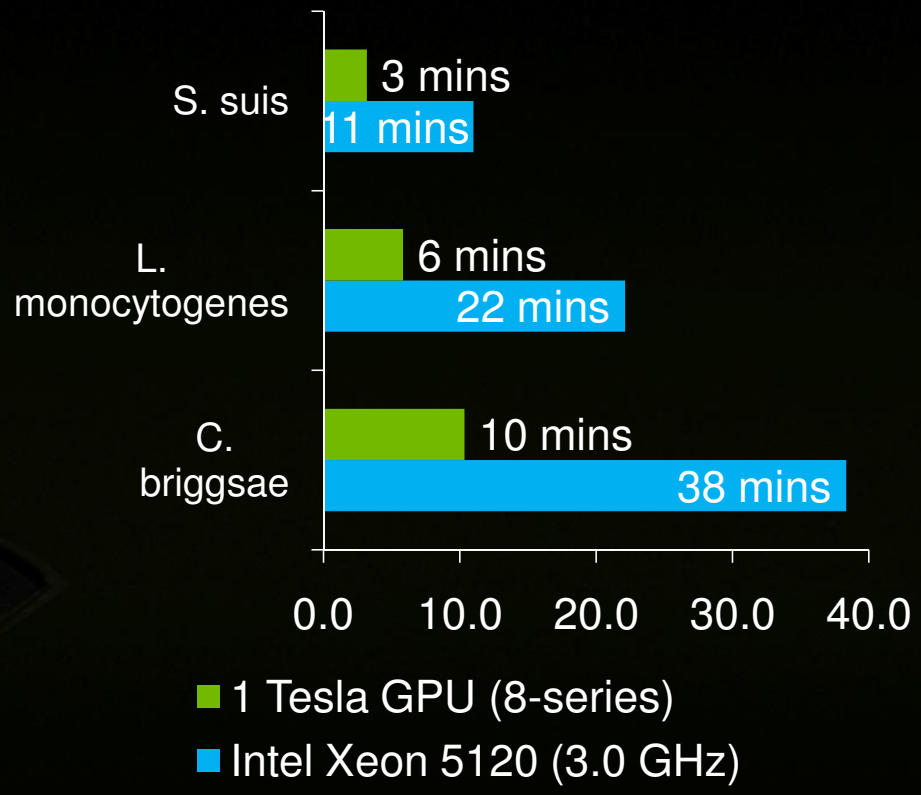
- Protein sequence alignment using profile HMMs
- Available now
  - Supports multiple GPUs
- Speedups range from 60-100x faster than CPU
- Download
  - http://www.mpihmmer.org/releases.htm



© NVIDIA Feb26th 2010

MUMmerGPU:
Genome Sequence Alignment

- High-throughput pair-wise local sequence alignment
  - Designed for large sequences
- Drop-in replacement for "mummer" component in MUMmer software
- Speedups 3.5x to 3.75x
- Download
  - http://mummergpu.sourceforge.net



© NVIDIA Feb26th 2010

More Coming Soon

- CUDA Smith Waterman: Sequence Alignment
- CUDA MEME: Motif Discovery
- GPU AutoDock: Protein Docking

© NVIDIA Feb26th 2010

Backup

Lessons Learned

Although seemingly harmless at first glance, one must not recycle random numbers in any way or conformational sampling is greatly reduced. In addition, doing so, even if the numbers are permuted, causes positional and/or rotational aliasing of the molecule because they do not quite sum to zero.

On a related note, one cannot skimp on random number quality or the system will rapidly heat up. CUDA is sophisticated enough to allow generation of high quality random numbers

Summation of forces *must* be deterministic. This means one cannot use atomic ops to perform this sum unless one first converts to a fixed point representation, and one cannot use atomic ops to order neighbor lists or bins or simulations will not be reproducible. Floating point math is not associative.

Verification - a lot of time was put into validating this work. The goal of this project was to port production quality code to the GPU not to do research but to accelerate the use of this code for every day science.

© NVIDIA Feb26th 2010

Fermi

3x the shared memory and allows bonded force kernel to cache incoherent stores, and faster sorting leads to faster nonbond cell calculation

Accelerated 64-bit shared memory atomic operations allow for faster (and simpler) charge interpolation as well as local accumulation of forces into shared memory (if converted to 64-bit fixed point first)

Vastly improved double-precision performance accelerates FFTs, force accumulation, and integration

© NVIDIA Feb26th 2010