SOS14 Panel: Future of GPU HW

Jeffrey Vetter

Presented to

SOS14, Savannah

Oak Ridge National Laboratory

Georgia Tech College of Computing
Computational Science and Engineering

Panelists

- Sadaf Alam, CSCS
- Scott Hemmert, Sandia
- John Morrison, LANL
- Jeffrey Vetter, ORNL/Georgia Tech, (moderator)
- Sriram Swaminarayan, LANL, ex officio

Charge

- **Future directions for GPU HW and systems built from them***
  - Status: Is your organization currently pursuing heterogeneous computing? Small scale, large scale? Why or why not?
  - Apps: Which specific applications at your site (could) exploit heterogeneous manycore, if any?
  - Software: What programming models and paradigms do you use/need for these architectures to be successful?
  - Arch: What specific architectural features do/could your applications exploit?

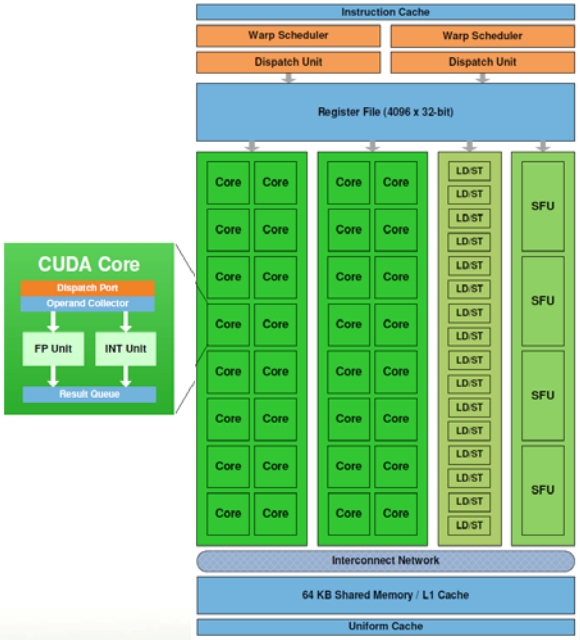
*In a public forum with multiple vendors present ☺

Keeneland – An NSF-Funded Partnership to Enable Large-scale Computational Science on Heterogeneous Architectures

- NSF Track 2D System of Innovative Design
  - Georgia Tech
  - NICS
  - ORNL
  - UTK
- Two GPU clusters
  - Initial delivery – Spring 2010
  - Full scale – Spring 2012
  - NVIDIA, HP, Intel, Qlogic
- Software tools, application development
- Operations, user support
- Education, Outreach, Training for scientists, students, industry

- Exploit graphics processors to provide extreme performance and energy efficiency

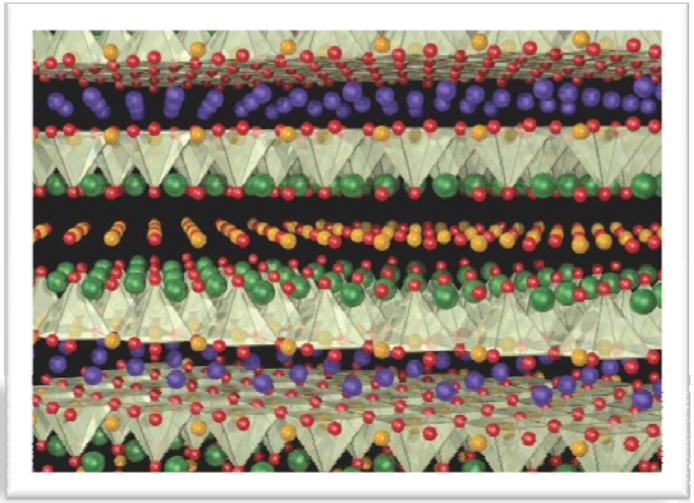
NVIDIA's new Fermi GPU

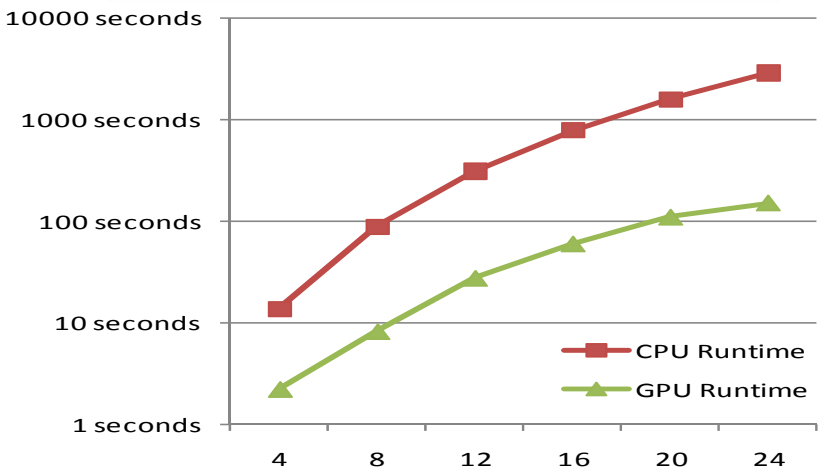


Jeffrey	Vetter
Jack	Dongarra
Richard	Fujimoto
Thomas	Schulthess
Karsten	Schwan
Phil	Andrews
Troy	Baer
Kathlyn	Boudwin
Mark	Fahey
Jim	Ferguson
Ursula	Henderson
Doug	Hudson
Ron	Hutchins
Patricia	Kovatch
Bruce	Loftis
Nathaniel	Mendoza
Jeremy	Meredith
Terry	Moore
Tracy	Rafferty
Don	Reed
Jim	Rogers
Philip	Roth
Arlene	Washington
Sudha	Yalamanchili

Computational Materials - Case Study

- Quantum Monte Carlo simulation
  - High-temperature superconductivity and other materials science
  - 2008 Gordon Bell Prize
- GPU acceleration speedup of 19x in main QMC Update routine
  - Single precision for CPU and GPU: target single-precision only cards
  - Required detailed accuracy study and mixed precision port of app
- Full parallel app is 5x faster, start to finish, on a GPU-enabled cluster



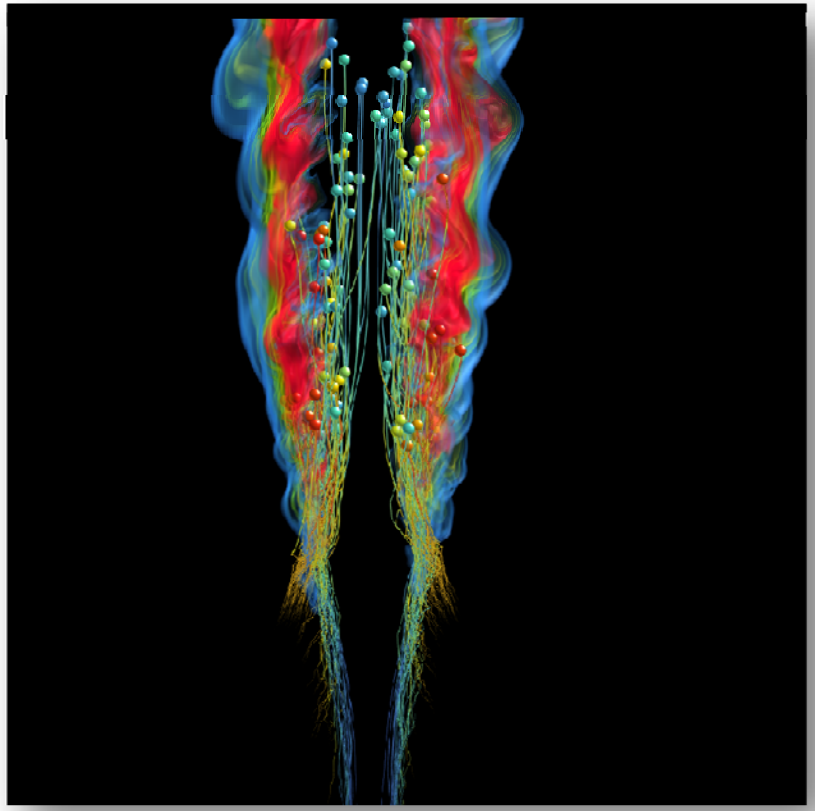


GPU study: J.S. Meredith, G. Alvarez, T.A. Maier, T.C. Schulthess, J.S. Vetter, "Accuracy and Performance of Graphics Processors: A Quantum Monte Carlo Application Case Study", *Parallel Comput., 35(3):151-63, 2009.*

Accuracy study: G. Alvarez, M.S. Summers, D.E. Maxwell, M. Eisenbach, J.S. Meredith, J. M. Larkin, J. Levesque, T. A. Maier, P.R.C. Kent, E.F. D'Azevedo, T.C. Schulthess, "New algorithm to enable 400+ TFlop/s sustained performance in simulations of disorder effects in high-Tc superconductors", SuperComputing, 2008. [*Gordon Bell Prize winner*]

Combustion with S3D – Case Study

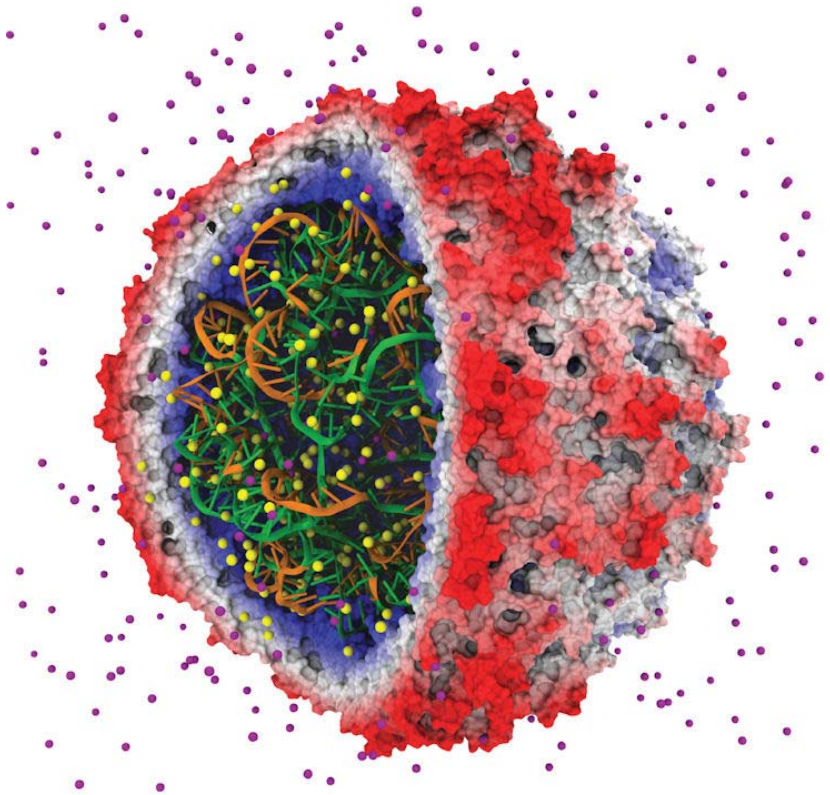
- Application for combustion - S3D
  - Massively parallel direct numerical solver (DNS) for the full compressible Navier-Stokes, total energy, species and mass continuity equations
  - Coupled with detailed chemistry
  - Scales to 150k cores on Jaguar
- Accelerated version of S3D's Getrates kernel in CUDA
  - 31.4x SP speedup
  - 16.2x DP speedup



K. Spafford, J. Meredith, J. S. Vetter, J. Chen, R. Grout, and R. Sankaran. Accelerating S3D: A GPGPU Case Study. Proceedings of the Seventh International Workshop on Algorithms, Models, and Tools for Parallel Computing on Heterogeneous Platforms (HeteroPar 2009) Delft, The Netherlands.

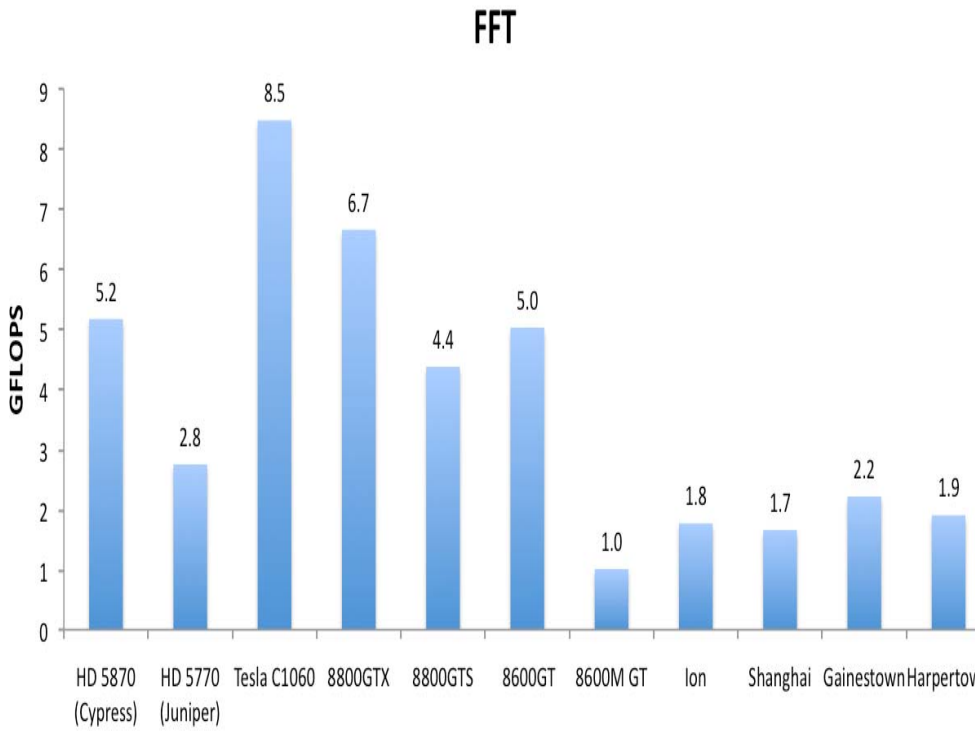
Biomolecular systems from NAMD Team

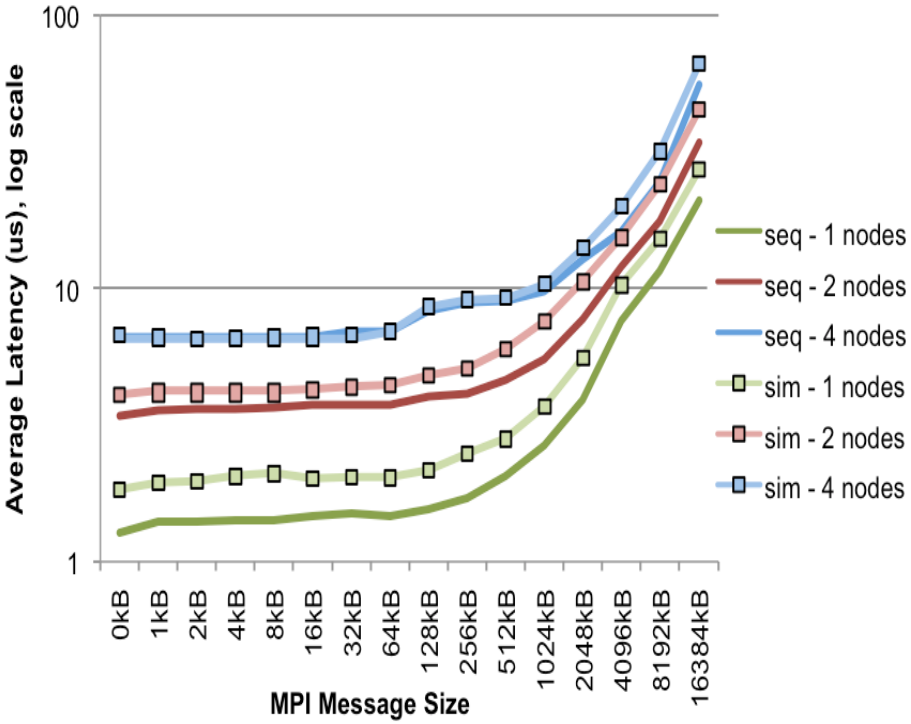
- NAMD, VMD
  - Study of the structure and function of biological molecules
- Calculation of non-bonded forces on GPUs leads to 9x speedup
- Framework hides most of the GPU complexity from users



J.C. Phillips and J.E. Stone, "Probing biomolecular machines with graphics processors," *Commun. ACM, 52(10):34-41, 2009.*

Scalable Heterogeneous Computing (SHOC) Benchmark Suite





A. Danalis, G. Marin, C. McCurdy, J. Meredith, P.C. Roth, K. Spafford, V. Tipparaju, and J.S. Vetter, "The Scalable HeterOgeneous Computing (SHOC) Benchmark Suite," in Third Workshop on General-Purpose Computation on Graphics Processors (GPGPU 2010)`. Pittsburgh, 2010

Paper also includes energy and CUDA comparisons.

HW Wish List

- **Programmable architectures**
  - Correct and efficient hardware support for apps
  - Tune-abled architectures
  - Compiler support
- **Chip IO, Chip IO, Chip IO**
- **Tight integration of heterogeneous cores**
  - Eliminate PCIe latency wall
- **Better understanding of resiliency**
  - Revisiting precision has provided some upside
- **Shared resources/memory between heterogeneous cores – maybe**
  - Leads to complexity and unpredictability

THANKS!