Toward Exascale Computational Science with Graphics Processors

Jeffrey Vetter
With contributions from many...

Presented to
SAAHPC, Knoxville
Overview

- Predictive scientific simulation is important for scientific discovery
- HPC systems have been highly successful
- The HPC community has several (new) constraints
- Heterogeneous computing with GPUs offers some opportunities and challenges
- Newly awarded NSF partnership will provide heterogeneous supercomputing for open science
- Exascale computing will require fundamental changes to our systems
Scientific Simulation is an accepted form of science

- Theory fails for
  - Complex phenomena

- Experiment fails for
  - Ethics
    - Testing on humans, animals
  - Experimentation limits
    - Climate modeling
    - Atomistic scale materials

Simulation is Tackling Compelling Problems

- **Energy for environmental sustainability**
  - Climate change: Carbon sequestration, weather event impacts on global climate, decadal climate predictive skill in aerosol forcing, global climate at unprecedented resolution, mantle convection
  - Energy storage: Charge storage and transfer in nano-structured supercapacitors
  - Combustion: Stabilizing diesel jet flames for increased efficiency and decreased emissions
  - Bioenergy: Recalcitrance in cellulosic ethanol
  - Solar: Nonequilibrium semiconductor alloys
  - Energy transmission: Role of inhomogeneities in high-temperature superconducting cuprates
  - Fusion: ITER design, optimization, and operation
  - Nuclear energy: Fully resolved reactor core neutron state
Tackling Compelling Problems (2)

- **Materials and nanoscience**
  > Structure of nanowires, nanoelectronics, nanorods, and strongly correlated materials (magnets)

- **Fundamental science**
  > Astrophysics: Decipher core-collapse supernovae and black hole mergers
  > Chemistry: Water structure in biological and aqueous-phase systems; water bonding on graphene and carbon nanotubes
  > Nuclear physics: Probe the anomalously long lifetime of carbon-14
  > High energy physics: Generate lattices at the physical values of the light quark (u, d, s) masses
  > Turbulence: Dispersion relative to air quality modeling and bioterrorism; physics of shock/turbulence interaction
  > Biology: Effect of an embolization on hemodynamics in the entire neurovasculature
Resulting in Highly Visible Science

**Physical Review Letters:**
High temperature superconductivity

**Combustion and Flame:**
3D flame simulation

**Nature:**
Astrophysics

**Physics of Plasmas:**
ICRF heating in ITER

Transformational science enabled by advanced scientific computing
Motivating Example: Climate Modeling

- Assess scientific basis, impact, adaptation, vulnerability, mitigation
  - Observation and simulation play critical roles

- Profound impact on worldwide socioeconomic policy, energy, etc
  - UN Conference - Sep 09
  - Copenhagen Climate conference – Dec 09

- Intergovernmental Panel On Climate Change
  - Sponsored by UNEP and WMO
  - Thousands of scientists from around the world

Computational science roadmap for a predictive, integrated climate model

**Climate Drivers:**
- Quantify predictive skill for hydrologic cycle on sub continental space scales
- Accurate and bounded representation of climate extremes
- Exploration of climate system sensitivities
- Exploration of climate system sensitivities
- Development of accurate system initialization techniques
- Direct incorporation of “human-dimensions” feedback mechanisms
- Global cloud-system resolving model formulations
- Non-hydrostatic adaptive formulations

**Science simulations:**
- Global Simulations with 1st Generation coupled carbon and nitrogen components
- Accurate treatment of 200 Km hydrological features
- Global simulations with full chemistry/biogeochemistry and fully coupled stratosphere
- Global simulations capable of accurate treatment of 100 Km hydrological features
- Atmospheric composition
- Prediction experiments on decadal time scales, 100 Km hydrological features
- Large-ensemble multi-decadal prediction experiments
- Global cloud-system resolving ensemble prediction experiments

Source: DOE Exascale Initiative
http://extremecomputing.labworks.org/
Today’s applications are quickly becoming more complex than earlier applications

- DOE Exascale Initiative [1]

Increased fidelity

Answering more complex questions

- Scientists [2,3] working on model that combines
  - Climate model
  - Energy economics
  - Population movements, demographics

Applications design and implementation are already complex; Writing and optimizing code for each new architecture and programming model is impractical (and only going to happen w/ heroic efforts/funding.)

HPC Systems Today

- **TOP500 #1: Jaguar @ ORNL**
  - Cray XT5
  - 2.3 Petaflops
  - 224k cores, AMD Istanbul
  - 300 TB Main memory
  - Seastar 2.1+
    - 9.6 GBps/link
  - Storage system
    - Lustre
    - 10 PB, 240 GBps
  - ~200 cabinets
  - 4,400 sq ft
  - 7 MWs during HPL run
  - Liquid cooled

http://nccs.gov
Typical Jaguar XT5 science workload

- Current: Transition to operations
  - 27 early-access projects, January–July
    - Total allocation: 540.8M hours
    - Usage since January: 247M hours (159 users)
- ALCC project: FY09 Joule metric applications
  - Total allocation: 25M hours
  - Usage since January: 17M hours (17 users)
- Fall 2009: General availability
  - 38 INCITE projects transitioned to system
    - Total allocation: 469M hours
    - Usage (XT4) since January: 91M hours (448 users)
- Discretionary projects: Climate AR5 production
  - Total allocation: 80M hours
  - 2 projects: NCAR/DOE and NOAA (~50 users)
<table>
<thead>
<tr>
<th>Science area</th>
<th>Code</th>
<th>Contact</th>
<th>Cores</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Materials</td>
<td>DCA++</td>
<td>Schulthess</td>
<td>150,144</td>
<td>1.3 PF MP</td>
</tr>
<tr>
<td>Materials</td>
<td>LSMS</td>
<td>Eisenbach</td>
<td>149,580</td>
<td>1.05 PF</td>
</tr>
<tr>
<td>Seismology</td>
<td>SPECFEM3D</td>
<td>Carrington</td>
<td>149,784</td>
<td>165 TF</td>
</tr>
<tr>
<td>Weather</td>
<td>WRF</td>
<td>Michalakes</td>
<td>150,000</td>
<td>50 TF</td>
</tr>
<tr>
<td>Climate</td>
<td>POP</td>
<td>Jones</td>
<td>18,000</td>
<td>20 simulation years/day</td>
</tr>
<tr>
<td>Combustion</td>
<td>S3D</td>
<td>Chen</td>
<td>144,000</td>
<td>83 TF</td>
</tr>
<tr>
<td>Fusion</td>
<td>GTC</td>
<td>PPPL</td>
<td>102,000</td>
<td>20 billion particles/second</td>
</tr>
<tr>
<td>Materials</td>
<td>LS3DF</td>
<td>Lin-Wang Wang</td>
<td>147,456</td>
<td>442 TF</td>
</tr>
<tr>
<td>Chemistry</td>
<td>NWChem</td>
<td>Apra</td>
<td>96,000</td>
<td>480 TF</td>
</tr>
<tr>
<td>Chemistry</td>
<td>MADNESS</td>
<td>Harrison</td>
<td>140,000</td>
<td>550+ TF</td>
</tr>
</tbody>
</table>

Science applications are scaling on Jaguar.
THE ROAD AHEAD
ORNL Roadmap to Exascale

Increasing computation requirements and resources

- 0.6 -> 1 PF Cray XT (NSF-1)
- 1 -> 2 PF Cray (LCF-2)
- 50 TF > 100 TF > 250 TF Cray XT4 (LCF-1)
- 170 TF Cray XT4 (NSF-0)
- 18.5 TF Cray X1E (LCF-0)
- 20 PF > 40 PF
- 100 PF > 250 PF
- 1 EF

Increasing power and facilities

ORNL Multi-Agency Computer Facility
260,000 ft²

ORNL Multipurpose Research Facility

ORNL Computational Sciences Building


13 July 2010
## Notional System Architecture Targets and “swim lanes”

<table>
<thead>
<tr>
<th>System attributes</th>
<th>2010</th>
<th>“2015”</th>
<th>“2018”</th>
</tr>
</thead>
<tbody>
<tr>
<td>System peak</td>
<td>2 Peta</td>
<td>200 Petaflop/sec</td>
<td>1 Exaflop/sec</td>
</tr>
<tr>
<td>Power</td>
<td>6 MW</td>
<td>15 MW</td>
<td>20 MW</td>
</tr>
<tr>
<td>System memory</td>
<td>0.3 PB</td>
<td>5 PB</td>
<td>32-64 PB</td>
</tr>
<tr>
<td>Node performance</td>
<td>125 GF</td>
<td>0.5 TF</td>
<td>7 TF</td>
</tr>
<tr>
<td>Node memory BW</td>
<td>25 GB/s</td>
<td>0.1 TB/sec</td>
<td>1 TB/sec</td>
</tr>
<tr>
<td>Node concurrency</td>
<td>12</td>
<td>O(100)</td>
<td>O(1,000)</td>
</tr>
<tr>
<td>System size (nodes)</td>
<td>18,700</td>
<td>50,000</td>
<td>5,000</td>
</tr>
<tr>
<td>Total Node Interconnect BW</td>
<td>1.5 GB/s</td>
<td>150 GB/sec</td>
<td>1 TB/sec</td>
</tr>
<tr>
<td>MTTI</td>
<td>day</td>
<td>O(1 day)</td>
<td>O(1 day)</td>
</tr>
</tbody>
</table>

DOE Exascale Initiative
Tough Decisions for Managing Power and Facilities Demands

- Build bigger buildings and plan to pay $$$ for ops
- Improve efficiencies
  - Power distribution
  - Workload scheduling
  - Software
- Use architectures that ‘match’ your workload
- Design new underlying technologies
  - Optical networks
  - 3D stacking
  - MRAM, PCM, nanotubes

Comparison of Projected Electricity Use, All Scenarios, 2007 to 2011.

Heterogeneous architectures can offer better performance, power.

- **Tukwila – First 2 billion transistor chip**
  - 80486 had ~1.2M transistors, ~50MHz, 1989
  - Specialization can be free

- *Industry has debated merits of each architecture for decades…*
- *Combination of all approaches optimizes power and performance*

---

### No single architecture solves all power problems

<table>
<thead>
<tr>
<th>METHOD</th>
<th>BITRATE (MBITS/S)</th>
<th>RELATIVE PERFORMANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel Nehalem</td>
<td>252914</td>
<td>-</td>
</tr>
<tr>
<td>Bitslice</td>
<td>63262</td>
<td>4.0</td>
</tr>
<tr>
<td>8-bit LUT</td>
<td>42834</td>
<td>5.9</td>
</tr>
<tr>
<td>Naïve</td>
<td>3368</td>
<td>75.1</td>
</tr>
</tbody>
</table>

*Source: Delagi, ISSCC 2010*
ENABLING HETEROGENEOUS COMPUTING FOR THE OPEN SCIENCE COMMUNITY
NSF Office of Cyber Infrastructure RFP

- NSF 08-573 OCI Track 2D RFP in Fall 2008
  - Data Intensive
  - Experimental Grid testbed
  - Pool of loosely coupled grid-computing resources
  - Experimental HPC System of Innovative Design

An experimental high-performance computing system of innovative design. Proposals are sought for the development and deployment of a system with an architectural design that is outside the mainstream of what is routinely available from computer vendors. Such a project may be for a duration of up to five years and for a total award size of up to $12,000,000. It is not necessary that the system be deployed early in the project; for example, a lengthy development phase might be included. Proposals should explain why such a resource will expand the range of research projects that scientists and engineers can tackle and include some examples of science and engineering questions to which the system will be applied. It is not necessary that the design of the proposed system be useful for all classes of computational science and engineering problems. When finally deployed, the system should be integrated into the TeraGrid. It is anticipated that the system, once deployed, will be an experimental TeraGrid resource, used by a smaller number of researchers than is typical for a large TeraGrid resource. (Up to 5 years duration. Up to $12,000,000 in total budget to include development and/or acquisition, operations and maintenance, including user support. First-year budget not to exceed $4,000,000.)
Oct 2008 Alternatives Analysis

- STI Cell (?)
- FGPAs
- Cyclops64 (?)
- Cray XMT
- Sun Rock/Niagara (?)
- ClearSpeed (?)
- Tensilica
- Tilera
- Anton
- SGI Molecule (?)
- Intel Larrabee (?)

- Graphics processors
- Others...

- Performance
  - Must show reasonable performance improvements at scale on real scientific applications of interest

- Programmability
  - Must be easy to re-port and re-optimize applications for each new architecture (generation) without large effort, delays

- Precision - Accuracy
  - Must provide impressive performance accurately

- Reliability
  - Must get high scientific throughput without job failures or inaccurate results

- Power and Facilities Cost
  - Must be reasonably affordable in terms of power and facilities costs
Alternatives analysis concluded GPUs were a competitive solution

- Success with various applications at DOE, NSF, government, industry
  - Signal processing, image processing, etc.
  - DCA++, S3D, NAMD, many others
- Community application experiences also positive
  - Frequent workshops, tutorials, software development, university classes
  - Many apps teams are excited about using GPGPUs
Batman: Arkham Asylum with PhysX!

- Smoke reacts with Batman, provides cover
- Walls explode
- Glass shatters
- Tattered curtains react with characters

Highest rated PC game since 2007!

Courtesy of NVIDIA
GPU Rationale – What’s different now?

Heterogeneous Computing with Graphics Processors

- Very High Memory Bandwidth
- High SP Flop Rate
- High Flop per Watt
- Productivity CUDA OpenCL
- Reliability at Scale
- High DP Flop Rate

Leverage commodity

Fermi
And Power Efficiency!

Source: http://www.realworldtech.com
## Keeneland Partners

<table>
<thead>
<tr>
<th>Georgia Institute of Technology</th>
<th>National Institute of Computational Sciences</th>
<th>Oak Ridge National Laboratory</th>
<th>University of Tennessee, Knoxville</th>
<th>NVIDIA</th>
<th>HP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Project management</strong></td>
<td><strong>Operations and TG/XD Integration</strong></td>
<td><strong>Applications</strong></td>
<td><strong>Scientific Libraries</strong></td>
<td><strong>Tesla</strong></td>
<td><strong>HPC Host System</strong></td>
</tr>
<tr>
<td><strong>Acquisition and alternatives assessment</strong></td>
<td><strong>User and Application Support</strong></td>
<td><strong>Facilities</strong></td>
<td><strong>Education, Outreach, Training</strong></td>
<td><strong>Applications optimizations</strong></td>
<td><strong>Training</strong></td>
</tr>
<tr>
<td><strong>System software and development tools</strong></td>
<td><strong>Operational Infrastructure</strong></td>
<td><strong>Education, Outreach, Training</strong></td>
<td><strong>Education, Outreach, Training</strong></td>
<td><strong>Training</strong></td>
<td><strong>System integration</strong></td>
</tr>
<tr>
<td><strong>Education, Outreach, Training</strong></td>
<td><strong>Education, Outreach, Training</strong></td>
<td><strong>Education, Outreach, Training</strong></td>
<td><strong>Training</strong></td>
<td><strong>Training</strong></td>
<td><strong>Training</strong></td>
</tr>
</tbody>
</table>
Keeneland Team

- Jeffrey Vetter, Dick Glassbrook, Jack Dongarra, Richard Fujimoto, Thomas Schulthess, Karsten Schwan, Sudha Yalamanchili, Kathlyn Boudwin, Jim Ferguson, Doug Hudson, Patricia Kovatch, Bruce Loftis, Jeremy Meredith, Jim Rogers, Philip Roth, Arlene Washington, Phil Andrews, Mark Fahey, Don Reed, Tracy Rafferty, Ursula Henderson, Terry Moore, Kyle, Spafford, and many others

- NVIDIA
- HP

- Keeneland Sponsor: NSF
- Other sponsors: DOE, DOD, DARPA
Keeneland Initial Delivery (KID) System

• KID Architecture
  – HP Linux Cluster
  – Fermi 6 GB
  – New integrated CPU/GPU design
  – Details to be announced soon
  – ~200 CPUs + ~300 GPUs

• Traditional Linux software stack augmented with GPU compilers, software tools, libraries

• Delivery and acceptance in Sept 2010
KID will use NVIDIA’s Fermi

- 3B transistors
- ECC
- 8x the peak double precision arithmetic performance over NVIDIA's last generation GPU.
- 448 CUDA Cores featuring the new IEEE 754-2008 floating-point standard
- NVIDIA Parallel DataCache
- NVIDIA GigaThread Engine
- Debuggers, language support
Many GPU-enabled systems blooming everywhere ....

China’s Defense University builds World Third fastest supercomputer

October 29 (China Military News Reporting by Johnathan Weng) — The Chinese National University of Defense Technology (NUDT) unveiled Thursday China’s fastest supercomputer, also the World Third fastest computer, which is able to do more than one quadrillion calculations per second theoretically at its peak.

1. Oak Ridge LCF3
2. More announcements expected in coming months ...

Tokyo Institute of Technology announces SSD-pack 2.39 petaflop supercomputer

By Joseph L. Flatley posted Jun 23rd 2010 2:06PM

China announces that “Tianhe-II” LINPACK score is 563.1 teraFLOP (1.0 petaFLOPS) of Jilin Research Institute.

IBM has announced plans to start using SandForce SSDs in its enterprise machines, and now it looks like the Tokyo Institute of Technology is doing one better, working with NEC and HP to produce Tybanne 2.0. This next-gen supercomputer will reportedly operate at 2.39 petaflops (that’s a lot of flops!) and uses a new multilevel storage architecture consisting of DRAM as well as SSDs. Not only will this bad boy have thirty times the computing capacity of Tybanne 1.0, but in part to its some 2,846 Intel Westmere microprocessors and 4,224 NVIDIA Tesla Mosaic GPUs, its power draw should be some 1/29th of its predecessor’s. If all goes according to plan, it should be in operation this fall, at a cost of ¥3.8 billion (approx $53.5 million).

Development and application of a HPC system for multi-scale discrete simulation—Mole-8.5

Xiaowei Weng, Wei Ge, Xianfeng He, FeiGuo Chen, Li Guo, Jinghai Li
Institute of Process Engineering, Chinese Academy of Sciences, Beijing, 100190

Mole-8.5 is the first GPGPU supercomputer (Rpeak of about 1100 Tflops) using NVIDIA Tesla C2050 in the world, which includes 372 nodes and is established in April 2010. It is the successor of the first supercomputer with 1.0 Petaflops peak performance in single precision in China, which was a hybrid system including four units integrating NVIDIA and AMD GPUs announced on April 20, 2009. Mole-8.5 was designed and established by Institute of Process Engineering (IPE), Chinese Academy of Sciences, one of the NVIDIA CCOEs. A designing philosophy utilizing the similarity between hardware, software and the problems to be solved is embodied, based on the multi-scale method and discrete simulation approaches developed at IPE. The whole system is connected with Gigabit Ethernet and QDR Infiniband network. Mole-8.5 has some unique advantages over the HPC system of same performance based on CPU, for example the high performance/price ratio, the area occupied by it is only about 150 M2. The impact result of 320 nodes of Mole-8.5 is 2.07×10^5 Gflops with a power consumption of about 480 kWatt, therefore the average power efficiency is 431 Mflops/Watt, manifesting an energy efficient supercomputer.
APPLICATIONS
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 on Tesla T10


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 on Tesla T10
  - 31.4x SP speedup
  - 16.2x DP speedup

Biomolecular systems from NAMD Team – Not just us

• NAMD, VMD
  – Study of the structure and function of biological molecules
• Calculation of non-bonded forces on GPUs leads to 9x speedup on Tesla T10
• Framework hides most of the GPU complexity from users

Many others ... Including Skeptics?

Debunking the 100X GPU vs. CPU Myth: An Evaluation of Throughput Computing on CPU and GPU

Victor W Lee†, Changkyu Kim†, Jatin Chhugani†, Michael Deisher†, Daehyun Kim†, Anthony D. Nguyen†, Nadathur Satish†, Mikhail Smelyanskiy†, Srinivas Chennupaty†, Per Hammarlund†, Ronak Singhal† and Pradeep Dubey†

victor.w.lee@intel.com

†Throughput Computing Lab, Intel Corporation
‡Intel Architecture Group, Intel Corporation

ABSTRACT
Recent advances in computing have led to an explosion in the amount of data being generated. Processing the ever-growing data in a timely manner has made throughput computing an important aspect for emerging applications. Our analysis of a set of important throughput computing kernels shows that there is an ample amount of parallelism in these kernels which makes them suitable for today’s multi-core CPUs and GPUs. In the past few years there have been many studies claiming GPUs deliver substantial speedups (between 10X and 1000X) over multi-core CPUs on these kernels. To understand where such large performance difference comes from, we perform a rigorous performance analysis and find that after applying optimizations appropriate for both CPUs and GPUs the performance gap between an Nvidia GTX280 processor and the Intel Core i7 960 processor narrows to only 2.5x on average. In this pa-

![Graph](image)

(a) Relative Performance

Figure 1: Comparison between Core i7 and GTX280 Performance.
Keeneland Software Environment

• Integrated with NSF TeraGrid/XD
  – Including TG and NICS software stack

• Programming environments
  – CUDA
  – OpenCL
  – Compilers
    • PGI
      – Accelerate, CUDA Fortran
    • OpenMP 3.0
  – Scalable debuggers
  – Performance tools

• Additional software activities
  – Benchmarks
  – Performance and correctness tools
  – Scientific libraries
  – Virtualization
Ocelot: Dynamic Execution Infrastructure

NVIDIA Virtual ISA

Use as a basis for

- Insight → workload characterization
- Performance tuning → detecting memory bank conflicts
- Debugging → illegal memory accesses, out of bounds checks, etc.

PTX 1.4 compliant Emulation
• Validated on full CUDA SDK
• Open Source version released

http://code.google.com/p/gpuocelot/

Gregory Diamos, Dhuv Choudhary, Andrew Kerr, Sudhakar Yalamanchili
Workload Analysis: Examples

Branch Divergence
- Study of control Flow behavior
- Motivate synchronization support

Inter-thread Data Flow
- Study of data sharing patterns
- Motivate architectural support

Gregory Diamos, Dhuv Choudhary, Andrew Kerr, Sudhakar Yalamanchili
Libraries: One and two-sided Multicore+GPU Factorizations

- These will be included in up-coming MAGMA releases
- Two-sided factorizations can not be efficiently accelerated on homogeneous x86-based multicores (above) because of memory-bound operations
  - MAGMA provided hybrid algorithms that overcome those bottlenecks (16x speedup!)

Multicore + GPU Performance in double precision

LU Factorization

Hessenberg Factorization

GPU: NVIDIA GeForce GTX 280
CPU: Intel Xeon dual socket quad-core @2.33 GHz

GPU BLAS: CUBLAS 2.2, dgemm peak: 75 GFlop/s
CPU BLAS: MKL 10.0, dgemm peak: 65 GFlop/s
Hybrid Virtual Machine: HyVM

- Uniform runtime model for heterogeneous platforms:
- Hybrid Virtual Machines
  - Uniformity:
    - Virtual Execution Unit (VEU): hypervisor-level, uniform runtime representation for program executables
    - Heterogeneity-aware hypervisors: VMM-level management methods for improved platform utilization (incl. cache and energy) and application performance (SLAs)
    - Dynamic platform emulation: runtime CK compilation or re-writing for diverse accelerator targets (via LLVM)
  - High performance:
    - Commodity and custom VEU ‘containers’: Virtual Machines (VMs) – processes/threads – commodity cores; Special Execution Environments (e.g., NVIDIA) - Computational Kernels (CKs) - accelerators
    - Runtime and adaptive {CK} optimization for parallelism
    - Standards-compliant CK programming and runtime APIs (OpenCL, CUDA)
    - Compiler-based optimization techniques for {CK}

Ada Gavrilovska, Karsten Schwan, Sudha Yalamanchili, and many PhD students
The Scalable HeterOgeneous Computing (SHOC) Benchmark Suite

- Benchmark suite with a focus on scientific computing workloads, including common kernels like SGEMM, FFT, Stencils
- Parallelized with MPI, with support for multi-GPU and cluster scale comparisons
- Implemented in CUDA and OpenCL for a 1:1 performance comparison
- Includes stability tests

<table>
<thead>
<tr>
<th>Level 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>BusSpeedDownload: measures bandwidth of transferring data across the PCIe bus to a device.</td>
</tr>
<tr>
<td>BusSpeedReadback: measures bandwidth of reading data back from a device.</td>
</tr>
<tr>
<td>DeviceMemory: measures bandwidth of memory accesses to various types of device memory including global, local, and image memories.</td>
</tr>
<tr>
<td>KernelCompile: measures compile time for several OpenCL kernels, which range in complexity</td>
</tr>
<tr>
<td>PeakFlops: measures maximum achievable floating point performance using a combination of auto-generated and hand coded kernels.</td>
</tr>
<tr>
<td>QueueDelay: measures the overhead of using the OpenCL command queue.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFT: forward and reverse 1D FFT.</td>
</tr>
<tr>
<td>MD: computation of the Lennard-Jones potential from molecular dynamics, a specific case of the nbody problem.</td>
</tr>
<tr>
<td>Reduction: reduction operation on an array of single precision floating point values.</td>
</tr>
<tr>
<td>SGEMM: single-precision matrix-matrix multiply.</td>
</tr>
<tr>
<td>Scan: scan (also known as parallel prefix sum) on an array of single precision floating point values.</td>
</tr>
<tr>
<td>Sort: sorts an array of key-value pairs using a radix sort algorithm</td>
</tr>
<tr>
<td>Stencil2D: a 9-point stencil operation applied to a 2D data set. In the MPI version, data is distributed across MPI processes organized in a 2D Cartesian topology, with periodic halo exchanges.</td>
</tr>
<tr>
<td>Triad: STREAM Triad operations, implemented in OpenCL.</td>
</tr>
</tbody>
</table>


Paper also includes energy and CUDA comparisons. Beta software available at [http://ft.ornl.gov/doku/shoc/start](http://ft.ornl.gov/doku/shoc/start)
SHOC Example #1: Longitudinal OpenCL Performance

- Single precision, Tesla C1060 GPU
- Comparing NVIDIA OpenCL implementation from 2.3 and 3.0 GPU Computing SDK
SHOC Example #2: Compare OpenCL and CUDA

- OpenCL improving, but still trailing CUDA
- Tesla C1060, Single Precision, CUDA and OpenCL 3.0
- FFT/MD/SGEMM – GFLOPS, Reduction/Scan – GB/s
SHOC Example #3: Compare Different GPUs

- Single Precision
- ECC On (for Tesla C2050)
- Radeon HD 5870: AMD OpenCL v2.1
- Tesla C2050 CUDA 3.1b
- Others CUDA 3.0

### FFT

<table>
<thead>
<tr>
<th>GPU</th>
<th>GFLOPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATI Radeon HD 5870</td>
<td>39.8</td>
</tr>
<tr>
<td>Tesla C2050</td>
<td>261.0</td>
</tr>
<tr>
<td>Tesla C1060</td>
<td>165.9</td>
</tr>
<tr>
<td>GeForce 8800GTX</td>
<td>130.6</td>
</tr>
<tr>
<td>NVIDIA ION</td>
<td>12.6</td>
</tr>
</tbody>
</table>

### MD

<table>
<thead>
<tr>
<th>GPU</th>
<th>GFLOPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATI Radeon HD 5870</td>
<td>340.4</td>
</tr>
<tr>
<td>Tesla C2050</td>
<td>468.5</td>
</tr>
<tr>
<td>Tesla C1060</td>
<td>292.9</td>
</tr>
<tr>
<td>GeForce 8800GTX</td>
<td>201.3</td>
</tr>
<tr>
<td>NVIDIA ION</td>
<td>18.9</td>
</tr>
</tbody>
</table>
SHOC Example #4 – Energy Efficiency

- Single precision, calculated using vendor’s TDP – Ion very efficient for bandwidth bound problems
SHOC 0.9 Beta Release

• Source code available: http://ft.ornl.gov/doku/shoc/downloads

• New Application-Level Benchmark S3D – combustion kernel

• Numerous bug fixes

• Coming Soon:
  – Sparse Matrix-Vector Multiply
  – More Double Precision Tests
Recap

- Predictive scientific simulation is important for scientific discovery
  - Advancing science, informing policymakers
- HPC systems have been highly successful
  - Decades of improvement
- The HPC community has several (new) constraints
  - Power, Facilities, Cost
- Heterogeneous computing with GPUs offers some opportunities and challenges
  - High performance; good performance per watt
  - Programmability; limited applicability
- Keeneland - Newly awarded NSF partnership will provide heterogeneous supercomputing for open science

We are hiring Research Scientists, Postdocs, Students

Develop software for advanced heterogeneous architectures
Scientific libraries
Software tools
Runtime systems
Assess new architectures
Identify and refactor promising applications for Keeneland platforms

http://keeneland.gatech.edu
http://ft.ornl.gov
Thank You!

Thanks to contributors, sponsors
Many collaborators across apps
teams, academia, labs, and industry
DOE, NSF, ORNL, DOD, DARPA

More information
http://ft.ornl.gov
vetter@computer.org
Publications: http://ft.ornl.gov/pubs

http://keeneland.gatech.edu
http://www.cse.gatech.edu
http://www.cercs.gatech.edu
http://icl.cs.utk.edu
http://www.nics.tennessee.edu/
http://ft.ornl.gov