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

KEENELAND - ENABLING HETEROGENEOUS COMPUTING FOR THE OPEN SCIENCE COMMUNITY
BACKGROUND – HOW DID WE GET HERE?
Oct 2008 alternatives analysis for NSF OCI RFP 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

• Programmability, Resilience?
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
### 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**
Exascale computing will require tough decisions and/or innovative technologies

- Build bigger buildings and plan to pay $$$ for ops
- Improve efficiencies
  - PUE
  - Power distribution
  - Workload scheduling
  - Software
- Use architectures that ‘match’ your workload
  - GPUs, FPGAs
- Design new underlying technologies
  - Optical networks
  - 3D stacking
  - MRAM, PCM, R-RAM
Heterogeneous architectures can offer better performance, power

No single architecture solves all power problems

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

Source: Delagi, ISSCC 2010
KIID ARCHITECTURE
Keeneland – Initial Delivery System Architecture

**Initial Delivery** system procured and installed in Oct 2010
- 201 TFLOPS in 7 racks (90 sq ft incl service area)
- 677 MFLOPS per watt on HPL

Final delivery system expected in early 2012

---

Keeneland System (7 Racks)

- **ProLiant SL390s G7** (2CPUs, 3GPUs)
- **S6500 Chassis** (4 Nodes)
- **Rack** (6 Chassis)

- **M2070**
- **Xeon 5660**

- **67 GFLOPS**
- **515 GFLOPS**
- **1679 GFLOPS**
- **6718 GFLOPS**
- **40306 GFLOPS**
- **201528 GFLOPS**

- Full PCIe X16 bandwidth to all GPUs

- **Integrated with NICS Datacenter GPFS and TG**

---

Images and logos from various technology companies.
NVIDIA Fermi

- 3B transistors!
- Error correction
- 448 CUDA Cores featuring the new IEEE 754-2008 floating-point standard
  - 8× the peak double precision arithmetic performance over NVIDIA's last generation GPU
  - 515 DP GF
  - 1030 SP GF
  - 32 cores per SM, 21k threads per chip
- 120-144 GB/s memory BW
- NVIDIA Parallel DataCache
- NVIDIA GigaThread Engine
- Debuggers, language support
HP ProLiant SL390s G7 2U half width tray

1 GPU module in the rear, lower 1U

2 Non-hot plug SFF (2.5") HDD

4 Hot plug SFF (2.5") HDDs

2 GPU modules in upper 1U

Dual 1GbE

Dedicated management iLO3 LAN & 2 USB ports

UID LED & Button

Health LED

VGA

Serial (RJ45)

Power Button

SFP+

QSFP (QDR IB)

Dual 1GbE
Keeneland Node Architecture SL390

- RAM
- DDR3
- CPU
- QPI
- I/O Hub
- PCIe x8
- Infiniband
- integrated
- PCIe x16
- GPU (6GB)
- DDR3
- CPU
- QPI
- I/O Hub
- PCIe x16
- GPU (6GB)
- QPI
- PCIe x16
- GPU (6GB)
New ProLiant SL6500 series

Highly Flexible s6500 Chassis

Multinode, Shared Power and Cooling Architecture

Benefits: Low Cost, High Efficiency Chassis

- 4U chassis for deployment flexibility
- Standard 19” racks, with front I/O cabling
- Unrestricted airflow (no mid-plane or I/O connectors)
- Reduced weight
  - Individually serviceable nodes
  - Variety of optimized node modules
- SL Advanced Power Manager support
  - Power monitoring
  - Node level power off/on

- Shared power and fans
- Optional hot-plug redundant PSU
- Energy efficient hot-plug fans
- 3-phase load balancing
- 94% platinum common slot power supplies
- N+1 capable power supplies (up to 4)
KID Installation

• From the dock to functioning system in 7 days!
  – HP Factory integration and testing prior to delivery contributed to quick uptime
• System delivered on Oct 27
• Installation completed on Oct 29
• Top500, Green500 results completed on Nov 1
• Finishing acceptance testing this week
Keeneland ID installation – 10/29/10
Installation Team has worked long hours

Clockwise from upper right: Stephen McNally, Kyle Spafford, Philip Roth, Jeremy Meredith, Dave Holton (HP), Jeffrey Vetter, Dale Southard (NVIDIA).

Thanks to many at HP, NVIDIA, Qlogic: Paul Salerno, Glen Lupton, etc.
Keeneland Partners

Georgia Institute of Technology
- Project management
- Acquisition and alternatives assessment
- System software and development tools
- Education, Outreach, Training

National Institute of Computational Sciences
- Operations and TG/XD Integration
- User and Application Support
- Operational Infrastructure
- Education, Outreach, Training

Oak Ridge National Laboratory
- Applications
- Facilities
- Education, Outreach, Training

University of Tennessee, Knoxville
- Scientific Libraries
- Education, Outreach, Training

NVIDIA
- Tesla
- Applications optimizations
- Training

HP
- HPC Host System
- System integration
- Training
Status

- Finish acceptance testing on KID
- Enter early science operation
- KID goals
  - Connected to TG/XD
  - Resource for applications teams with GPU codes
  - Resource for GPU software and tool development
- Larger, final delivery system planned for mid 2012
APPLICATIONS
Early Success Stories

Computational Materials

• Quantum Monte Carlo
  – 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
• Full parallel app is 5x faster, start to finish, on a GPU-enabled cluster on Tesla T10

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


Simulating Blood Flow with FMM

- Multiphysics, multiphysics particle flow of deformable cells in viscous fluid with non-uniform distribution

Preliminary results from KID

FMM results from KID

Graph showing performance metrics:
- **GFLOP/s** against **Spherical Harmonics’ Order**
- **Seconds/1M points** against **Number of MPI Processes**

Key metrics:
- **Nehalem**, **Tesla**, **Fermi**, **Fermi (Streams)**
- **1M Points**, **2M Points (NUMA control)**, **3M Points (NUMA control)**
NAMD

- NAMD, VMD
  - Study of the structure and function of biological molecules
- Calculation of non-bonded forces on GPUs leads to 6x on FERMI
- Framework hides most of the GPU complexity from users


Preliminary results from KID
GROMACS

- GROMACS (GROningen MAchine for Chemical Simulations) is a molecular dynamics simulation package

Preliminary results from KID

<table>
<thead>
<tr>
<th></th>
<th>ns/day on KID</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU 1 thread</td>
<td>0</td>
</tr>
<tr>
<td>CPU 6 threads</td>
<td>5</td>
</tr>
<tr>
<td>GPU</td>
<td>120</td>
</tr>
</tbody>
</table>
AMBER on FERMI (courtesy R. Walker, D. Poole et al.)
SOFTWARE
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

PTX Kernel

PTX Emulation

GPU Execution

LLVM Translation

x86

NVIDIA GPU

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

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

Use as a basis for
- Insight \(\rightarrow\) workload characterization
- Performance tuning \(\rightarrow\) detecting memory bank conflicts
- Debugging \(\rightarrow\) illegal memory accesses, out of bounds checks, etc.

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

<table>
<thead>
<tr>
<th>Matrix size x 1000</th>
<th>LU Factorization</th>
<th>Hessenberg Factorization</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>120</td>
<td>120</td>
</tr>
<tr>
<td>7</td>
<td>140</td>
<td>140</td>
</tr>
<tr>
<td>8</td>
<td>160</td>
<td>160</td>
</tr>
<tr>
<td>9</td>
<td>180</td>
<td>180</td>
</tr>
<tr>
<td>10</td>
<td>200</td>
<td>200</td>
</tr>
</tbody>
</table>

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

Jack Dongarra, Stan Tomov, and Rajib Nath
DOE Vancouver: A Software Stack for Productive Heterogeneous Exascale Computing

**Objectives**

- Enhance programmer productivity for the exascale
  - Increase code development ROI by enhancing code portability
  - Decrease barriers to entry with new programming models
- Create next-generation tools to understand the performance behavior of an exascale machine

**Approach**

- Programming tools
  - GAS programming model
  - Analysis, inspection, transformation
- Software libraries: autotuning
- Runtime systems: scheduling
- Performance tools
- Impact on DOE Applications

The proposed Maestro runtime simplifies programming heterogeneous systems by unifying OpenCL task queues into a single high-level queue.

**Impact**

- Reduced application development time
- Ease of porting applications to heterogeneous systems
- Increased utilization of hardware resources and code portability
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


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**

- ATI Radeon HD 5870: 39.8 GFLOPS
- Tesla C2050: 261.0 GFLOPS
- Tesla C1060: 165.9 GFLOPS
- GeForce 8800GTX: 130.6 GFLOPS
- NVIDIA ION: 12.6 GFLOPS

**MD**

- ATI Radeon HD 5870: 340.4 GFLOPS
- Tesla C2050: 468.5 GFLOPS
- Tesla C1060: 292.9 GFLOPS
- GeForce 8800GTX: 201.3 GFLOPS
- NVIDIA ION: 18.9 GFLOPS
Longitudinal OpenCL Performance

- Single precision, Tesla C1060 GPU
- Comparing NVIDIA OpenCL implementation from 2.3 and 3.0 GPU Computing SDK
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
Energy Efficiency

- Single precision, calculated using vendor’s TDP – Ion very efficient for bandwidth bound problems
FUTURE SYSTEMS
Echelon: Extreme-scale Compute Hierarchies with Efficient Locality-Optimized Nodes

Main Objectives

- Two orders of magnitude increase in application execution energy efficiency over today's CPU systems.
- Improve programmer productivity so that the time required to write a parallel program achieving a large fraction of peak efficiency is comparable to the time required to write a serial program today.
- Strong scaling for many applications to tens of millions of threads in UHPC system.
- High application mean-time to interrupt (AMTTI) with low overhead; matched to application needs.
- Machines resilient to attack; enables reliable software.

Key Innovations

- Programming systems that express concurrency/locality abstractly; autotuning for hardware mapping.
- Self-aware runtime reacts to changes in environment, workload (load-balance), fault states.
- Fine-grained, energy-optimized, multithreaded throughput cores + latency-optimized cores.
- Software selective memory hierarchy configuration; selective coherence for non-critical data.
- HW/SW cooperative resilience for energy- and performance-efficient fault protection.
- Guarded pointers for memory safety.
- Low-power, high speed communication circuits.

Echelon Execution Model

- Programmability: global address space, abstract memory hierarchy, autotuning; runtime task placement/scheduling.
- Efficiency: active messages, bulk transfer.
- Dependability: software selective redundancy, hardware accelerated guarded pointers.

Echelon System Diagram

Georgia Tech, Stanford, UC-Berkeley, UPenn, UT-Austin, U. Utah, Tennessee, Lockheed Martin
Recap

• The HPC community has several (new) constraints
  – Power, Performance, Facilities, Cost
• Emerging technologies will play a critical role
• Heterogeneous computing with GPUs offers some opportunities and challenges
  – High performance; good performance per watt
  – Programmability; limited applicability
• KID is up and running
• Keeneland Final System coming in 2012

For more information:
vetter@computer.org
http://keeneland.gatech.edu
http://ft.ornl.gov
http://www.cse.gatech.edu
http://www.cercs.gatech.edu
http://icl.cs.utk.edu
http://www.nics.tennessee.edu/