Genomics Moves Us into Supercomputing
Learn to grow your code

Fernanda Foertter
Oak Ridge Leadership Computing Facility (OLCF)
Presenting at PAGXXII Jan 2014

What is the Leadership Computing Facility?
Collaborative, multi-lab, DOE/SC initiative ranked top domestic priority in Facilities for the Future of Science: A Twenty-Year Outlook.

Mission: Provide the computational and data science resources required to solve the most important scientific & engineering problems in the world.

- Highly competitive user allocation program (INCITE, ALCC)
- Projects receive 100x more hours than at other generally available centers.
- LCF centers partner with users to enable science & engineering breakthroughs

Memory capacity: 5K – 10K (1024 x 40 bit words, later doubled to 2048)
In 20 minutes, ORACLE could work problems that would take two mathematicians with calculators three years.
In 1953, ORACLE was the fastest computer, with the biggest memory, in the world.

ORACLE (1953)
Oak Ridge Automatic Computer and Logical Engine

Distributed Computing

What is a supercomputer?

A tightly integrated machine composed of several processors that is capable of working on a single problem that would otherwise be impossible single machine.

27 Petaflops
World’s Most Powerful Computer
27 Petaflops
Theoretical Peak Performance

USA's
#2
World’s Most Powerful Computer

increased our system capability by 10,000X
since 2004

2004 was a very big year...

Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

per core Performance

Kogge and Shalf, IEEE CISE
Moore's Law continues, while CPU clock rates stopped increasing in 2003 due to power constraints.

Power is capped by heat dissipation and $$$

Performance increases have been coming through increased parallelism

**Power is the problem**

- Power consumption of 2.3 PF Jaguar 7 megawatts
  - equivalent to a small city (~7,000 homes)

Power consumption of a 27 PF CPU-only system 60-80 megawatts
  - equivalent to ~70,000 homes

Power consumption of a 27 PF Hybrid system 8.2 megawatts
  - equivalent to ~8,000 homes

**ACCELERATORS**
Why GPUs?
High performance and power efficiency on path to exascale

Optimized for multitasking
Optimized for throughput

10x performance per socket
10x the energy-efficiency

ORNL's "Titan" Hybrid System

SYSTEM SPECIFICATIONS:
- Peak performance of 27.1 PF
- 24 x GPU = 2.8 CPU
- 18,600 Compute Nodes each with:
  - 16 Core AMD Opteron® CPU
  - NVIDIA Tesla K20x GPU
  - 32 x 6 GB memory
- 912 Service and I/O nodes
- 320 Cabinets
- 710 TB total system memory
- Chp w/GMII 10 Teras in/terconnct
- 6.9 GW peak power

4,352 ft²
(404 m²)

Titan Nodes

<table>
<thead>
<tr>
<th></th>
<th>Node</th>
<th>3.2 GHz, 32 GB (DDR3)</th>
<th>CPU</th>
<th>732 MHz, 6 GB (DDR5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMD Opteron®</td>
<td>AMD Opteron®</td>
<td>3.2 GHz, 32 GB (DDR3)</td>
<td>732 MHz, 6 GB (DDR5)</td>
<td></td>
</tr>
<tr>
<td>6200 Interlagos</td>
<td>6200 Interlagos</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(16 cores)</td>
<td>(2688 CUDA cores)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Network</td>
<td>Gemini High Speed Interconnect</td>
<td>3D Teras</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Storage</td>
<td>Lustre Filesystem</td>
<td>40 PB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Archive</td>
<td>High-Performance Storage System (HPSS)</td>
<td>29 PB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Center for Accelerated Application Readiness (CAAR)

Material Science (WL-LSMS)
- Illuminating the role of material disorder, statistics, and fluctuations in nanoscale materials and systems.

Molecular (LAMMPS)
- A molecular description of soft materials, with applications in biotechnology, medicine and energy.

Combustion (S3D)
- Understanding turbulent combustion through direct numerical simulation with complex chemistry.

Climate Change (CAM-SE)
- Answering questions about specific climate change adaptation and mitigation scenarios; realistic represent features like precipitation patterns / statistics and tropical deserts.

Astrophysics (NRDF)
- Radiation transport – important in astrophysics, laser fusion, combustion, atmospheric dynamics, and medical imaging – computed on AMR grids.

Nuclear Energy (Denver)
- Discrete ordinates radiation transport calculations that can be used in a variety of nuclear energy and technology applications.

lessons learned
Action plan for code porting
We developed a plan for porting these applications, which involved the following steps:

1. Multidisciplinary code team for each code – OLCF application lead, Cray engineer, NVIDIA developer, also cross-cutting support from tool and library developers.
2. Early testing hardware - while use GPU cluster “pilot” for code development.
3. Code inventory for each code to understand characteristics, application code structure, code scalability for GPU port, algorithmic structure, data structures and data movement patterns. Also code execution profile – are there performance “hot spots” or is the profile flat?
4. Develop parallelization approach for each application – ascertain which algorithms and code components to port to GPU, how to map work to GPU threads, how to manage data motion CPU-GPU and between GPU main memory and GPU cached/shared memory.
5. Develop GPU programming model for port to GPU, e.g., CUDA for more coarse-to-the-metal programming, OpenACC for a higher abstraction level and a more incremental porting approach, OpenCL, for portability advantages, or licenses when appropriate.
6. Address code development issues – rework vs. refactoring, managing portability in other platforms, incorporating GPU code into build system, relationship to the code repository main trunk.
7. Representative test cases, e.g., early science problems, formulated as tasks for evaluating code performance and setting priorities for code optimization, also formulate comparison metrics to measure success, e.g., time to completion, comparison X86 vs. Titan Cray XE Interlagos/Kaifep.

Application characteristics inventory

<table>
<thead>
<tr>
<th>App</th>
<th>Science Area</th>
<th>Algorithm(s)</th>
<th>Grid type</th>
<th>Programming Environment</th>
<th>Compiler(s) supported</th>
<th>LOC</th>
<th>Math Libraries</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAM-SE</td>
<td>climate</td>
<td>CAM, climate</td>
<td>3D</td>
<td>Fortran, Python, MPI</td>
<td>IBM, Intel</td>
<td>500K</td>
<td>Trilinos</td>
</tr>
<tr>
<td>LANL-PSE</td>
<td>Biology, materials</td>
<td>LANL-PSE</td>
<td>3D</td>
<td>Fortran, Python, MPI</td>
<td>IBM, Intel</td>
<td>140K</td>
<td>PETSc, Trilinos</td>
</tr>
<tr>
<td>S3D</td>
<td>combustion</td>
<td>S3D, combustion</td>
<td>3D</td>
<td>Structured, MPI</td>
<td>IBM, Intel</td>
<td>10K</td>
<td>None</td>
</tr>
<tr>
<td>Dense</td>
<td>nuclear energy</td>
<td>Dense, nuclear energy</td>
<td>3D</td>
<td>Structured, MPI</td>
<td>IBM, Intel</td>
<td>46K</td>
<td>None</td>
</tr>
<tr>
<td>WL-LSIM</td>
<td>nanoscience</td>
<td>WL-LSIM</td>
<td>3D</td>
<td>Structured, MPI</td>
<td>IBM, Intel</td>
<td>70K</td>
<td>None</td>
</tr>
<tr>
<td>NIMF</td>
<td>radiation</td>
<td>NIMF, radiation</td>
<td>AMR</td>
<td>Structured, MPI</td>
<td>IBM, Intel</td>
<td>500K</td>
<td>PETSc, Trilinos</td>
</tr>
</tbody>
</table>

Hybrid Programming Model

- On Jaguar, with 299,008 cores, we were seeing the limits of a single level of MPI scaling for most applications.
- To take advantage of the vastly larger parallelism in Titan, users need to use hierarchical parallelism in their codes.
  - Distributed memory: MPI, SHMEM, PGAS
  - Node Local: OpenMP, Pthreads, local MPI communicators
  - Within threads: Vector constructs on GPU, libraries, OpenACC
- These are the same types of constructs needed on all multi-PFLOPS computers to scale to the full size of the systems!
CAAR: Selected Lessons Learned

- Repeated themes in the code porting work
  - Finding more threadable work for the GPU
  - Improving memory access patterns
  - Making GPU work (kernel calls) more coarse-grained if possible
  - Overlapping data transfers with other work (leverage HyperQ)
- Use as much asynchronicity as possible (CPU, GPU, MPI, PCIe-2)
- The difficulty level of the GPU port was in part determined by:
  - Structure of the algorithms—e.g., available parallelism, high computational intensity
  - Code execution profile—flat or hot spots
  - The code size (LOC)

CAAR: Selected Lessons Learned

- We estimate possibly 70-80% of developer time is spent in code restructuring, regardless of whether using CUDA / OpenCl / OpenACC / …
- More available flops on the node should lead us to think of new science opportunities enabled—e.g., more DOF per grid cell
- We may need to look in unconventional places to get another ~30X thread parallelism that may be needed for exascale—e.g., parallelism in time

Science requires exascale capability this decade

Mission: Deploy and operate the computational resources required to tackle global challenges
- Deliver transforming discoveries in climate, materials, biology, energy technologies, etc.
- Enabling investigation of otherwise inaccessible systems, from regional climate impacts to energy grid dynamics
- Maximize scientific productivity and progress on largest scale computational problems
- World class computational resources and specialized services for the most computationally intensive problems
- Stable hardware/software path of increasing scale to maximize productive applications development

Path to exascale

- Hierarchical parallelism
  - Improve scalability of applications
- Explicit data management
  - Between CPU and GPU memories
- Data locality: Keep data near processing
  - GPU has high bandwidth to local memory and large internal cache
- Expose more parallelism
  - Code refactoring and source code directives can double performance
- Heterogeneous multicore processor architecture
  - Using right type of processor for each task

How to get an allocation
DOE Allocation Policy for Leadership Facilities

- Primary Objective:
  - "Provide substantial allocations to the open science community through an peered process for a small number of high-impact scientific research projects"

OLCF allocation programs

<table>
<thead>
<tr>
<th>Mission</th>
<th>INCITE</th>
<th>ALCC</th>
<th>Director's Discretionary</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-risk, high-payoff science that requires LCF-scale resources</td>
<td>High-risk, high-payoff science aligned with DOE mission</td>
<td>Strategic LCF goals</td>
<td></td>
</tr>
<tr>
<td>Call Duration</td>
<td>1 year (Closes June)</td>
<td>1 year (Closes February)</td>
<td>Rolling</td>
</tr>
<tr>
<td>Typical Size</td>
<td>20M - 100M core-hours/yr</td>
<td>5 - 10 projects</td>
<td>1M - 75M core-hours/yr</td>
</tr>
<tr>
<td>Review Process</td>
<td>Strategic, Peer Review</td>
<td>Computation Readiness</td>
<td>Strategic, Peer Review Computation Readiness</td>
</tr>
<tr>
<td>Managed by</td>
<td>INCITE management committee</td>
<td>DOE Office of Science</td>
<td>OLCF management</td>
</tr>
<tr>
<td>Availability</td>
<td>Open to all scientific researchers and organizations including industry</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Expanding Access to Supercomputing

- Innovative and Novel Computational Impact on Theory and Experiment (INCITE)
- Awards time for "grand challenges" in science and engineering
- Open to researchers from academia, government labs, industry
- Overwhelming demand, record number of proposals
- Demand exceeds awards by 3x

Distribution of INCITE (2012)

<table>
<thead>
<tr>
<th>Hours allocated</th>
<th>Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.9M</td>
<td>3</td>
</tr>
<tr>
<td>6.5M</td>
<td>3</td>
</tr>
<tr>
<td>15.2M</td>
<td>15</td>
</tr>
<tr>
<td>925M</td>
<td>45</td>
</tr>
<tr>
<td>2505M</td>
<td>55</td>
</tr>
<tr>
<td>8.098M</td>
<td>66</td>
</tr>
<tr>
<td>1.6B</td>
<td>69</td>
</tr>
<tr>
<td>1.7B</td>
<td>57</td>
</tr>
<tr>
<td>1.7B</td>
<td>60</td>
</tr>
<tr>
<td>SB</td>
<td>61</td>
</tr>
</tbody>
</table>

2014 INCITE Call for Proposals

- Planning Request for Information (RFI)
- Call opens April, 2013. Closes June, 2013
- Expect to allocate more than 5 billion core-hours
- Expect 3X oversubscription
- Awards to be announced in November for CY 2014
- Average award to exceed 50 million core-hours
- INCITE Proposal Writing Webinars!

Contact information
Julia C. White, INCITE Manager
whitejc@DOEleadershipcomputing.org

Science breakthroughs at the LCF

<table>
<thead>
<tr>
<th>Hours allocated vs requested</th>
<th>Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.9M</td>
<td>6.5M</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Hot off the press...
Highly Scalable Parallel BLAST to Speedup Genomic Data Analysis

Sharan Rekepalli, University of Tennessee, and Richard Casey, Colorado State University

The Challenge
- Improve speed and scaling of downstream sequence analysis tools used to metagenomics data analysis of various micro-biomes

The Success
- Algorithmic and I/O improvements of NCBI-BLAST using HSP module versus the MPI-BLAST

Implications for Future Research
- The speedups achieved reduce the computational time needed for novel biological knowledge discovery from years to days on Titan supercomputer

Speedups achieved on Titan

With Algorithmic, I/O improvements and code optimizations we generated results for metagenomics sample set with 10 million sequences in one hour compared to months of computation using MPI-BLAST

Weak Scaling Study of HSP-BLAST

Time in seconds

Number of sequences

Number of cores

Conclusions
- Power is the primary constraint
- Clock speed trend flat to slower over coming years
- Increased parallelism lowers memory per core
- Accelerated/hybrid-multicore heterogeneity is the new norm
- Cost of data movement will dominate performance problems.
- Importance of data locality (compute where data resides)
- Memory/Network bandwidth need is increasing (aka “Big Data” needs)...
- ...But not fast enough
- Reliability will be a big problem going forward
- OLCF resources are available through open, peer-reviewed allocation mechanisms.

Acknowledgements
- OLCF-3 CAAR Team: Bronson Messer, Wayne Joubert, Mike Brown, Matt Norman, Markus Eisenbach, Ramanan Sankaran
- OLCF Users: Bhanu Rekepalli, Richard Casey
- OLCF-3 Hardware Vendor Partners: Cray, AMD, and NVIDIA

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

Questions