Argonne Leadership Computing Facility: Applications Performance and Site Update

Raymond Loy

Applications Performance Engineering and Data Analytics (APEDA)

Argonne Leadership Computing Facility
ALCF Mission

Provide the computational science community with a world leading computing capability dedicated to breakthrough science and engineering.

INCITE

The Department of Energy’s ASCR Innovative and Novel Computational Impact on Theory and Experiment Program (INCITE) annually awards researchers worldwide millions of supercomputer processor hours at DOE’s global flagship facilities. INCITE has enabled open science breakthroughs achieved on the IBM Blue Gene/L and IBM Blue Gene/P supercomputers at Argonne National Laboratory to grow tremendously in number and impact over the past seven years.
ALCF Timeline

2004
- DOE-SC selected the ORNL, ANL and PNNL team for Leadership Computing Facility award

2005
- Installed 5 teraflops Blue Gene/L for evaluation

2006
- Began production support of 6 INCITE projects, with BGW
- Continued code development and evaluation
- “Lehman” Peer Review of ALCF campaign plans

2007
- Increased to 9 INCITE projects; continued development projects
- Installed 100 teraflops BlueGene/P (late 2007)

2008
- Began support of 20 INCITE projects on BG/P
- Added 557 Teraflops BG/P

2009
- 28 Projects / 400 M CPU-hours

2010
- 35 Projects / 656 M CPU-hours
Goal: Acquire the next-generation Blue Gene, 10-20PF
January 2009 - CD0 approved
July 2009 - Lehman Review (CD1/2) passed
Contract review by Lehman Committee pending

*We expect to install Mira in early 2012*
TCS: ALCF’s New Home

• 7 stories
• 25,000 ft^2 computing center
• 18,000 ft^2 library
• 10,000 ft^2 advanced digital laboratory
• 7,000 ft^2 conference center
• 30 conference rooms
• 3 computational labs
• 700 employees from 6 divisions
Capability and Science

- Astrophysics: 17%
- Biological Sciences: 19%
- Chemical Sciences
- Climate Research: 11%
- Combustion: 2%
- Computer Sciences: 4%
- Engineering: 3%
- Environmental Sciences: 1%
- Lattice Gauge Theory: 17%
- Materials Sciences: 11%
- Nuclear Energy: 2%
- Nuclear Physics: 2%
- Plasma Physics: 7%
- Applied Math (CFD):
## Projects with Current Allocations Running at Capability

### 2010 INCITE Projects with Runs in Capability Range

<table>
<thead>
<tr>
<th>Project</th>
<th>Name</th>
<th>Project</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultiscaleCoupTurb</td>
<td>&lt;kuri&gt;</td>
<td>Turbulent_Mixing</td>
<td>&lt;lel&gt;</td>
</tr>
<tr>
<td>LatticeQCD</td>
<td>&lt;mac&gt;</td>
<td>SSSPP</td>
<td>&lt;lus&gt;</td>
</tr>
<tr>
<td>CCESDev</td>
<td>&lt;washi&gt;</td>
<td>ProteinLigand_In</td>
<td>&lt;bin&gt;</td>
</tr>
<tr>
<td>CompNuclStruct</td>
<td>&lt;dean&gt;</td>
<td>HydrideNano</td>
<td>&lt;vol&gt;</td>
</tr>
<tr>
<td>LSCondMatterSim</td>
<td>&lt;coven&gt;</td>
<td>SupernovaModels</td>
<td>&lt;lam&gt;</td>
</tr>
<tr>
<td>PlasmaMicroturb</td>
<td>&lt;tang&gt;</td>
<td>ProteinStruct</td>
<td>&lt;bak&gt;</td>
</tr>
<tr>
<td>PEACEndStation</td>
<td>&lt;wor&gt;</td>
<td>High_Reynolds</td>
<td>&lt;mos&gt;</td>
</tr>
</tbody>
</table>

### Discretionary Projects with Runs in Capability Range

<table>
<thead>
<tr>
<th>Project</th>
<th>Name</th>
<th>Project</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>IonPermeation</td>
<td>&lt;rou&gt;</td>
<td>Nwchem4bgp</td>
<td>&lt;ham&gt;</td>
</tr>
<tr>
<td>boeingIDDES</td>
<td>&lt;spal&gt;</td>
<td>NEK5000</td>
<td>&lt;fis&gt;</td>
</tr>
<tr>
<td>Virtual_Reactor</td>
<td>&lt;kau&gt;</td>
<td>PhastaIO</td>
<td>&lt;jan&gt;</td>
</tr>
<tr>
<td>QLG-Turbulence</td>
<td>&lt;vah&gt;</td>
<td>CharmRTS</td>
<td>&lt;kale&gt;</td>
</tr>
</tbody>
</table>

*NEKTAR-G2 Karniadakis, Scalable_Matching Pothen*
Usage 2009-08-01 to 2010-03-31

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>% of Total Usage</th>
<th>Capability</th>
<th>% of Time in Capability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cumulative</td>
<td>677M</td>
<td>100%</td>
<td>162M</td>
<td>23.9%</td>
</tr>
<tr>
<td>INCITE</td>
<td>434M</td>
<td>64.1%</td>
<td>92.6M</td>
<td>21.3%</td>
</tr>
<tr>
<td>Discretionary</td>
<td>209M</td>
<td>30.9%</td>
<td>56.1M</td>
<td>26.8%</td>
</tr>
<tr>
<td>ALCC*</td>
<td>12M</td>
<td>1.83%</td>
<td>0M</td>
<td>0%</td>
</tr>
<tr>
<td>Support</td>
<td>21M</td>
<td>3.16%</td>
<td>13.2M</td>
<td>61.7%</td>
</tr>
</tbody>
</table>
Achieving Capability

Social
- Leap to Petascale
- Personal visits to focus on targeted issues

Catalyst/PE collaborations
- Explore project plans and needs to encouraging capability plans
  - Help identify and eliminate blocks
- Explicit assistance for expanding to capability
  - Algorithm/method assistance
  - Performance assessments
  - Recommendations for improvements
  - Funding explored for longer term efforts

Technical
- Jobs accrue priority as a function of job size and queued time weighted by requested run length
- Big Run Monday
  - Reintroducing early successful model
- Restrict non-capability run-times
- Backfill and small-size jobs have reduced run-times to shorten waiting times for capability jobs
  - 90% of all capability jobs run in 36 hrs

Intrepid’s quick turnaround on very large ($O(32K)$ tasks) jobs has allowed us to rapidly discover and fix the many issues that show up when running a large code like the CCSM at scale. In FY10, very large CCSM jobs will be routine instead of heroic one-time-only efforts.

- Warren Washington

90% of all capability jobs run is less than 36 hours. The last 5 32-rack runs were queued for:
- 8:54:31
- 8:10:24
- 6:47:46
- 9:37:19
- 1 day, 4:12:33
Capability Usage (con’t)

Capability

Target: 250000000
Current total: 210200415.1 (84.1%)
2009-08-01 to 2010-05-08
Examples of Collaborations

**INCITE**
- NEK5000
  - Resolved MPI comm_dup issue with IBM bringing scaling to full 32-racks
  - 30% performance improvement remapping application
  - Recent success at 72-racks at Juelich
- NWChem
  - eliminated numerous IO bottlenecks incompatible with BGP filesystem
  - implemented passive progress in ARMCI (~170x BW increase => 2x application speedup)
  - first multi-threaded kernels in NWChem
- DNS3D
  - Identified key error in code specific implementation of globals enabling 64K and 131K cores
- GPAW/AVBP/FLASH

**Pre-INCITE/Discretionary**
- HIRAM - GFDL
  - ALCF LDRD
- HSCD
  - Aggressive rework of parallel algorithms, threading, I/O, etc
  - Increased scaling from 512 cores to 32K cores in 6 months
- NEKTAR
  - 30% performance improvement at large scales
- MADNESS
  - Ran within 2 days of starting port
  - Threads scale near perfectly
  - Spring 2010 to address larger MPI scaling challenges
- MPQC
GAMESS
(General Atomic and Molecular Electronic Structure System)

- *Ab Initio* Quantum Chemistry package
- Implements all major QC methods (Hartree-Fock, MP2, Coupled Cluster, MCSCF, CI, DFT, etc)
- Maintained by the research group of Prof. Mark Gordon at Iowa State University ([http://www.msg.chem.iastate.edu/gamess/](http://www.msg.chem.iastate.edu/gamess/))
- Highly scalable, including many Distributed Data algorithms
- Ported to all major architectures
- Implements the Fragment Molecular Orbital (FMO) method for large systems
  - Divide into fragments, *Ab Initio* calculations on fragments and their dimers
  - Coulomb field of entire system is included
  - Compute time scales linearly with system size
  - Efficiently parallelized (to 32-racks of BG/P and beyond), see:
    - [http://staff.aist.go.jp/d.g.fedorov/fmo/main.html](http://staff.aist.go.jp/d.g.fedorov/fmo/main.html)
GAMESS - BG/P Performance enhancements

All about the I/O

- Originally, the FMO code in GAMESS was written for cluster MPPs with local disc.
  - *Every process writing 4 or 5 files each*

Fixes:

- Store fragment densities in DDI arrays to remove a major I/O burden. [Dmitri Fedorov (chief developer of FMO, at AIST, Japan)]
- Further modifications so that only group master processes open files that might be written to. (Previously, all processes still opened files, even if subsequently unused)
- Use the 'RAMDISK' feature on to store the remaining files opened by the 'master' processes.
GAMESS on BlueGene/P 'Intrepid', Leadership Computing Facility, Argonne National Laboratory (water) 12,288 atoms Ab Initio (FMO)/aug-cc-pVDZ
GAMESS on BlueGene/P 'Intrepid', Leadership Computing Facility, Argonne National Laboratory
(water) 12,288 atoms Ab Initio (FMO)/aug-cc-pVDZ
ALCF Quantum Chemistry Application Progress

- NWChem
  - 170x ARMCI bandwidth improvement giving 20%-50% application speedup
  - First multi-threaded and PPC tuned NWChem kernels (18.7%)
    • Increases scaling
    • Alleviate memory bottleneck
    • Match for future architectures
    • Double-hummer + other optimizations
  - Reduced startup time
    • 20min -> 40sec
  - Runs to 64K-128K cores
    • Further ARMCI work underway to improve 64K and 128K scaling

- MPQC
  - 100x faster than 2 weeks ago through ALCF work
  - Working on dynamic load balancing to further improve past 64K cores
Large Eddy Simulation of Two Phase Flow Combustion in Gas Turbines

**Science**

- Improve airplane and helicopter engine designs – reduce design cost, improve efficiency
- Study the two phase flows and combustion in gas turbine engines
- 2009 Accomplishments
  - *First and only simulations of a full annular chamber*
  - Identified instability mechanisms reducing efficiency in gas turbines

**Code and Challenges**

- AVBP
  - Large Eddy Simulation (LES) technique
  - 3D Navier Stokes, unstructured (*full tetrahedral*) mesh
  - Scaling growth from 256 cores in 2006 to 16,384 cores in 2009
    - Restructuring code parallelization paradigm
    - Restricted by grid generation for scale
  - I/O work underway

“ALCF support was also key to improve the communication pattern used by the code thus minimizing the hops for communications and increasing scalability significantly. “
FLASH Project on Intrepid

**Science**
- Answered critical question on critical process in Type Ia supernovae
  - First simulated buoyancy-driven turbulent nuclear combustion in the fully-developed turbulent regime while also simultaneously resolving the Gibson scale
  - Showed that ultra-high resolution studies are not critical for SN Ia models
- Comparing methods of detonation for Type Ia

**Code and Challenges**
- Methods/Algorithms
  - Operator split, multi-physics, navier-Stokes finite-volume
  - Block structured adaptive mesh
  - Multi-pole and multi-grid gravity solvers

**Comparing methods of detonation for Type Ia**

**Don Lamb, University of Chicago**

**ALCF/MCS/University of Chicago collaboration to modernize 9 year-old I/O approach.**
- Reduced memory footprint
- Increased scalability
Simulation of "High" Reynolds Number Turbulent Boundary Layers

**Science**
- Probe the turbulent flow phenomena over solid boundaries to improve designs of vehicles
- Poor understanding of the turbulence in these flows limits improved designs
- 2010 target
  - Single campaign at 32K cores
  - 16384x715x3840 mesh, 0.37GB/core
  - Wall clock ~2000 hours

**Code and Challenges**
- Time-dependent 3D incompressible Navier-Stokes
- Runge-Kutta timestep; spectral element/finite different in space

<table>
<thead>
<tr>
<th>2010Q1</th>
<th>2010Q2</th>
<th>2010Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fine tune/Prep</td>
<td>Production campaign</td>
<td>Initial Data analysis</td>
</tr>
</tbody>
</table>
Probing the Non-Scalable Nano Regime in Catalytic Nanoparticles

**Science**
- Gold has exciting catalytic capabilities that are still poorly understood
  - e.g. CO into CO$_2$
- First principle calculations of the actual nano-particles are needed

**Code and Challenges**
- Project spent 2 years as ALCF discretionary
  - Physics and performance work
- Post-doc in ALCF working with research group developing GPAW capability
  - Reduced memory footprint
  - New layer of parallelization
  - Expanded production from 128 cores to 16,384; 64K cores at EOY

*Necessary to translate grids between 2 methods*
Computational Nuclear Structure

- Green’s Function Monte Carlo (GFMC)
  - *Ab initio* calculation of properties of light nuclei
  - New common benchmark for other methods

- Calculations of $^{12}$C with complete Hamiltonian
  - INCITE time, 32K cores
  - Best converged *ab initio* calculations of $^{12}$C ever
  - Key calculation to launch into study of effects of different terms in the nuclear interaction and compute excited states

- ADLB
  - ANL-MCS (Lusk) developed multipurpose library for distributing load
    - SciDAC project
  - Required for $^{12}$C simulation
  - OpenMP + MPI programming model

![Graph showing density of $^{12}$C](image)
Computational Nuclear Structure

- Green’s Function Monte Carlo (GFMC)
  - *Ab initio* calculation of properties of light nuclei
  - New common benchmark for other methods
- Calculations of $^{12}$C with complete Hamiltonian
  - INCITE time, 32K cores
  - Best converged *Ab initio* calculations of $^{12}$C ever
  - Key calculation to launch into study of effects of different terms in the nuclear interaction and compute excited states

- ADLB
  - SciDAC funded ANL-MCS developed multipurpose library for distributing load
  - Required for $^{12}$C simulation
  - OpenMP + MPI programming model
Computational Protein Structure Prediction and Protein Design

- Computationally design protein-based inhibitors towards pathogens like H1N1
- Rapid turn around of huge campaigns on ALCF reinvented how the science is done and enables new research
- Three Key 2009 Results
  - *Determination of protein structures* 100-200 amino acids in size with sparse NMR constraints, using a new genetic algorithm – *Science*, 19Feb2010
  - Novel method for design of protein inhibitors - generated 2 candidates to serve as inhibitors of pandemic influenza
  - *Structure determination and refinement into low-resolution (3.5-8A) density data from cryo-electron microscopy and X-ray crystallography* - *J. Mol. Biol. Sept. 2009*
  - Key new algorithms developed, to be released in Rosetta3

Genetic algorithm optimization and memory driving size of simulations.

Large protein structures can now be computationally determined by incorporating backbone-only NMR data into Rosetta. Shown here is the structural comparison of ALG13 (201 amino acids) determined by INCITE. Determination of protein structures now also include using RDCs and backbone NH-NH NOEs (B) experimentally by conventional NMR methods (PDB id: 2jzc).
Designing Better Materials for Nuclear Reactors

Priya Vashista
University of Southern California

Key to design of next-generation nuclear reactor
- Revealed a missing link between sulfur-induced inter-granular amorphization and embrittlement
- 47 million-atom ReaxFF MD simulations exploring fracture modes in materials
- 50M CPU-Hours

“. the proposed upgrade of the ALCF resources to a 10-20 Petaflops Blue Gene Q system would enable our research group to make major breakthrough simulations on a number of applications of high relevance to DOE. We look forward to working with the ALCF on their new platform. “

With Sulfur
Color: common-neighbor parameter that characterizes atomistic defects

(A) close-up fracture simulation in nanocrystalline Ni with amorphous sulfide GB layers
Laser-Driven Coherent Betatron Oscillation in a Laser-Wakefield Cavity

- Develop laser-plasma accelerators with improved emittance
- Possibility of generating coherent x-ray radiation from laser-plasma accelerators
  - imaging single large molecules, such as for example proteins in solutions
- Vorpal in collaboration with Tech-X
- ~1M core-hours discretionary

Vorpal in collaboration with Tech-X
~1M core-hours discretionary

DOE Office of Science DE-AC02-06CH11357, DE-FG02-01ER41178, DE-FC02-07ER41499, DE-FG02-04ER41317

Electron-beams in plasma bubbles take up a sinusoid form as a consequence of overlap with the drive-laser's electromagnetic field. Trajectories of individual electrons will become coherent in a frame moving with the phase-velocity of the laser. This phenomenon has potential to become a source of coherent x-rays, useful e.g. for imaging single large molecules, such as proteins in solutions.
Leap To Petascale Workshops

- Annual multi-day workshops to focus on scaling and performance
  - Current INCITE projects
  - INCITE applicants to prepare proposals
- ALCF staff focus entirely on workshop
- External expertise for in-depth dives
  - Performance tools
  - Debuggers
  - IBM personnel

L2P 2009
- Significant progress on 8 projects
- 7 INCITE proposals
- Ex: Boldyrev new INCITE project
  - Scaled code from 4-32 racks
  - 40% performance improvement with ESSL implementation
- Ex: Georgie Tech LBMFEA
  - Identified the key limiting methods to scalability and a path to fix

L2P 2008
- 21 projects attended
- 7 are current INCITE projects
ALCF-2 Early Science Program
Early Science Program

- “In early 2012 the ALCF will be installing at least 10PF of a next-generation Blue Gene. We are asking the community to help us make this deployment as successful and productive as possible.”

- Goals
  - Help us shake-out the system and software stack using real applications
  - Develops community and ALCF expertise on the system
  - A stable and well-documented system moving into production
  - Exemplar applications over a broad range of fields
  - At least 2 billion core-hours to science

- 2010 ESP Proposal Timeline
  - January 29th - Call for Proposals Issued
  - April 29th – Call for Proposals Closes
  - Mid June – ESP Awards Announced
  - July – “Getting Started” Workshop for ESP Teams
Next-Generation Blue Gene

- Nodes
  - Higher levels of HW parallelism
  - Applications expecting approximately 10x performance/node
  - Twice the memory/core

- Interconnect
  - Continuing the Blue Gene tradition of a very strong interconnect

- I/O
  - Optimizing the end-to-end I/O software stack
  - Targeting 240 GB/s aggregate peak BW
Early Science Program Timeline

- Call Opens January 2010
- Proposals due April 29, 2010
- Rapid ramp up in activity through end of 2010
Conclusion

- A broad variety of INCITE projects running at “capability” scale
  - Usage can vary from intense campaigns to a sub-component of regular usage
- Discretionary projects are key in developing new science and scale
  - Develop INCITE and future INCITE capability
  - Explore science questions to identify large scale directions for INCITE
- ALCF is actively engaged in the next generation Blue Gene machine
- The Early Science Program will bridge the gap from the current machine to the next