Experiences in Optimizing a $250K Cluster for High-Performance Computing Applications

Kevin Brandstatter
Dan Gordon
Jason DiBabbo
Ben Walters
Alex Ballmer
Lauren Ribordy
Ioan Raicu

Illinois Institute of Technology

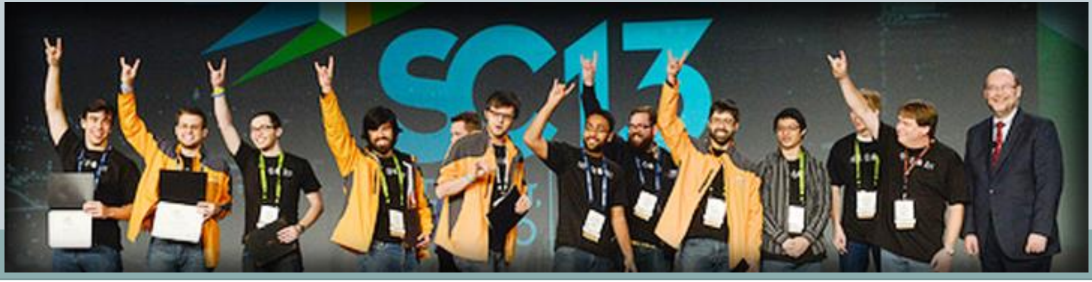
December 3rd, 2014

IEEE/ACM Supercomputing/SC 2014

- One of the largest conferences on SuperComputing and HPC
- Over 10000 Registered attendees
- 356 Exhibitors on over 140000 sq ft of exhibit space
- Scinet, the worlds fastest network at 1.5 Tb of bandwidth
- Technical program selected from more than 394 submissions (21%)
- Bi Annual Top 500 Awards

Student Cluster Competition (SCC)

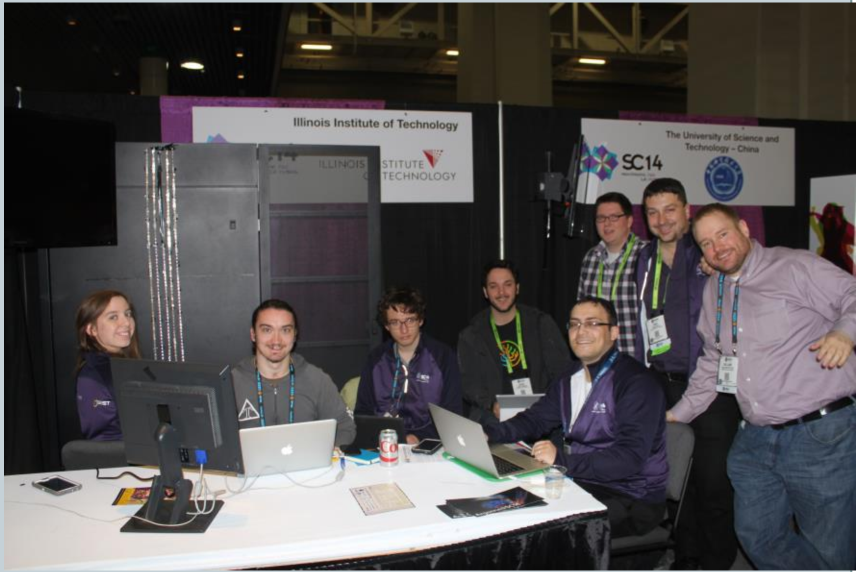
- 6 Undergraduates (And staff advisor)
- 26 Amp power limit
- 4 applications (plus Linpack)
- Optimize applications
- Top of the line hardware (no price limit)
- 48 hour competition (~6 months preparation)
- Interview scores
- HPC Impact showcase presentation



K

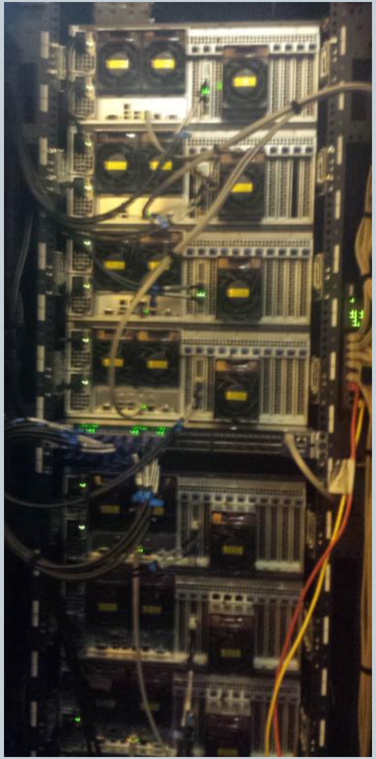
Chicago Fusion Team Members

- Students:
  - Alex Ballmer (1st year UG)
  - Ben Walters (2nd year UG)
  - Dan Gordon (4th year UG)
  - Jason DiBabbo (4th year UG)
  - Kevin Brandstatter (4th year UG)
  - Lauren Ribordy (Highschool)
- Advisor:
  - Ioan Raicu (IIT/Argonne)
- Others:
  - **William Scullin (Argonne), Ben Allen (Argonne)**, Cosmin Lungu (1st year UG) Andrei Dumitru (1st year UG), Adnan Haider (1st year UG), Dongfang Zhao (4th year PhD), Tonglin Li (6th year PhD), Ke Wang (5th year PhD), Scott Krieder (4th year PhD)


Illinois Institute of Technology
ILLINOIS INSTITUTE TECHNOLOGY
SC14
The University of Science and Technology – China

Hardware

- 6 node cluster (originally 8)
- 2x Infiniband 56Gb/s
- 36-port Infiniband switch
- 2x Intel Xeon E5-2699 v3 (Haswell) 18-core CPUs @ 2.3 GHz per node
- 10 Nvidia K40 GPUs (2 per node on 5 nodes)
- 128 GB RAM per node
- ~3TB of SSD storage



Software

- CentOS 7
- Warewulf (cluster management)
- Intel Compilers
- MVAPICH2 / Intel MPI
- CUDA
- Slurm (job scheduler)
- GPFS (parallel file system)
- Matlab

Sponsors

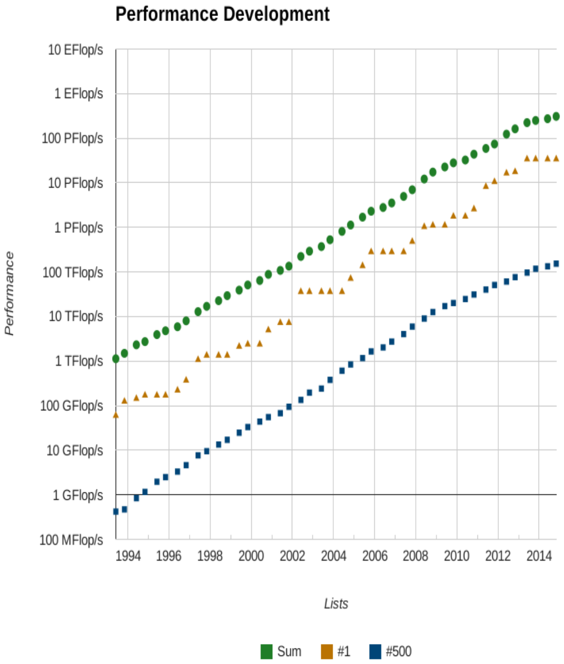
intel®
Argonne
NATIONAL LABORATORY
Mellanox®
TECHNOLOGIES
NVIDIA®

Applications

- Linpack/HPCC
- NAMD: Not (just) Another Molecular Dynamics program
- ADCIRC: ADvanced CIRCulation model
- Matlab: Seismic Analysis by Reverse-Time Migration
- Enzo (Mystery Application): Scientific Cosmological Simulation Application

Linpack/HPCC

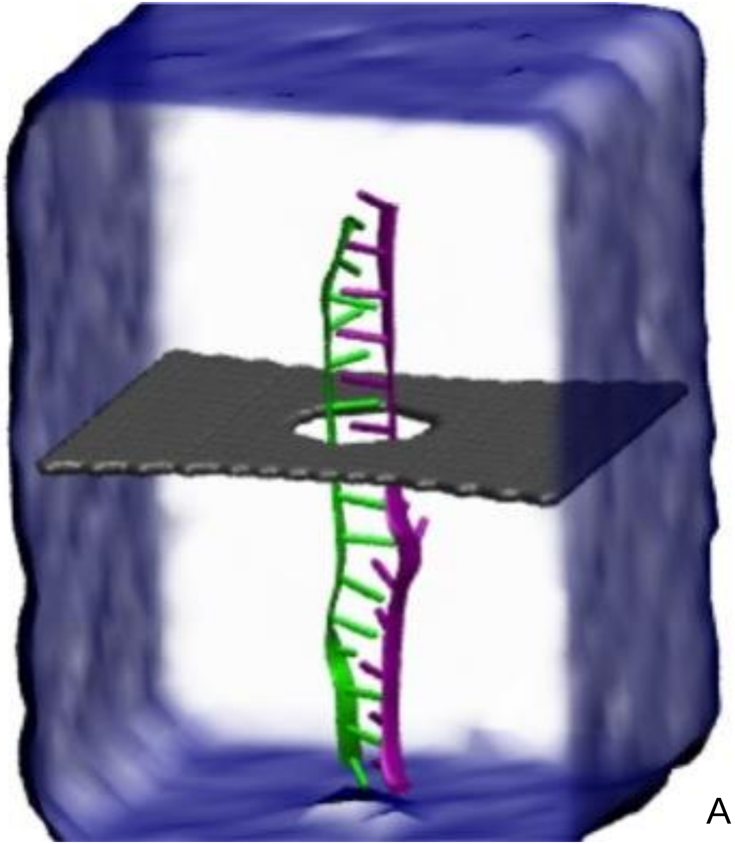
- Standard set of benchmarks for HPC systems
- Key benchmark is linpack, a measure of compute [Flops] performance (CPU Intensive)
- 1st Place in Unmodified Linpack benchmark
- 4th Overall HPCC score
- Dynamic frequency scaling to keep within power constraints



K

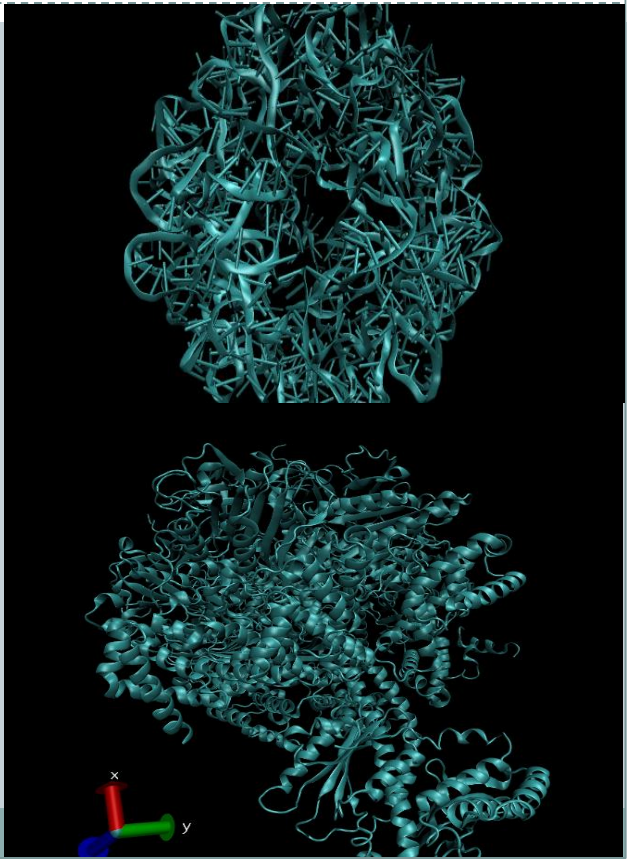
NAMD

- NAnoscale Molecular Dynamics Simulator
- Simulates interactions between very large molecules
- Used for modeling folding proteins
- Capabilities – Standard mpi, CUDA, SMP (not used)
- Runs from the charm parallelism framework (compiled with the intel c++ compiler, then with charmc)



NAMD

- Inputs: tcl coordinates file
- Outputs: coordinates file and trajectory file for animation
- CPU intensive application
- Visualization can be animated using trajectories applied to coordinate files



ADCIRC

- Fluid dynamics and circulation simulator
- Ex: Hurricane Katrina simulation
- Disk intensive (Must be careful about I/O)
- Runtimes 2 minutes to ~40 hours

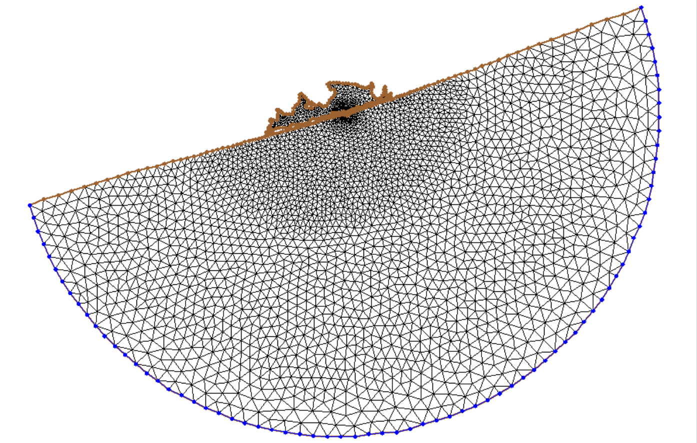
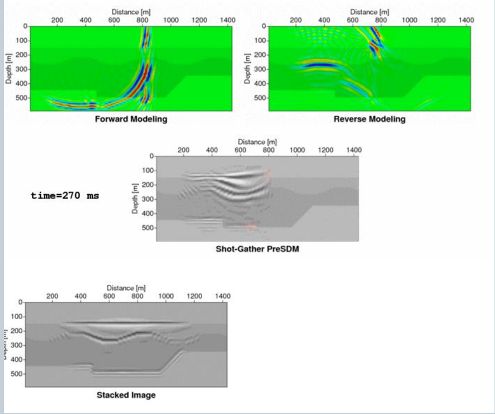
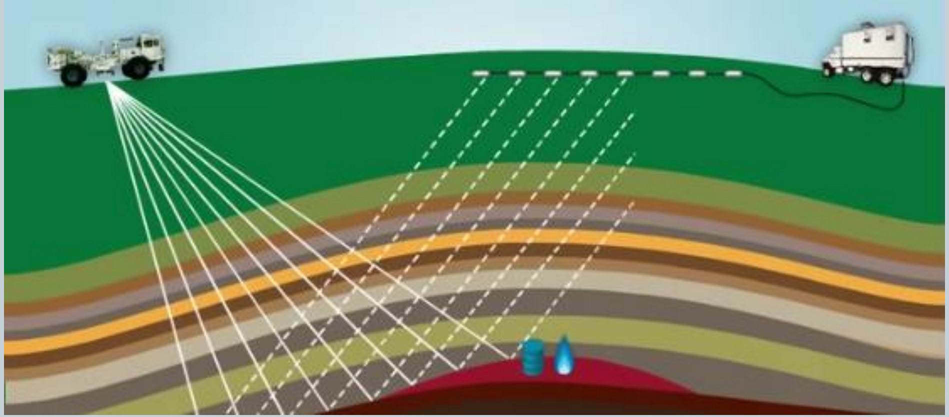


Figure E2.2a Shinnecock Inlet Example Problem Finite Element Grid

Matlab

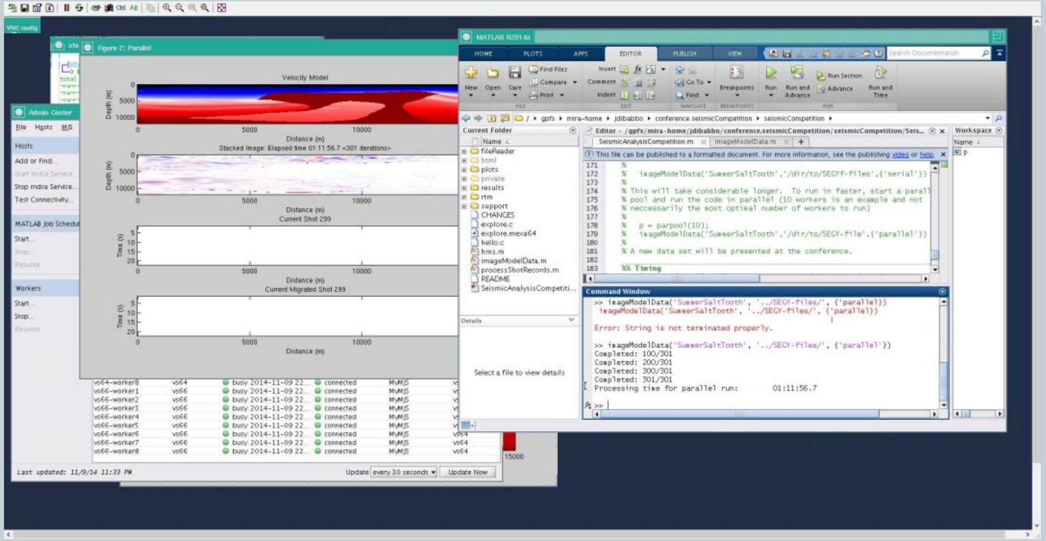
- Seismic Analysis by Reverse-Time Migration
- Datasets ranging from 16MB to 3.4GB
- Memory Intensive Application
- Runtime from 5 minutes to 10 hours





Matlab

- Optimization Opportunity: Add GPU/CUDA code to measure against CPU version



gpfs ▸ mira-home ▸ jdibabbo ▸ conference.seismicCompetition ▸ seismicCompetition ▸

Current Folder: fileReader, html, plots, private, results, rtm, support, CHANGES, explore.c, explore.mexa64, hello.c, hms.m, imageModelData.m, processShotRecords.m, README, SeismicAnalysisCompetiti...

Select a file to view details

Editor - /gpfs/mira-home/jdibabbo/conference.seismicCompetition/seismicCompetition/Seis...

SeismicAnalysisCompetition.m | imageModelData.m

This file can be published to a formatted document. For more information, see the publishing video or help.

```
    %
    %   imageModelData('SummerSaltTooth','/dir/to/SEGYf-files',{'serial'})
    %
    % This will take considerable longer.  To run in faster, start a parall
    % pool and run the code in parallel (10 workers is an example and not
    % neccessarily the most optimal number of workers to run)
    %
    %   p = parpool(10);
    %   imageModelData('SummerSaltTooth','/dir/to/SEGY-file',{'parallel'})
    %
    % A new data set will be presented at the conference.

    %% Timing
```

Command Window

```
>> imageModelData('SummerSaltTooth', '../SEGY-files/', {'parallel})
 imageModelData('SummerSaltTooth', '../SEGY-files/', {'parallel})
                                                      |
Error: String is not terminated properly.

>> imageModelData('SummerSaltTooth', '../SEGY-files/', {'parallel'})
Completed: 100/301
Completed: 200/301
Completed: 300/301
Completed: 301/301
Processing time for parallel run:       01:11:56.7

>>
```

Workspace: p

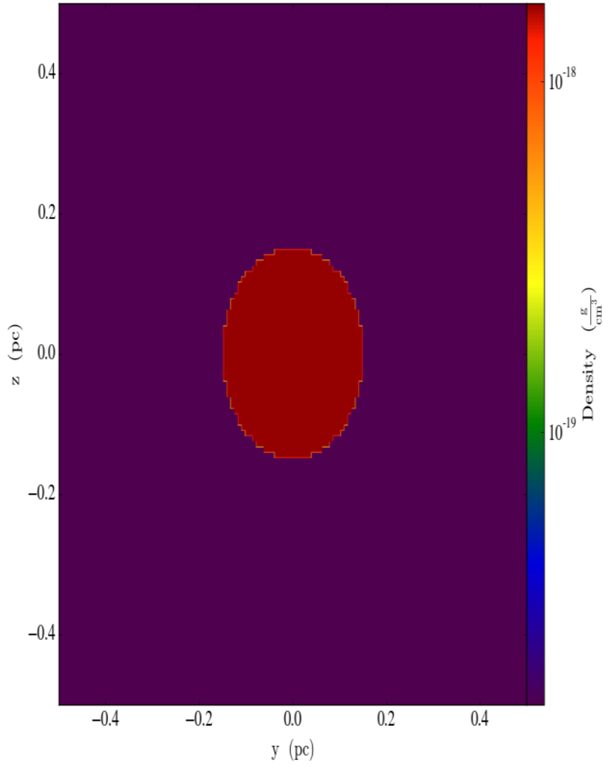
vs64-worker8	vs64	busy 2014-11-09 22...	connected	MyMJS	v
vs66-worker1	vs66	busy 2014-11-09 22...	connected	MyMJS	v
vs66-worker2	vs66	busy 2014-11-09 22...	connected	MyMJS	v
vs66-worker3	vs66	busy 2014-11-09 22...	connected	MyMJS	v
vs66-worker4	vs66	busy 2014-11-09 22...	connected	MyMJS	v
vs66-worker5	vs66	busy 2014-11-09 22...	connected	MyMJS	v
vs66-worker6	vs66	busy 2014-11-09 22...	connected	MyMJS	vs64
vs66-worker7	vs66	busy 2014-11-09 22...	connected	MyMJS	vs64
vs66-worker8	vs66	busy 2014-11-09 22...	connected	MyMJS	vs64

Last updated: 11/9/14 11:33 PM

Update every 30 seconds ▾ | Update Now

Enzo

- Scientific Cosmological Simulation Application
- Application announced at start of competition
- Simulations run for several minutes to several hours
- Outputs from a few gigabytes to several terabytes
- Visualization of dataset output



K

Visualizations

- LED lights
- Programmable with Java and Python
- Plan: display real-time power readings from PDUs on LED lights; if power limit breached, code red!
- Worked well to draw people over
- Synced well with the sirens
- Hardware:
  - Individually programmable LED light strips
  - Fade candyboard
  - A soldering iron
  - Power supply
  - Lots of wires and electrical tape!!!

What We’ve Learned

- Automating processes will save your life
- Stateless provisioning is priceless
- The wonders of resource management (Slurm is still tempermental)
- How to (not) break electrical circuits and how to solder circuits
- Older hardware (e.g. SSD drives) are not worthwhile due to issues in reliability
- The error-prone process of managing a computing cluster
- How to tune the OS, storage, network, and HPC apps

Our Biggest Challenge

- Change in complete architecture and software 5 weeks before
  - Chasis → challenged us in low level support for power management
  - CPUs → Ivy bridge to Haswell
  - GPUs → K20 to K40
  - Network → 40Gb/sec Ethernet to 56Gb/sec Infiniband
  - OS → CentOS to Warewolf
  - MPI → OpenMPI to MVAPICH2
- Hardware arrived unassembled 10 days before we shipped (overnight)
- Allowed team only a few days to debug the new environment and tune the code

- Change in complete architecture and software 5 weeks before
  - Chasis → challenged us in low level support for power management

What would we do differently next time

- Run apps simultaneously (possibly on single node)
- Focus on apps
  - Work on them sooner
  - Scaling studies
  - More datasets
- Long-term preparation
- Heterogonous (1~2 nodes with GPUs, 6~10 nodes with CPUs)
- Overclock CPUs
- Automation
  - Job scheduler
  - Power management
- Reduce idle power (GPU/fans)

Thanks! (all)

- A big thanks to the SC14 and its organizers
- Our steadfast advisor Ioan Raicu
- Our tireless helper from Argonne (William Scullin, Ben Allen)
- And Wanda (Argonne) who made it possible for us to ship a 1500 lb crate overnight
- Without them, our cluster would never have reached the epic proportions of awesomeness it has

J