Illinois Institute of Technology – SC15 Student Cluster Competition

Ben Walters*, Alexander Ballmer*, Adnan Haider*, Andrei Dumitru*, Keshav Kapoor*, Calin Segarceau*, William Scullin+, Ben Allen+, Ioan Raicu*+
*Department of Computer Science, Illinois Institute of Technology
*+Argonne Leadership Computing Facility, Argonne National Laboratory
{bwalter4, aballmer, ahaider3, adumitru, kkapoor2, csegarce}@hawk.iit.edu, {wscullin, bsallen}@alcf.anl.gov, iraicu@cs.iit.edu

Background

Abstract

This work investigated power management through CPU frequency scaling as well as managing the CPU idle states. In addition to hardware controls, we use parameter sweeps along with Allinea MAP and Performance Reports to better predict ideal conditions for peak performance. We put all these techniques together to optimize a cluster of 5 resource-heavy nodes for 4 scientific applications and the High Performance Linpack HPC benchmark.

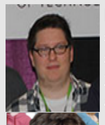
Illinois Institute of Technology (IIT)

Ilinois Institute of Technology (IIT) is a private, technology-focused, research university offering undergraduate and graduate degrees in engineering, science, architecture, business, design, human sciences, applied technology, and law. IIT is centrally located in Chicago.

Our team is comprised of five IIT undergraduates, along with one Naperville Central High School student. The team, has a great deal of experience and skills that will help us win this competition:

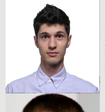
- 5 years collective research experience in IIT's DataSys Lab
- 3 summer research internships at Argonne National Laboratory
- 1 summer research internship at University Corporation for Atmospheric Research
- 2 members of the Student Cluster Competition at SC14 (and 2 backup members)
- 10 months of preparation for SCC at SC15

Team

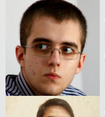


Ben Walters (Team Captain) is a 3rd year undergraduate student in CS at IIT. He has worked in the DataSys lab since June 2013. He was an official member of the SCC 2014 team. His responsibilities include WRF and systems administration.

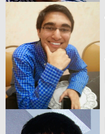
Alexander Ballmer is a 2nd year CS student at IIT. He is a CAMRAS scholar with a full ride scholarship. He was an official member of the SCC 2014 team. His focus is on the HPC Repast and systems administration.



Calin Segarceau is a 1st year CS student at IIT. He's the team's HPL guru. His backup duties include power management and system administration.



Andrei Dumitru is a 2nd year student studying CS. He has been working in the Datasys lab since August 2014. He was a backup member of the SCC 2014 team. His duties include focusing on the MILC application



Adnan Haider is currently a 2nd year student in CS. His research interests include distributed computing, architecture optimization, and parallel network simulation. He was a backup member of the SCC 2014 team. He is point on Trinity

Keshav Kapoor is an 11th grade student at Naperville Central High School. He participated in an REU at IIT during summer 2015. His focus is system resource visualization as well as scientific visulization.

Hardware / Software

Cluster

	Node Description	Aggregate over a 5 node system
Chassis	Supermicro SYS-4048B-TR4FT	
CPU	4 Intel Xeon E7 8867 v3 CPUs	640 HT over 320 Intel x86 cores at 2.5 GHz (12.8TFlops)
Memory	1024GB DDR4 RAM (32x32GB DIMMS)	5TB RAM delivering 2TB/sec bandwidth
Storage	1.9TB SSD per node composed of 200GB SATA SSD OS drive and 1.7TB Intel NVMe PCIe SSD Storage drive	9.5TB of raw storage delivering 14GB/s reads and 9GB/s writes
Network	100Gbs x 4 per node (MCX455A-ECAT CX-4 VPI adapter)	Mellanox EDR InfiniBand switch with 800Gb/sec bisection bandwidth
Power	Maximum peak power draw per node is about 1000 watts	3120 watts with power management tuning and selective use of hardware

Hardware Justification

We decided to partition our resources over fewer nodes with more cores and memory per node:

- Ability to run larger workloads within a single node
- Allows us to eliminate network overhead from small to medium sized workloads
- We can also run many applications on the cluster simultaneously without interference by running each application on a separate node

- By managing CPU frequency and idle states, we can minimize power consumption per work performed
  - This mitigates the power inefficiency that arises in applications that cannot utilize every core
- Having a large amount of memory and high number of cores per node will give us flexibility in how we schedule applications
  - We could run both a memory intensive application and a compute intensive application on the same node
  - Since we have a large amount of memory, we could dedicate a very large percentage towards one app and still have plenty leftover to run a non-memory intensive application

Software Stack

- CentOS 7
- Intel Compilers (C, C++, Fortran)
- Intel MPI
- Intel Math Kernel Library (MKL)
- IBM General Parallel File System (GPFS)
- Slurm job scheduler
- Spack package manager
- Warewulf systems management suite

slurm workload manager

CentOS

Spack

Approaches

Preparation

Lessons learned from last year's competition:

- Run multiple applications simultaneously
- Applications may not necessarily perform most efficiently using all cores on a node
- Applications with different power intensity can be run at the same time
- CPU frequency can be scaled down to reduce power consumption

General preparation Strategies:

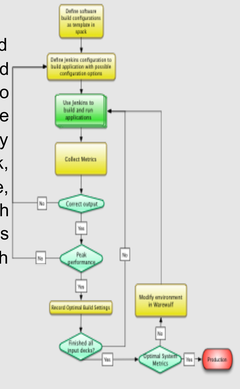
- Learn the applications very well
  - Knowledge of the application can be more useful than system efficiency in some cases
- Find applications that can be run simultaneously
  - Match applications with different power needs
  - Match applications with resource requirements that do not overlap
- Investigate how applications behave at different CPU frequencies
  - If an application performs well (in terms of work per watt) regardless of CPU frequency, then we have more flexibility to lower CPU frequency to fit within our power budget

Application Optimization

- Use Allinea to profile applications and identify resource bottlenecks
- Run each application with Allinea Performance Reports
  - Extremely easy to use
  - Identifies primary resource used (CPU, Network, Disk I/O)
- Identify important application parameters
  - Identify parameters that may affect how the application performs with regards to its critical resource
  - Run small parameters sweeps to find good parameter settings for critical parameters
- Run applications with Allinea MAP to identify functions in which the application spends a lot of time
  - Read application documentation about this function to find its primary use
  - This often leads to insight about which resource its most important

Automated Testing

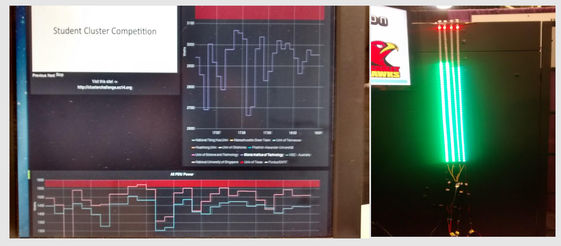
The process of determining the "best" possible combination of application build options, environmental settings, and system configurations to reduce time to solution while minimizing resource utilization and contention is laborious. By utilizing community tools like Spack, Jenkins, Performance Co-Pilot, OProfile, Tau, and Warewulf, we can walk through many combinations and permutations without human input and assure high utilization.



Power Management

The main constraint of the Student Cluster Competition is a total cluster power consumption limit of 26 Amps. To maintain this while maintaining high performance on scientific applications, we used the following methods:

- Record per-node power consumption for application test runs
  - Allows comparison among any dataset or parameter configuration at any point in time
- Overprovision power consumption for the entire cluster
  - Allow the theoretical maximum power consumption of the entire cluster be greater than the power limit
  - No application will use every resource to the fullest, so something can always be turned down to reduce power consumption
- Do not try to use every core for each application
  - For non-CPU intensive applications, cores can be idled to save power consumption without losing much performance



Why Our Team Will Win

- We have a wealth of research and internship experience among the team members
- We have a primary and secondary student that has spent several months learning as much as possible about each application.
- In addition to applications, we assigned primary and secondary students to become experts in several critical areas
  - Power management
  - System administration
  - Parallel file systems
- We utilized Allinea MAP and Allinea Performance Reports to identify critical resources for each application
  - By knowing the main bottleneck of each application, we were able to focus our efforts to tune both application and system parameters
- The team met on a weekly basis since January 2015 (except the summer) to discuss results and experiment strategies moving forward
- The team practiced giving oral presentations to practice communicating their methodologies, results, and knowledge of the applications

Acknowledgements

This work would not be possible without the generous support of Intel and Argonne National Laboratory. This research used resources of the ANL's Leadership Computing Facility (ALCF), a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357. Special thanks goes out to the staff of the ALCF.

Contact: Ioan Raicu iraicu@cs.iit.edu; for more information, see http://sc15.supercomputing.org/conference-program/student-programs/sc15-student-cluster-competition