15 TFlops Haswell vs. 60 TFlops Knight Landing for HPC Scientific Computing Applications

Ben Walters[*], Alex Ballmer[*], Andrei Dumitru[*], Adnan Haider[*], Serapheim Dimitropoulos[*], Ariel Young[*],
William Scullin[‡], Ben Allen[‡], Ioan Raicu[*‡]

[*]Department of Computer Science, Illinois Institute of Technology, Chicago IL, USA
[‡]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne IL, USA

{bwalter4, aballmer, adumitru, ahaider3, sdimitro, ayoung11}@hawk.iit.edu, {wscullin, bsallen}@alcf.anl.gov iraicu@cs.iit.edu

Abstract**—***In order to achieve the next level of performance from large scale scientific computing, we have been exploring ways to optimize cluster utilities and applications in order to maximize performance, while minimizing power usage. To do this we investigated automated ways to manage cluster power consumption through management of processor per-core frequencies (both to over boost and under clock), fan speeds, water cooling, and powering down of accelerators when not in use. In addition to hardware controls, we are exploring automating auto building multiple configurations of applications, and parameter sweeps in order to better predict ideal conditions for peak performance. Four of the six team members were involved in the Student Cluster Competition at Supercomputing/SC 2014, where many lessons were learned, and the team came out of the competition energized to refine the process, and try again This year's team consists of 6 undergraduate students from the Illinois Institute of Technology (IIT), two mentors and staff support from Argonne National Laboratory's (ANL) Leadership Computing Facility, and a faculty advisor with a joint appointment between IIT and ANL. Sponsorship will be from Intel and Mellanox. We propose a cluster configuration leveraging five 4U nodes (using 4 sockets per node with Intel Xeon E7-4880 v3 processors with 18-cores at 2.3GHz, and 4 Intel Knights Landing Xeon Phi accelerators per node) connected by 4 Infiniband (IB) 100Gb/sec full duplex ports per node and creating a direct interconnect between the 5 nodes while eliminating a power hungry IB switch. We estimate that this 20U half rack will deliver approximately 75 double precision TFlops/sec with less than 6200 watts of power, and that MIC or CPU only usage of the system will result in 60TFlops/sec or 15TFlops/sec in less than 3120 watts respectively; we will use aggressive power management techniques to reach these targets. The latest Knights Landing Xeon Phi accelerators have a flop/watt advantage over both NVIDIA and AMD GPUs, and their x86 programmability makes them a more likely candidate to be useful for a wide range of real applications (beyond Linpack).*

I. TEAM MEMBERS

This year's team consists of 6 undergraduate students from the Illinois Institute of Technology (IIT), two mentors from Argonne National Laboratory (ANL) who are system administrators on the IBM BlueGene/Q supercomputer, and a faculty advisor with a joint appointment between IIT and ANL.

Ben Walters (*Team Captain*) is a 2nd year undergraduate student in CS at IIT. He is the recipient of the University Scholarship, a nearly full ride award. He has worked in the DataSys lab since June 2013 working on the deployment of OpenStack on a 12-node cluster and CUDA profiling on NVIDIA GPUs. He was an intern at Argonne National Laboratory in 2014. He participated in the SCC competition at SC14 working on system administration and ADCIRC. His duties in SCC will include system administration and specializing in the WRF application. He has attended SC13 and SC14.

Alexander Ballmer is a 1st year student at IIT. He is a CAMRAS scholar with a full ride scholarship. He has worked in the DataSys lab since June 2014 working on SCC related activities from SC14. He will be an intern at Argonne National Laboratory in 2015. He participated in the SCC competition at SC14 as a team member. His duties in SCC will include specializing in the HPC Repast application. He has attended SC14. His interests in research are distributed systems, HPC file systems, and peer-to-peer networking. Other interests include building R/C aircraft, model rocketry, and filmmaking.

Andrei Georgian Dumitru is a 1st year student studying CS. He has been working in the Datasys lab since August 2014, and he helped the Cluster Competition team to tune the MATLAB application. In the last two years, he has interned with Microsoft and BitDefender, and in the summer 2015 he will be working for HERE Maps. His research interests include parallel computing and information security. His duties in SCC will include specializing in the HPL benchmark.

Adnan Haider is currently a 1st year student in CS. He currently works in the Scalable Computing Software Laboratory. His research interests include distributed computing, computer architecture optimization, and parallel network simulation. He has implemented optimizations in the Hadoop Distributed File System, developed a parallel network simulator, and has designed and implemented cache optimizations in memory simulator, GEM5. He will work at the University Corporation of Atmospheric Research during the summer, where he will look at how to use SSD's as burst buffers for climate modeling data. In addition, he has attended SC 14 with the HPC for undergrads program. He also works outside the computer science field, as the External Fundraising Chair for IIT's Engineers Without Borders Chapter.

Serapheim Dimitropoulos is a 4th year student in CS at IIT. He is interested in high-performance accelerators and distributed operating systems. He has relevant research experience with Intel Xeon Phi accelerators, SCIF, OpenMP, and OpenCL. His duties in SCC will include tuning and porting applications to the Xeon Phi accelerators.

Ariel Young is a 2nd year student in CS at IIT. She is interested in distributed systems and big data. Her duties in SCC will include visualizations of the scientific output as well as visualizations of the system monitoring.

William Scullin is a member of the Catalyst team at the Argonne Leadership Computing Facility. William is a computational generalist who enables scientific discovery through systems and software engineering at scale. With a strong background in both systems administration and computational science, he helps the team find resources and guides their research.

Ben Allen is the lead administrator for the Joint Laboratory for Systems Evaluation at the Argonne Leadership Computing Facility. He designs and maintains an environment that serves as a testbed for new and interesting hardware and software and supports HPC application and performance engineers using it. Ben has been key in helping the team realize their hardware plans in a safe and sane manner.

Dr. Ioan Raicu is an assistant professor in CS at IIT, as well as a guest research faculty in MCS at ANL. He received his PhD from UChicago, and has worked at NASA and Northwestern prior to IIT. He is the recipient of the NSF/CRA CIFellowship and the NSF CAREER award. His research work and interests are in distributed systems, emphasizing large-scale resource management in supercomputing, cloud computing, and many-core computing.

II. Why are we Participating?

Dr. Raicu has experience in mentoring 17 undergraduate students as well as 2 high-school students at IIT in the DataSys laboratory. Through research, young and diverse students have learned important skills early in their undergraduate studies (often times starting in their freshman year): teamwork, written and oral skills, computational thinking, interdisciplinary skills, and experimental skills. Dr. Raicu has strived to establish an interdisciplinary undergraduate research program with computational thinking at its core. The students were exposed to tutorials, research papers, programming languages, and distributed systems, as well as to other research environments in the Chicago area, through research meetings and presentations at Argonne, UChicago, and IIT. The undergraduate students in the DataSys laboratory have published numerous papers over the past several years [2-15].

The students from this team were hand-picked by Dr. Raicu because of their level of performance in relevant coursework, their maturity, their enthusiasm, and their technical skills. Dr. Raicu has been working in the supercomputing space since 2003, and has tackled some of the most challenging problems in resource management at extreme scales. Although much of the research performance in the DataSys laboratory is quite challenging, the undergraduate coursework in distributed systems can at times seem inadequate to keep some of the brightest undergraduate students engaged and stimulated. Dr. Raicu has spent a significant amount of time engaging the brightest undergraduate students through REU supplements to keep their thirst for knowledge and challenge adequately supplied. Furthermore, Dr. Raicu has established a new REU site called BigDataX at IIT and University of Chicago focusing on Big Data research at extreme scales (http://datasys.cs.iit.edu/grants/BigDataX/).

Having undergraduate students attend the Supercomputing/SC conference and to have the possibility to participate in the Cluster Challenge, is by far the best positive impact on their education, and on opening their eyes to the exciting field of distributed and high-performance computing. If we are successful in engaging our brightest undergraduate students, we will surely have a positive impact on their ultimate achievable potential. This is an excellent opportunity to show these students how exciting high-performance computing can be, to hopefully convince them that graduate school is not only worthwhile, but highly desirable!

III. Winning is our Middle Name

We have assembled an excellent team of undergraduate students who are passionate about distributed system, and have the necessary skills to succeed in the cluster challenge! First of all, these students are all top notch students with nearly straight A in their academic work. Furthermore, 3 of the 6 students have full ride merit-based scholarships with another 2 students having nearly full ride scholarships. Four of the students have been involved with the SCC competition at SC14. Some of these students have been involved in multiple years of research work in the DataSys laboratory, where the students were exposed not only to cutting edge research in distributed systems, scientific computing, scheduling, and storage, but also to real practical issues in running at extremely large scales at 16K-node scales on an IBM BlueGene/P supercomputer. The students have in fact been exposed to some hybrid systems as well, such as the Bluewater Cray system with NVIDIA K20 GPUs, as well as to a local cluster of 10-nodes with desktop GPUs and multiple K80 Tesla GPUs.

IV. Team Skills Diversity

The team is diverse in the sense that it includes undergrads across different years, from 1st year to 4th year. All of the students have Linux experience through research in the DataSys laboratory, internships, or coursework. All the students have been exposed to a variety of programming models such as multi-threading, OpenMP, MPI, CUDA, OpenCL, MapReduce, workflows, client/server architectures, sockets, and event-driven concurrent programming. All students have been working for many years with C/C++ as well as Java. They have all used batch schedulers (e.g. Slurm and SGE) and are proficient in bash scripting, low level OS kernel tuning for process management and network tuning, and using profiling tools to analyze performance bottlenecks and issues. They have also been exposed to a variety of clouds from Google, Microsoft, and Amazon, and are familiar with everything from user-level virtualization, to para-virtualization, to hardware-based network virtualization. They have also used both Ethernet and Infiniband networks and are familiar with advanced features that could affect network performance (e.g. frame size in Ethernet, Single Root Input/Output Virtualization SRIOV for Infiniband). Two of the students (Ben and Alex) have also attended the SC conference in 2014.

In addition to the experience that 4 of the 6 students gained by participating in SCC in SC14, these 4 students have already begun training for SCC at SC15 through a new course that William and Ioan put together in the Spring 2015 semester. The students spent the semester studying the announced applications, in doing code review of other scientific applications, and in setting up and configuring an old cluster of machines, including installing and configuring Linux, a shared file system, MPICH, HPL, and all the announced applications. Three of the students on the team will spend the summer working full time as interns at ANL or IIT towards the exploration of the large parameter space that governs the possible configurations that will yield the best performance per watt of power consumed. All 6 students have agreed to take another course in the Fall semester that is specifically designed to prepare them for the competition. Over the course of 11 months, each student would have spent over 800 hours preparing for this competition, in addition to the time 4 team members spent last year preparing for half a year in 2014!

V. Teamwork

They all know each other very well, some have known each other for years, some are roommates, and others are lab-mates. They have been working in teams (in both course-work and in research in the DataSys laboratory) already, and have already learned the importance of teamwork, good communication, and knowing each other's strengths and weaknesses. No matter on the outcome, I am 110% positive that they will have a great time, they will surely learn many things along the way, and they will come out of this competition energized and full of ideas. Working with such talented and collegial undergraduate students must be one of the best feelings a professor can have – it is just an awesome feeling to see all the hard work invested in both teaching and research pay off.

VI. Team Experience

The assembled team have built three separate 8~12 node clusters from scratch (loaded with AMD and NVIDIA GPUs, SSDs, multi-core CPUs from both AMD and Intel, and multiple Gb/s Ethernet adaptors as well as Infiniband 56Gb/sec networks), and have configured them with both a traditional HPC software stack (Slurm batch-scheduler with GPFS parallel filesystem) and with a newer cloud software stack, such as OpenStack and virtualization through XEN; they have managed these clusters for use in research and teaching activities now for several years. Some of the students from the team have been managing these clusters for the entire DataSys laboratory as well as for students taking courses in distributed systems. All the students have spent time in internships, such as Argonne National Laboratory, Nokia HERE, and NCAR, they have done research and written papers [2-15], they have attended conferences the Supercomputing/SC conferences in 2014, and have been part of major research initiatives that have developed and evaluated distributed systems at petascales levels. They are all quite proficient in all the skillset that is critical to carrying out an in-depth performance evaluation and tuning of HPC applications.

VII. Tuning and Optimizing the Application Set

We are excited for this year's application set, and believe that our hardware and skill set will contend for first place. This section describes how we plan to tune and optimize the application set to make use of our hardware.

Trinity (http://trinityrnaseq.github.io/) is a bioinformatics application that is composed of three software modules, Inchworm, Chrysalis, and Butterfly. They are applied sequentially to process large volumes of RNA sequence reads. It partitions sequence data into graphs, and then processes each graph independently. Inchworm and Chrysalis are both written in C++ while Butterfly is written in Java. Inchworm and Chrysalis are both memory intensive but the intensity depends on the data being processed. Trinity is memory intensive as are most bioinformatics applications. We plan to equip our nodes with large amounts of memory in addition to large lower level caches in order to mitigate the data movement between the last level cache and memory.

Repast HPC (http://repast.sourceforge.net/repast_hpc.html) is an agent based modeling program written in C++. The Repast program itself is a framework for controlling user written "agents." as the simulation runs, Repast directs the agents to interact with one another in various ways, like creating network links, changing state, or otherwise storing or exchanging information. Data collected by agents is cached in memory and then funneled by MPI to a centralized output file at the end of the simulation or checkpoint. Data output has the option of using netCDF for better disk performance. Possible performance characteristics include the lack of heavy double precision usage, which would mean that Repast will probably not be as CPU intensive as other applications. Repast (like all other agent based modelers) will not run on GPUs, but we hope we can port the code to the Xeon Phi given the MIMD architecture.

The Weather Research and Forecasting model (WRF) is a next-generation mesoscale numerical weather prediction system designed to serve both atmospheric research and operational forecasting needs (http://www.wrf-model.org/index.php). WRF has two main parts: data pre and post-processing, and the WRF model solver. The data processing sections will most likely run best in serial or in parallel with few cores, since it is not very computationally intensive. The model solver involves hefty parallel computation. The WRF model solver can run in for different modes: serial, shared memory parallel, distributed memory parallel, and hybrid. The hybrid mode combines distributed and shared memory parallelism using MPI and OpenMP. This will give us great flexibility on how to run WRF on our CPUs and the Xeon Phi accelerators.

MILC (http://www.physics.utah.edu/~detar/milc/) is a set of codes written in C developed by the MIMD Lattice Computation (MILC) collaboration for doing simulations of four dimensional SU(3) lattice gauge theory on MIMD parallel machines. From preliminary investigation, it seems that there is some support for OpenMP in the MILC codebase, leading us to be optimistic about the potential of running MILC on the Xeon Phi accelerators.

We believe the proposed testbed with its x86 architecture on both CPUs and accelerators have a good chance to work across a diverse set of applications, including the mystery application. The team is prepared to write OpenMP code for applications that don't support it natively.

VIII. General Overview of Proposed Cluster

We propose a cluster configuration leveraging five 4U nodes (using 4 sockets per node with Intel Xeon E7-4880 v3 processors with 18-cores at 2.3GHz, and 4 Intel Knights Landing Xeon Phi accelerators per node) connected by 4 Infiniband (IB) 100Gb/sec full duplex ports per node and creating a direct interconnect between the 5 nodes while eliminating a power hungry IB switch. We estimate that this 20U half rack will deliver approximately 75 double precision TFlops/sec with less than 6200 watts of power, and that MIC or CPU only usage of the system will result in 60TFlops/sec (from 20 Xeon Phi accelerators, rated at 3Tflops/sec in a 150 watt envelope) or 15TFlops/sec (20 Xeon E7-4880 v3 CPUs, rated at 0.75Tflops/sec in a 145 watts envelope) in less than 3120 watts respectively. More details can be found in Table 1. The latest Knights Landing Xeon Phi accelerators have a flop/watt advantage over both NVIDIA and AMD GPUs as well as CPUs, and their x86 programmability makes them a more likely candidate to be useful for a wide range of real applications (beyond Linpack).

Table 1: Summary of proposed cluster hardware

	Description	**Aggregate over a 5 node system**
CPU	Quad Socket with Intel Xeon E7-4880v3 CPUs	720 HT over 360 Intel x86 cores at 2.3GHz delivering 15TFlops/sec
Memory	512GB DDR4 RAM (32x16GB)	2.5TB RAM delivering 1.7TB/sec bandwidth
Accel.	Quad Xeon Phi Knights Landing accelerators	5760 HT over 1440 Intel x86 cores at 1.3GHz delivering 60TFlops/sec, coupled with 320GB DDR4 RAM @ 8TB/sec
Storage	1TB SSD storage per node	5TB SSD delivering 10GB/sec of persistent disk storage I/O
Network	Direct switchless Infiniband network with 400Gb/sec bandwidth per node	800Gb/sec bisection bandwidth, avoiding about 200 watts of power consumption by skipping the switch
Power	Maximum peak power draw of all hardware per node is 1200 watts	Operating nodes with only MICs or CPUs, will results a total power drat less than 3120 watts

In last years competition, we had an IB switch that consummed about 100 watts of power under load, and it caused an additional 100 watts of power inbalance. In order to reduce this 200 watts power loss, we decided to create a direct mesh interconnect with 4 IB ports per node, and direct cables from every node to every other node. See

Figure 1 for the pictorial represenation. This not only reduces power consumption, but will also deliver 7X the bandwidth compared to the more traditional siwtch based 56Gb/sec IB topology, and it should also reduce network latency for small packets.

Admittedly our approach is quite different than the traditional winner of the competition, in the sense that we are using the power efficient Intel Xeon Phi accelerators, as well as bypassing the switch saving about 4% of power plus another 4% due to the load imbalance for a total of 8% power savings. We further expect to reduce power consumption by replacing stock air cooling solutions with water-cooled ones, which we expect to save an additional 10% of power. The Intel Xeon Phi Knights Landing are also 2X more power efficient than the best GPUs from AMD and NVIDIA, and the high-end Intel processors we plan on using are also 10%~20% more power efficient than the more mainstream processors with fewer cores. There are many savings we can have with the proposed hardware, allowing us to pack more than 2X the hardware than a typical setup in the power envelope of 3120 watts. The IIT team together with Argonne will drive the proposed cluster to victory in both the modified Linpack and the overall performance categories in the Standard Track at SC15.

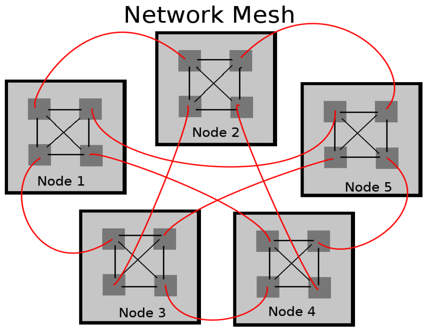

Figure 1: Network topology overview

IX. Demonstrations to Impress

Dr. Raicu firmly believes in live visualizations as one of the best mechanisms to understand the performance and bottlenecks of an application. We will use a combination of monitoring tools, such as the Darshan project [16] being developed at Argonne National Laboratory. We will also leverage the innovative work in distributed filesystems FusionFS [17-21] to outperform more traditional HPC filesystems such as PVFS or GPFS. Dr. Raicu has also made much progress in the design and implementation of distributed key/value storage systems (ZHT [2]) which might come in valuable to further accelerate the respective HPC applications. Much of the research happening in the DataSys laboratory could be put to the test in accelerating these HPC applications.

We also built our own custom system status display last year from LED strips – a unique feature of our rack that earned a lot of attention from SC14 attendees. IIT has a first rate architecture school and we intend to build on what we learned from last year's status display, suggestions from architecture faculty and students, and the industrial design of historical systems to produce a system that will be visually engaging and functional. We believe this will help put us apart from other team's booths and highlight what we can do on our own.

X. Institutional Commitment

The DataSys laboratory at IIT has access to a cluster of workstations composed of 25 nodes with 220-cores, 614GB of memory, NVIDIA GPUs, SSDs, and 1Gb/s Ethernet network. The SCS laboratory at IIT (where Dr. Raicu is also a member of) has an 85 node Sun Microsystems ComputeFarm which has been used extensively in both research and teaching. The team also has access to the cluster that was build for SCC @ SC14, which consists of 8-nodes with dual socket motherboards and Intel Xeon E5-2699v3 processors (same ones we plan to use this year at SC15), NVIDIA K40 GPUs, and a dual port 56Gb/sec IB network.

To conduct state-of-the-art research in distributed systems, Dr. Raicu has also been active in writing proposals to get access to some of the largest supercomputers in the world. He has written multiple proposals to get access to a variety of systems. These awards have summed up to over 7M CPU hours. Some of the systems Dr. Raicu and his students have access to, with ranking in the Top500 in November 2014 [1], are: Titan@ORNL (Top500 #2), Mira@ANL (Top500 #5), Stampede@TACC (Top500 #7), Bluewaters@NCSA (likely top 10), Cielo@LANL (Top500 #40), and Gaea@ORNL (Top500 #87).

Dr. Raicu enjoys teaching and works hard to make the courses interesting, current, useful, and engaging. Many students who take his classes generally translate the course knowledge into real world skills that helps students find jobs. His courses are often a mixture of theory (of distributed systems) and practice (real implementations of distributed systems). He has taught 3 grad-courses (Advanced Operating Systems, Cloud Computing, and Data-Intensive Computing) and an undergrad-course (Intro to Parallel and Distributed Computing). These were all new courses he designed and taught since 2011, and all courses covered in depth different aspects of distributed and high-performance computing.

XI. Sponsorship Support

ANL is in the early stages of confirming sponsorship from Intel and Mellanox, which both supported the team in 2014. This work is also supported in part by NSF award OCI-1054974 and REU supplements. This work will also use resources of the ALCF at ANL, which is supported by the DOE contract DE-AC02-06CH11357.

References

[1] Top500, November 2014; http://www.top500.org/lists/2014/11/; 2014
[2] Tonglin Li, et al. "ZHT: A Light-weight Reliable Persistent Dynamic Scalable Zero-hop Distributed Hash Table", IEEE IPDPS 2013
[3] Ke Wang, et al. "SimMatrix: Simulator for MAny-Task computing execution fabRIc at eXascales", ACM HPC 2013
[4] Tonglin Li, et al. "Distributed Kev-Value Store on HPC and Cloud Systems", GCASR 2013
[5] Kevin Brandstatter, et al. "NoVoHT: a Lightweight Dynamic Persistent NoSQL Key/Value Store", GCASR 2013
[6] Ke Wang, et al. "Paving the Road to Exascale with Many-Task Computing", Doctoral Showcase, IEEE/ACM Supercomputing/SC 2012
[7] Tonglin Li, et al. "Exploring Distributed Hash Tables in High-End Computing", ACM Performance Evaluation Review (PER), 2011
[8] Ke Wang, et al. "SimMatrix: Simulator for MAny-Task computing execution fabRIc at eXascales", GCASR 2012
[9] Scott J. Krieder, et al. "Design and Evaluation of the GeMTC Framework for GPU-enabled Many-Task Computing", ACM HPDC14
[10] Benjamin Grimmer, et al. "Enabling Dynamic Memory Management Support for MTC on NVIDIA GPUs", EuroSys 2013
[11] Scott Krieder, et al. "Early Experiences in running Many-Task Computing workloads on GPGPUs", XSEDE 2012
[12] Dustin Shahidehpour, et al. "Accelerating Scientific Workflow Applications with GPUs", GCASR 2013
[13] Ben Grimmer, et al. "Enabling Dynamic Memory Management Support for MTC on NVIDIA GPUs", GCASR 2013
[14] Scott J. Krieder, et al. "Towards Efficient Many-Task Computing on Accelerators in High-End Computing Systems", GCASR 2013
[15] Jeff Johnson, et al. "Understanding the Costs of Many-Task Computing Workloads on Intel Xeon Phi Coprocessors", GCASR 2013
[16] Darhsan, http://www.mcs.anl.gov/research/projects/darshan/, 2014
[17] Dongfang Zhao, et al. "HyCache+: Towards Scalable High-Performance Caching Middleware for Parallel File Systems", IEEE CCGrid 2014
[18] Dongfang Zhao, Ioan Raicu. "HyCache: A User-Level Caching Middleware for Distributed File Systems", IEEE HPDIC 2013
[19] Dongfang Zhao, et al. "Improving the I/O Throughput for Data-Intensive Scientific Applications with Efficient Compression Mechanisms", IEEE/ACM Supercomputing 2013
[20] Dongfang Zhao, Ioan Raicu. "Distributed File Systems for Exascale Computing", Doctoral Showcase, IEEE/ACM Supercomputing/SC 2012
[21] Dongfang Zhao, et al. "Distributed Data Provenance for Large-Scale Data-Intensive Computing", IEEE Cluster 2013