Delivering 3.5 Double Precision GFlops/Watt and 200Gb/sec Bi-Section Bandwidth with Intel Xeon Phi-based Cisco Servers

Kevin Brandstatter[*], Jason DiBabbo[*], Daniel Gordon[*], Ben Walters[*], Alex Ballmer[*], Lauren Ribordy[#], Ioan Raicu[*‡]

[*]Department of Computer Science, Illinois Institute of Technology, Chicago IL, USA
[†]Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago IL, USA
[‡]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne IL, USA
[#]Glenbrook South High School, Glenview IL, USA

{kjbrandstatter,jdibabbo)@gmail.com, {dgordon5,bwalter4}@hawk.iit.edu, alexandersballmer@gmail.com, 156614@gmail.com, iraicu@cs.iit.edu

***Abstract*—This document is the team proposal for the SC14 Student Cluster Competition Standard Track. The team consists of 5 undergraduate students from the Illinois Institute of Technology (IIT), a high-school student from Glenbrook South High-School (GBS), a faculty advisor with a joint appointment between IIT and Argonne National Laboratory, and sponsorship from Cisco. We propose a cluster configuration leveraging four to six Cisco servers (using power efficient Haswell processors) connected by a dual 40Gb/s full-duplex Ethernet network, each with two Intel Xeon Phi coprocessor 5110P (with 600 x86 cores delivering an aggregate of 10 double precision TFlops/sec), achieving a competitive 3.5 DP GFlops/watt. Our configuration should best other systems equipped with GPU accelerators from NVIDIA and AMD due to the x86 architecture of the Xeon Phi making obtaining flops ratings close to theoretical peaks easier to obtain. The achieved power efficiency of the proposed cluster is between 1.16X and 4X better than the best power efficiency achieved in 2013 for double precision operations, arguably one of the most important metrics for scientific applications. The IIT team together with Cisco will drive the proposed cluster to victory in the Standard Track at SC14.**

I. Team Members

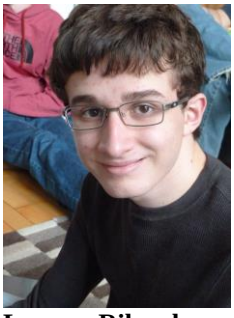
The team consists of 5 undergrads from IIT (Brandstatter, DiBabbo, Gordon, Walters, and Ballmer) ranging from freshmen to seniors, a high-school female student (Ribordy) from GBS, and a CS faculty advisor (Dr. Raicu).

Kevin Brandstatter is a 4th year undergraduate student in CS at IIT, as well as a research assistant in the DataSys Lab for 3 years. He is a CAMRAS scholar with a full ride scholarship. He has interned at Accenture Technology Labs and the Max Plank Institute. His research work focuses on distributed storage applications such as file systems, fault tolerance, and key/value stores. He attended SC12.

Jason DiBabbo is a 4th year CS major at IIT. He is the recipient of the Collens Scholarship, a nearly full-ride award. He has interned at Commonwealth Edison and Coyote Logistics, LLC as a software engineering intern, and will be interning at Microsoft in the summer of 2014 as a software development engineer in the Cloud and Enterprise department. His passions include enterprise software solutions and distributed computing.

Daniel Gordon is a 4th year majoring in CS with a specialization in distributed and cloud computing. He assisted in developing a dynamic volatility calculator at an internship with the Chicago Mercantile Exchange. He is currently working on an iOS development application, and distributed applications using AWS and Hadoop, and will be interning with Nokia over the summer 2014.

Ben Walters is a 1st 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. His current research project involves CUDA profiling on NVIDIA GPUs. He attended SC13 in Denver.



Alexander Ballmer is going to be a freshman at Illinois Institute of Technology in the Fall 2014 semester. His hobbies include Linux and distributed systems. He is a CAMRAS scholar with a full ride scholarship.

Lauren Ribordy spent the summer 2012 at the Illinois Institute of Technology (IIT) and learned the basics of Java and cloud computing. She constructed a word count program and performed latency and bandwidth tests on the Amazon AWS

cloud. She presented her work and results at both DePaul University and IIT. While at IIT, she realized her interest in human computer interaction (HCI), finding that it combined her creativity and her love of computers. She is interested in the development and design of operating systems and user interfaces.

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 6 undergraduate students at IIT in the DataSys laboratory. Through research, young and diverse students have learned important skills early in their undergraduate studies (often times during 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 [4]-[18].

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 for over a decade, 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.

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, 2 of the 6 students have full ride merit-based scholarships with another 2 students having nearly full ride scholarships. Two of the students have been involved with research in the DataSys laboratory spanning multiple years, 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 one K20 Tesla GPU.

IV. TEAM SKILLS DIVERSITY

The team is diverse in the sense that it includes undergrads across different years, from 1st year, to 3rd year and 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. Some of the students (e.g. Kevin) have significant experience in developing distributed storage systems at large scale (he was involved in developing and evaluating distributed storage systems on an IBM BlueGene/P supercomputer at up to 16K-node scales). 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 Kevin) have also attended the SC conference in 2012 and 2013 (funded by Dr. Raicu's NSF CAREER award), and they had followed the cluster competition closely.

V. TEAMWORK

They all know each other very well, some have known each other for years, some are room-mates, 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 team work, 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 two separate 10-node clusters from scratch (loaded with NVIDIA GPUs, SSDs, multi-core CPUs, and multiple Gb/s Ethernet adaptors), and have configured them with both a traditional HPC software stack (batch-scheduler with parallel filesystems) 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 at the Max Plank Institute and Accenture Technology Laboratory, they have done research and written papers [4]-[11], they have attended conferences the Supercomputing/SC conferences in 2012 and 2013, and have been part of major research initiatives that have developed and evaluated distributed storage systems at petascale and beyond on up to 16K-node scales. 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.

1) NAMD is a parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems.[24, 28] In addition, we've already identified research projects based on providing Xeon Phi acceleration for NAMD. Our team already has some application experience with Molecular Dynamics simulations [23] and we expect to leverage OpenMP or OpenCL to execute MD simulations on both the CPUs and Xeon Phi accelerators of our cluster.

2) MATLAB is a numerical computing environment for performing matrix manipulations and visualizing data. Again, we expect to leverage the Xeon Phi accelerators to offload matrix computations for the fastest possible time to completion. Our research shows that there are existing tools for running OpenCL with MATLAB. [25]

3) *ADCIRC* is a system of programs for solving free surface circulation and transport problems. We've identified research projects that have enabled ADCIRC applications to run on accelerators [26] and plan to leverage them with our testbed.

We believe the proposed testbed with its x86 architecture on both CPUs and accelerators have a good chance to seamlessly work across a diverse set of applications, including the mystery application. In conclusion, we expect this suite of applications to be excellent candidates for our Intel Xeon Phi accelerators, and look forward to the mystery application.

VIII. General Overview of Proposed Cluster

We propose a cluster configuration leveraging five Cisco servers (using power efficient Haswell processors) connected by a dual 40Gb/s full-duplex Ethernet network, each with two Intel Xeon Phi coprocessor 5110P (with 600 x86 cores delivering an aggregate of 10 double precision TFlops/sec), achieving a competitive 3.5 DP GFlops/watt. To put this into perspective, the most power efficient system in the Top500 (November 2013) [1] is ranked #311 in performance, and achieves 3.4 double precision GFlops/Watt. The most power efficient system in the Green500 (November 2013) list [2] achieves an improved 4.5 double precision GFlops/Watt. The winning team for the Standard Track Cluster Challenge at SC13 was able to achieve less than 3 double precision GFlops/Watt (total of less than 9 double precision TFlops across 8-nodes with NVIDIA K20c GPUs). The power efficiency of the proposed cluster is 1.16X to 4X better than the best power-efficiency achieved in 2013.

The Cisco servers each contain Haswell low power processors (with as low as 50watt envelope) and 2 Intel Xeon Phi 5110P coprocessor providing over 1 double precision TFlops/sec. In addition, the Cisco servers support networking over dual 40Gb/sec Ethernet ports, configured as a multi-rail scenario to aggregate the bandwidth across both ports into a seamless 80Gb/sec network connection (full-duplex). Open MPI can transparently bond all available local interfaces and stripe large messages across both ports. The five servers and 10 network ports will be interconnected by 10 ports running at 40Gb/sec each, for a bi-section bandwidth of 200Gb/sec.

Table 1: Summary of proposed cluster hardware

CPU	Low Power Intel Xeon Haswell Processor
Memory	64GB DDR3 (4x16GB)
Accelerator	Dual Intel Xeon Phi 5110P
Storage	1TB SSD storage
Network	Dual 40Gb/sec Ethernet
Power	Maximum peak power draw of 600w

While admittedly our approach is a more *off the wall solution*, we expect to not only keep up with the competition but we expect to win. While historically many teams have opted for high-end accelerator based solutions compressed into rack mounted units, our solution provides 10 double precision TFlops of performance and consumes a maximum of 3,000 watts. The estimated cluster costs are $80K, including all hardware, cables, and network ➔ costing $8 per GFlop. With an excellent Flops/Watts, an x86 architecture (across both processors and accelerators), and an excellent bi-section bandwidth, we expect to drive our systems at near peak performance and have an excellent chance at winning the standard track.

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 [19] being developed at Argonne National Laboratory. We will also leverage the innovative work in distributed filesystems FusionFS [20][21][22] to outperform more traditional HPC

filesystems such as PVFS or clustered filesystems such as NFS. Dr. Raicu has also made much progress in the design and implementation of distributed key/value storage systems (ZHT [4]) 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.

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. Dr. Raicu also has access to pre-production Intel Xeon Phi accelerators at Argonne National Laboratory.

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 are: Titan@ORNL (Top500 #2), Mira@ANL (Top500 #5), Stampede@TACC (Top500 #6), Bluewaters@NCSA (likely top 10), Cielo@LANL (Top500 #22), Kraken@NICS/UT (Top500 #30), Gaea@ORNL (Top500 #48), and Keeneland@GeargiaTech (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

Dr. Raicu has secured $4K in funding from the university to help with travel for the undergraduate team. The estimated cost of the proposed cluster is about $80K. Dr. Raicu's team has partnered with Cisco to provide all the needed hardware to prepare for the competition and to compete in the final challenge. The team will be given remote access to Cisco servers comparable to the ones to be used in the competition throughout the summer months, while multiple servers are procured and shipped to Illinois Institute of Technology to explore hands-on. Dr. Raicu will be visiting Ven Immani from Cisco on May 23rd for a site visit, and Jeff Squyres from Cisco will be visiting the team at IIT on June 6th. These visits are in addition to regular virtual meetings to ensure adequate progress is being made, and that we have access to any technical information we might require. Several pre-production servers will be sent to IIT to help prepare for the challenge. Part of the preparation involves getting the challenge applications running, tuning them, and measure the actual power consumption during the application runs. This is critical to ensure we have the best possible setup at the challenge in November.

Acknowledgements

This work is sponsored by Cisco, we are grateful to their generosity that has made this competition possible. We also want to thank Ven Immani and Jeff Squyres from Cisco for their useful feedback on the proposed cluster. This work was supported in part by the National Science Foundation under awards OCI-1054974 (CAREER). This research will use resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under contract DE-AC02-06CH11357.

References

[1] Top500, November 2013; http://www.top500.org/lists/2013/11/; 2014
[2] Green500, November 2013 List; http://www.green500.org/lists/green201311; 2014
[3] Mac Pro Specifications, https://www.apple.com/mac-pro/specs/; 2014
[4] Tonglin Li, et al. "ZHT: A Light-weight Reliable Persistent Dynamic Scalable Zero-hop Distributed Hash Table", IEEE IPDPS 2013
[5] Ke Wang, et al. "SimMatrix: Simulator for MAny-Task computing execution fabRIc at eXascales", ACM HPC 2013
[6] Tonglin Li, et al. "Distributed Kev-Value Store on HPC and Cloud Systems", GCASR 2013
[7] Kevin Brandstatter, et al. "NoVoHT: a Lightweight Dynamic Persistent NoSQL Key/Value Store", GCASR 2013
[8] Ke Wang, et al. "Paving the Road to Exascale with Many-Task Computing", Doctoral Showcase, IEEE/ACM Supercomputing/SC 2012
[9] Kevin Brandstatter, Ioan Raicu. "CiteSearcher: A Google Scholar frontend for Mobile Devices", IIT Research Day, 2012
[10] Tonglin Li, et al. "ZHT: a Zero-hop DHT for High-End Computing Environment", GCASR 2012
[11] Ke Wang, et al. "SimMatrix: Simulator for MAny-Task computing execution fabRIc at eXascales", GCASR 2012
[12] Scott J. Krieder, et al. "Design and Evaluation of the GeMTC Framework for GPU-enabled Many-Task Computing", ACM HPDC14
[13] Benjamin Grimmer, et al. "Enabling Dynamic Memory Management Support for MTC on NVIDIA GPUs", EuroSys 2013
[14] Scott Krieder, et al. "Early Experiences in running Many-Task Computing workloads on GPGPUs", XSEDE 2012
[15] Dustin Shahidehpour, et al. "Accelerating Scientific Workflow Applications with GPUs", GCASR 2013
[16] Ben Grimmer, et al. "Enabling Dynamic Memory Management Support for MTC on NVIDIA GPUs", GCASR 2013
[17] Scott J. Krieder, et al. "Towards Efficient Many-Task Computing on Accelerators in High-End Computing Systems", GCASR 2013
[18] Jeff Johnson, et al. "Understanding the Costs of Many-Task Computing Workloads on Intel Xeon Phi Coprocessors", GCASR 2013
[19] Darhsan, http://www.mcs.anl.gov/research/projects/darshan/, 2014
[20] Dongfang Zhao, et al. "HyCache+: Towards Scalable High-Performance Caching Middleware for Parallel File Systems", IEEE CCGrid 2014
[21] Dongfang Zhao, et al. "Improving the I/O Throughput for Data-Intensive Scientific Applications with Efficient Compression Mechanisms", IEEE/ACM Supercomputing 2013
[22] Dongfang Zhao, Ioan Raicu. "Distributed File Systems for Exascale Computing", Doctoral Showcase, IEEE/ACM Supercomputing/SC 2012
[23] Scott J. Krieder, et al. "Design and Evaluation of the GeMTC Framework for GPU-enabled Many-Task Computing", ACM HPDC14
[24] NAMD, http://www.ks.uiuc.edu/Research/namd/, 2014
[25] OpenCL, https://code.google.com/p/opencl-toolbox/, 2014
[26] http://www.sciencedirect.com/science/article/pii/S187705091100278X, 2014