Maximizing Computation per Power Ratios in High-Performance Computing: from Aggressive Power Management to Approximate Computing

Ioan Raicu[*‡+], William Scullin[^], Ben Allen[^], Kyle Hale[*], Kyle Chard[‡#], Simone Campanoni[+]

[*]Department of Computer Science, Illinois Institute of Technology, Chicago IL, USA
[‡]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne IL, USA
[^]Advanced Leadership Computing Facility, Argonne National Laboratory, Argonne IL, USA
[#]Computation Institute, University of Chicago, Chicago IL, USA
[+]Electrical Engineering and Computer Science, Northwestern University, Evanston IL, USA

iraicu@cs.iit.edu, {wscullin,bsallen}@alcf.anl.gov, khale@cs.iit.edu, chard@uchicago.edu, simonec@eecs.northwestern.edu

***Abstract*—** *To achieve the next level of performance from large scale scientific computing, we have been exploring ways to optimize cluster utilities and applications 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), water cooling, eliminating fans, powering down of accelerators when not in use, and compiler level optimizations. The Chicago Fusion Team is composed of six mentors and eleven students from various Chicago area institutions, such as Illinois Institute of Technology (IIT), University of Chicago (UChicago), Northwestern University (NU), University of Illinois at Chicago (UIC), and Argonne National Laboratory (ANL). We propose a cluster configuration leveraging nine nodes and a switch, enclosed in a 9U rack. The rack will be water cooled, and contain Intel Xeon processors, Intel Xeon Phi processors, Intel FPGAs, a NVIDIA GPU (for visualization), Intel Optane NVMe SSDs, and Omni-Path network. We estimate that this 9U rack will deliver approximately 54 double precision TFlops/sec with about 5kW of power. With aggressive power management and compiler optimizations, this hybrid cluster will fit in the 3kW power envelope and due to its variety of hardware, we expect it will be extremely flexible to efficiently support a wide range of real applications (beyond Linpack). Sponsorship will be from Intel and Argonne National Laboratory.*

I. Chicago Fusion Team

A. Mentors

The new Chicago Fusion Team has 6 mentors from various Chicago area institutions, such as Illinois Institute of Technology (IIT), University of Chicago (UChicago), Northwestern University (Northwestern), and Argonne National Laboratory (ANL); these mentors include faculty members who work in scientific computing, data science, resource management at extreme scales, distributed storage systems and run-time systems, parallel languages, compilers, operating systems, and virtual machines, as well as system administrators who run the Advanced Leadership Computing Facilities (ALCF) at ANL.

Ioan Raicu is an Associate Professor in CS at IIT, a guest research faculty in MCS at ANL, as well as a guest faculty in EECS at Northwestern. He received his PhD from UChicago in 2009. He is the recipient of the NSF/CRA CIFellow and the NSF CAREER awards. His research interests are in distributed systems, emphasizing large-scale resource management in supercomputing, cloud computing, and many-core computing. He will serve as the team primary advisor.

William Scullin is a member of the Catalyst team at the ALCF at ANL. 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 ALCF at ANL. 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.

Kyle C. Hale is an Assistant Professor in CS at IIT. His research interests lie at the intersection of operating systems, parallel computing, and computer architecture, with a focus on building experimental systems. His work ranges from Networks-on-Chip to specialized operating systems and parallel languages. He earned his Ph.D. in CS from Northwestern University in 2016. He will help tune the OS parameters for best performance, as well as explore the potential use of research light-weight OS for the Intel Xeon Phi.

Kyle Chard is a Senior Researcher and Fellow in the Computation Institute at UChicago and ANL. His research focuses on applying computational and data-intensive approaches to solve scientific problems. His background has primarily focused on distributed systems, related to scheduling of resources in Grid and Cloud environments. He received his PhD from Victoria University of Wellington in 2011. He will mentor students in data-management related topics.

Simone Campanoni is an Assistant Professor in EECS at Northwestern. Simone's main research areas are compilers and virtual machines, with special interest in computer architecture, runtime systems, operating systems, and programming languages. Simone

addresses research challenges through vertical specialization of the hardware/software stack. Simone received his Ph.D. degree with highest honors from Politecnico di Milano University in 2009. He will mentor students in compiler optimizations.

B. Students

Through an open call for participation, the mentors have identified 11 students from the greater Chicago area (from an initial pool of 21 students interested), ranging from freshmen in high school to junior undergraduate students, and they include both male and female students; the students represent 5 different institutions, from IIT, UChicago, UIC, and two different high schools. The students are ordered below by seniority.

Alexander Ballmer (IIT) is a junior in CS at IIT. He participated in the IIT's SCC team in 2014 and 2015. He is a CAMRAS scholar with a full ride scholarship. He has worked in the DataSys lab since June 2014. He has interned at Argonne National Laboratory in 2015. He has attended the SC conference in 2014, 2015, and 2016. 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.

Khashkhuu Otgontulga (IIT) is a junior in CS at IIT. He is the recipient of the Phi Theta Kappa scholarship and United Minds Inspiring Innovation scholarship. He is interested in distributed systems and artificial intelligence. In his free time, he plays basketball.

Nathan Seitz (UIC) is a Junior in CS at UIC. He has participated in the Midwestern Robotics Design Competition. He is interested in distributed systems in general.

Andrew Tolentino (IIT) is a sophomore in CS at IIT. He is interested in distributed systems and its application to real world dilemmas. In his free time, he enjoys volunteering at the Greater Chicago Food Depository, repackaging and delivering food for the hungry.

Min Seok Choi (IIT) is a sophomore in CS at IIT. He is the recipient of the International Scholarship and Leaders in Science & Tech Scholarship. He is interested in distributed systems, big data, and machine learning.

Shahzil Sheikh (IIT) is a sophomore in CS at IIT. He enjoys participating in hackathons. He is interested in distributed systems in general.

Hasan Rizvi (IIT) is a freshman in CS and Applied Math minor at IIT. He is an avid programmer and an absolute Math enthusiast who loves competitions which is reflected by the fact that he is a National Math Olympiad finalist. He also regularly attends weekend long hackathons in search of new roles and skills, and to meet smart people.

Rohan Kumar (UChicago) is a freshman at UChicago. He has participated in National Programming Competitions in India (geared towards IOI) and the Intel Ideathon, an Arduino-based Inter-School Competition. He is interested in distributed systems in general.

Bhoj Rani Soopal (IIT) is a freshman at IIT. She is an enthusiast learner and she is passionate for coding.

Ryan Prendergast (high-school) is a junior at Maine South High School in Park Ridge IL. He is interested in parallel and distributed computing.

Alexander Bernat (high-school) is a freshman at Rochelle Zell Jewish High School in Deerfield IL. He is a member of the golf and math teams and is a founding member of his high school's robotics team. He enjoys building computers and is a licensed Ham radio operator. He is also an avid drone builder and pilot and his other interests include cybersecurity and networking.

C. Prior Participation

IIT had two teams compete in 2014 [4, 7] and 2015 [5, 6, 8] when 5 IIT undergrads and 1 high school student participated in the student cluster challenge. The IIT team came in 9th (from 12 teams in 2014) and 4th (from 9 teams in 2015). With all the lessons learned from the prior two competitions, the new Chicago Fusion Team that spans 5 institutions aims for 1st place. Except for IIT, none of the other institutions have had teams compete in SCC in the past.

Figure 1: IIT SCC Team at IEEE/ACM SC 2015 [5, 6, 8]

Some of the lessons we learned from our prior competitions are:

1. Having the entire student team working on preparing for the SCC competition during the summer is critical due to the likely avalanche of student commitments when the Fall semester/quarters begin; in the past, we typically had 1 or 2 students work on SCC related training, but the team would only start working together at the start of the Fall semester
2. Having the actual hardware from the vendors available to the students starting as early as the summer is critical to intimately understanding the hardware and how to tune it; in our past two competitions, we only had the hardware for about 1 week prior to the competition
3. Having additional mentors (increased from 3 to 6) specializing in various parts of the HPC software stack is important to best tune these applications on a variety of different hardware

II. Hardware and Software

A. Hardware

We propose a cluster configuration leveraging nine nodes and a switch, enclosed in a 9U rack. The nodes will be separated into 8 nodes (in two 2U chassis) with one Intel Xeon Phi Processor 7290 (72 cores @ 1.5 GHz - 245W TDP) per node, and 192GB RAM, and Omni-Path 100 interconnect. The 9th node will be a multi-purpose node, that will serve as the head node, resource manager node, storage node, and accelerator node loaded with high-end consumer NVIDIA GPUs (for visualization) and Intel FPGA accelerators; it will be encased in a 4U chassis, and have 2 sockets using the Intel® Xeon® Processor E5-2699 v4 (22 cores @ 2.2 GHz – 145W TPD), 512GB of RAM, 8TB NVMe SSD, eight Intel Nallatech 510T FPGAs (100W TPD), and one Omni-Path 100 network adaptor. We estimate that this 9U rack will deliver approximately 54 double precision TFlops/sec with about 5kW of power.

Using only the Xeon Phi resources or the FPGA resources will easily keep the power envelope under 3120 watts. We will use aggressive power management techniques as well as research-rooted compiler optimizations to reach these targets while using as much of the hardware as possible. We will use water cooling to reduce cooling power costs, and we will use redundant power supplies to load balance power consumption across the two PDUs. The hybrid cluster will have the latest Knights Landing Xeon Phi accelerators, some general-purpose Intel Xeon processors, Intel FPGAs, and a consumer grade NVIDIA GPU for visualization, making the system extremely flexible and useful for a wide range of real applications (beyond Linpack).

Table 1: Summary of proposed cluster hardware

	Description	Aggregate over a 9-node system
Computing	8x Intel Xeon Phi Processor 7290; 2x Intel® Xeon® Processor E5-2699 v4; Nallatech 510T Cards with 16x Altera/Intel Arria 10 FPGAs, 32GB RAM	2304 HT over 576-cores @ 1.5GHz and 128GB 3D MCDRAM => 27 TFlops/sec DP; 44-cores @ 2.2GHz => 2 TFlops/sec DP; 16x FPGAs 512GB RAM => 24 TFlops/sec DP
Visualization	1x NVIDIA GeForce GTX Titan XP (12GB RAM, 12TFlops SP, 0.4TFlops DP)	This high-end consumer grade GPU should give us enough visualization power for lively interactive demos
Memory	192GB DDR4 RAM (6x32GB) on Phi nodes and 512GB DDR4 RAM (16x32GB) on Xeon node	2TB RAM achieving 1TB/sec aggregate I/O
Storage	375GB Intel NVMe DC P4800X SSD storage per node	3.75TB SSD => 20GB/sec I/O using GPFS parallel file system
Network	Intel® Omni-Path Edge Switch 100 Series (24-ports); Intel® Omni-Path Host Fabric Interface	900Gb/sec bisection bandwidth in a total of 300-watt envelope (switch and NICs combined)
Power	Maximum peak power draw of all hardware per node is 420 watts for the Phi nodes and 1600 watts for the head node	Total power for the 9-node cluster should be 5139 watts; we expect not all hardware to be used to capacity at any given point in time

B. Software

The software stack includes the use of Linux CentOS 7, Warewulf (cluster management), Slurm (job management), IBM GPFS (parallel file system), Intel compiler, MVAPICH2 / Intel MPI, and Allinea power monitoring. We plan on setting up a small power efficient Ethernet network with a 12-port switch and a MinnowBoard MAX acting as a provisioning agent, host for the scheduler, and monitoring host.

C. Power Management Optimizations

To manage power per node, we plan on leveraging the Intel Intelligent Power Node Manager most modern Intel hardware supports, which will be used to maintain the power budget; we will explore dynamic frequency scaling, processor state management, and elimination of cooling fans, and water cooling.

We also plan on using advanced compiler research to find the best compiler optimizations that delivers the best flops/watt. One such research project we will investigate is Softener, a compiler for C/C++ nondeterministic applications. These applications are often designed to be resilient to nondeterministic variations of intermediate data values. We will investigate to what degree the SCC applications have this behavior of nondeterminism. Softener leverages such resiliency to profitably satisfy producer-consumer dependences that involve such intermediate values. This work is led by one of the mentors Simone Campanoni. He found that some of these dependences can be satisfied with alternative producers, which are automatically generated by Softener. This creates a new degree of freedom that his compiler exploits within the set of code optimizations included in the industrial-strength compiler LLVM. Doing so, Softener significantly increases the performance/power consumption pareto frontier for nondeterministic C++ programs.

III. VENDOR/INSTITUTION SUPPORT

ANL is in the early stages of confirming sponsorship from Intel, which supported the IIT team in both 2014 and 2015. Our Intel sponsor point of contacts are Chris Allison and Keith Kirkendall. 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. We are aiming for the competition hardware to be available to the students in June 2017. Our ANL sponsor points of contacts are William Scullin and Ben Allen. The team primary advisor is Ioan Raicu.

We have partnered with Intel and ANL to leverage their experience and development efforts in porting software to the many-core Intel Xeon Phi x200-series processor. While the Xeon Phi is still a relatively exotic architecture, Intel has been working with vendors and application authors to ensure common applications work well on the platform. Luckily, this includes all the known SC17 SCC applications [1].

As they are frequently used reference points, Intel provides optimized HPL and HPCG binaries, as well as guidance for users looking to compile the software for themselves. Initial research and compiling the reference binary for HPCG made us assume that we would be at a disadvantage on the Xeon Phi given its smaller cache sizes and very high core count, particularly in the Symmetric Gauss-Seidel smoother. We were pleasantly surprised that by carefully choosing problems sizes and parameters, setting processor and memory affinity, and controlling for power consumption, we could still outperform a relatively similar Intel Xeon E5-2697 v4 cluster on a per-watt basis by about 15%. With a tuned binary, we suspect that we could further out perform CPU-based competitors. While the top-of-the-line GPU solutions do outperform the Intel Xeon Phi x200 on a unit basis, we still believe we'll be highly competitive in terms of performance-per-watt.

The RWTH Aachen team responsible for the Tersoff multi-body potential solver in LAMMPS had a strong focus on Xeon Phi performance. Their paper, "The Vectorization of the Tersoff Multi-Body Potential: An Exercise in Performance Portability", includes existing Xeon Phi 7250 benchmarks that showed the Xeon Phi as a competitive option. The authors noted "no optimization specific to KNL was incorporated in our code" with the compiler performing vectorization which makes us relatively optimistic that improvements in the Intel compiler and some minor tuning should allow us to do very well with this application. This is opposed to their description of GPU support via KOKKOS, which is noted as requiring further development.

Finally, we feel that our Xeon Phi-based platform will give us a leg-up on the mystery application. SCC organizers have historically chosen mystery applications with reasonable CPU and GPU support. However, the corpus of applications has included software written in Matlab and Java which tend to have more options for CPU-based clusters. Likewise, some community codes are very difficult to compile or only work with specific versions of compilers and libraries. While there is only one CUDA toolkit, there are at least four shipping compilers with support for the Intel Xeon Phi, which improves our odds of being able to produce a working binary quickly. We had debated inclusion of a GPU for rendering tasks, but with the creation of Embree, we're confident that an all-cpu solution is up for answering rendering tasks.

IV. DIVERSITY

The team is diverse in the sense that it includes high-school students from freshman to juniors, and undergrads across different years, from 1st year to 3rd years. Furthermore, the 11 students span 5 different institutions, and represent both male and female students. Most of the students are straight A students with full-ride scholarships.

Nearly all the students have Linux experience through research in the DataSys laboratory, internships, or coursework. Most of 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. Most have been exposed to 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). One of the students (Alex Ballmer) has also participated in SCC in both 2014 and 2015, and has attended the SC conference every year since 2014.

V. TEAM PREPARATION

The mentors have organized a 10-week summer program that will run at IIT from June 12th 2017 to August 18th 2017 [3], where the 11 identified students will spend 20 hours a week preparing for the SCC competition (paid hourly with funds from a NSF REU supplement). The students will be mentored by the 6 mentors as well as graduate students in the DataSys lab at IIT. The students will spend the summer studying the announced applications, in doing code review of other scientific applications, and in setting up and configuring the proposed cluster, including installing and configuring Linux, a shared file system, MPICH, HPL, and all the announced applications. Students will also use ANL resources such as Chameleon, Cooley, Theta, and the 2014/2015 SCC clusters as additional testbeds to help us better understand the application performance and scalability. They will explore a large parameter space that governs the possible configurations that will yield the best performance per watt.

The team mentors firmly believe 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 [17] being developed at Argonne National Laboratory. We will also leverage the innovative work in distributed filesystems FusionFS [13] to outperform more traditional HPC filesystems such as PVFS or GPFS. Dr. Raicu's group has made much progress in the design and implementation of distributed key/value storage systems (ZHT [14]) which might come in valuable to further accelerate the respective data-intensive HPC applications. Much of the research [9, 10, 11, 12, 15, 16] happening in the DataSys laboratory could be put to the test in accelerating these HPC applications.

In both 2014 and 2015, the IIT team built their own custom system status display from LED strips – a unique feature of our rack that earned a lot of attention from the SC14 and SC15 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.

The students will participate in daily meetings with graduate students, and weekly meetings with their mentors. They will join in various activities from a parallel running NSF REU program at IIT called "BigDataX: From theory to practice in Big Data computing at eXtreme scales" [2]; some of these activities are "Life as a PhD Student", "Careers in Research", "Women in Computing", picnic with other REU programs in Chicago, and a day-long field trip to ANL where students can meet researchers in a variety of fields using computing to solve grand challenges.

The final students who will make up the Chicago Fusion Team will continue training in the Fall semester. They will take 3-credits of independent study to allow them to have sufficient time in their schedule to continue preparing for SCC with at least 10 hours/week. Over the course of 6 months (June to November), each student would have invested nearly 500 hours in this competition!

REFERENCES

[1] Student Cluster Competition (SCC) @ IEEE/ACM Supercomputing/SC 2017; http://www.studentclustercompetition.us/2017/applications.html, 2017

[2] NSF REU Site: BigDataX: From theory to practice in Big Data computing at eXtreme scales; http://datasys.cs.iit.edu/grants/BigDataX/index.html, 2017

[3] BigDataX SCC Program -- Summer 2017; http://datasys.cs.iit.edu/grants/BigDataX/2017/scc.html, 2017

[4] Student Cluster Competition (SCC) @ IEEE/ACM Supercomputing/SC 2014; http://datasys.cs.iit.edu/events/SCC-SC14/index.html, 2017

[5] Student Cluster Competition (SCC) @ IEEE/ACM Supercomputing/SC 2015; http://datasys.cs.iit.edu/events/SCC-SC15/index.html, 2017

[6] Ben Walters, Alex Ballmer, Andrei Dumitru, Adnan Haider, Serapheim Dimitropoulos, Ariel Young, William Scullin, Ben Allen, Ioan Raicu. "15 TFlops Haswell vs. 60 TFlops Knight Landing for HPC Scientific Computing Applications", Student Cluster Competition (SCC), IEEE/ACM Supercomputing/SC 2015

[7] Kevin Brandstatter, Jason DiBabbo, Daniel Gordon, Ben Walters, Alex Ballmer, Lauren Ribordy, Ioan Raicu. "Delivering 3.5 Double Precision GFlops/Watt and 200Gb/sec Bi-Section Bandwidth with Intel Xeon Phi-based Cisco Servers", Student Cluster Competition (SCC), IEEE/ACM Supercomputing/SC 2014

[8] Kevin Brandstatter, Ben Walters, Alexander Ballmer, Adnan Haider, Andrei Dumitru, Serapheim Dimitropoulos, Ariel Young, William Scullin, Ben Allen, Ioan Raicu. "Experiences in Optimizing Cluster Performance For Scientific Applications: Controlling Configuration, Utilization, and Power Consumption", GCASR 2015

[9] Dongfang Zhao, Da Zhang, Ke Wang, Ioan Raicu. "Exploring Reliability of Exascale Systems through Simulations", ACM HPC 2013

[10] Ioan Raicu, Ian Foster, Yong Zhao, Alex Szalay, Philip Little, Christopher M. Moretti, Amitabh Chaudhary, Douglas Thain. "Towards Data Intensive Many-Task Computing", book chapter in "Data Intensive Distributed Computing: Challenges and Solutions for Large-Scale Information Management", IGI Global Publishers, 2012

[11] Dongfang Zhao, Ioan Raicu. "HyCache: A User-Level Caching Middleware for Distributed File Systems", IEEE HPDIC 2013

[12] Michael Wilde, Ioan Raicu, Allan Espinosa, Zhao Zhang, Ben Clifford, Mihael Hategan, Kamil Iskra, Pete Beckman, Ian Foster. "Extreme-scale scripting: Opportunities for large task-parallel applications on petascale computers", Scientific Discovery through Advanced Computing Conference (SciDAC09) 2009

[13] Dongfang Zhao, Ning Liu, Dries Kimpe, Robert Ross, Xian-He Sun, and Ioan Raicu. "Towards Exploring Data-Intensive Scientific Applications at Extreme Scales through Systems and Simulations", IEEE Transaction on Parallel and Distributed Systems (TPDS) Journal 2015

[14] Tonglin Li, Xiaobing Zhou, Ke Wang, Dongfang Zhao, Iman Sadooghi, Zhao Zhang, Ioan Raicu. "A Convergence of Key-Value Storage Systems from Clouds to Supercomputers", Concurrency and Computation: Practice and Experience (CCPE) Journal 2015

[15] Ke Wang, Abhishek Kulkarni, Michael Lang, Dorian Arnold, and Ioan Raicu. "Exploring the Design Tradeoffs for Extreme-Scale High-Performance Computing System Software", IEEE Transaction on Parallel and Distributed Systems (TPDS) 2015

[16] Dongfang Zhao, Ioan Raicu. "Distributed File Systems for Exascale Computing", Doctoral Showcase, IEEE/ACM Supercomputing/SC 2012

[17] Darhsan, http://www.mcs.anl.gov/research/projects/darshan/, 2017