CUG.2017.CAFFEINATED COMPUTING

Redmond, Washington May 7-11, 2017

Welcome To CUG 2017

The Cray User Group board would like to invite you to Redmond Washington for CUG 2017 hosted by Cray and the CUG Board May 7-11, 2017.

The 60th CUG has us returning to the Seattle area near Cray headquarters and in the heart of Washington wine-tasting country. As we look forward to 2017 the goal of the CUG Board is to expand the horizons of our attendees and our program to include even more data, storage, and analytics discussions while continuing to support high quality HPC content that is at the core of CUG.

There are many challenges in our field from rising power utilization, increasing core counts and scalability limitations, to the integration of new approaches that couple data analytics, HPC, and containerization in single environments. It is through events like CUG that we can share our approaches, expertise, and failures in order to continue to push boundaries. The relatively young history of Seattle has striking parallels to our own industry, in how difficult it was to first settle and the success it has achieved in the Pacific Northwest. Home to leading industry such as Boeing, Microsoft, and Amazon there's much we can learn from the area. Most importantly, to many sleep-deprived tech workers, it also has a thriving coffee culture, including being the birthplace of Starbucks. The theme of Caffeinated Computing touches on that tradition, and it may not be a coincidence that Seattle has such a storied history with coffee along with being the headquarters to our dear colleagues at Cray.
We value your contributions to our technical program and look forward to the progress that members and sponsors will share at CUG 2017.

David Hancock
President, Cray User Group

Invited Speakers
Keynote

DOUGLAS B. (Doug) KOTHE, Ph.D.
Oak Ridge National Laboratory

What are the Opportunities and Challenges for a new Class of Exascale Applications? What Challenge Problems can these Applications Address through Modeling and Simulation & Data Analytic Computing Solutions?

Abstract: The Department of Energy's (DOE) Exascale Computing Project (ECP) is a partnership between the DOE Office of Science and the National Nuclear Security Administration. Its mission is to transform today's high performance computing (HPC) ecosystem by executing a multi-faceted plan: developing mission critical applications of unprecedented complexity; supporting U.S. national security initiatives; partnering with the U.S. HPC industry to develop exascale computer architectures; collaborating with U.S. software vendors to develop a software stack that is both exascale-capable and usable on U.S. industrial and academic scale systems, and training the next-generation workforce of computer and computational scientists, engineers, mathematicians, and data scientists. The ECP aims to accelerate delivery of a capable exascale computing ecosystem that will enable breakthrough modeling and simulation (M&S) and data analytic computing (DAC) solutions to the most critical challenges in scientific research, energy assurance, economic competitiveness, and national security.

The computer and computational science and engineering communities in the public, private, and government sectors have been arguably thinking about exascale-class modeling and simulation technologies and capabilities for almost a decade. With exascale platforms becoming more certain and finally within sight, application developers and users must "get real" now to adequately take advantage of this opportunity. The hardware and software technologies currently envisioned in exascale platforms will present new challenges for application developers that could be disruptive relative to current approaches. New algorithms, for example, that communicate infrequently and store very little, may be critical for applications to move forward or even "hold pace". Hybrid node architectures with hierarchical memory and compute technologies will likely be the norm, and applications may face comprehensive restructuring to exploit more appropriate task-based programming models and new data structures.

Given these challenges, tremendous opportunity nevertheless exists for science-based computational applications that can deliver, through effective exploitation of exascale HPC technology, breakthrough M&S and DAC solutions that yield high-confidence insights and answers to the nation's most critical problems and challenges in scientific discovery,

energy assurance, economic competitiveness, and national security. While reflecting on some of my own person R&D experiences, I will survey these application opportunities, where I will also touch upon challenges, decadal challenge problems, and prospective outcomes and impact.

Bio: Douglas B. Kothe (Doug) has over three decades of experience in conducting and leading applied R&D in computational applications designed to simulate complexphysical phenomena in the energy, defense, and manufacturing sectors. Doug is currently the Deputy Associate Laboratory Director of the Computing and Computational Sciences Directorate (CCSD) at Oak Ridge National Laboratory (ORNL). Prior positions for Doug at ORNL, where he has been since 2006, were Director of the Consortium for Advanced Simulation of Light Water Reactors, DOE's first Energy Innovation Hub (2010-2015), and Director of Science at the National Center for Computational Sciences (2006-2010).

Before coming to ORNL, Doug spent 20 years at Los Alamos National Laboratory, where he held a number of technical and line and program management positions, with a common theme being the development and application of modeling and simulation technologies targeting multi-physics phenomena characterized in part by the presence of compressible or incompressible interfacial fluid flow. Doug also spent one year at Lawrence Livermore National Laboratory in the late 1980s as a physicist in defense sciences.

Doug holds a Bachelor in Science in Chemical Engineering from the University of Missouri – Columbia (1983) and a Masters in Science (1986) and Doctor of Philosophy (1987) in Nuclear Engineering from Purdue University.

Invited Speakers

Invited Talk

Arno Kolster

Co-Founder, Providentia Worldwide

Perspectives on HPC and Enterprise High Performance Data Analytics

Abstract: Mr. Kolster will present his experience of blending HPC and enterprise architectures to solve real-time, web-scale analytics problems and discuss the need to bridge the gap between HPC and enterprise. His unique perspective illustrates the need for enterprise to embrace HPC technologies and vice versa.

Bio: Arno was born in The Netherlands and grew up in Canada where he received a degree in Computer Science from The University Of Calgary. Currently residing in San Francisco, his main career focus over the past 30 years has been database architecture, database administration and operations architecture for industries that include oil and gas, emergency services, finance and until recently, 14 years at PayPal. His extensive knowledge of relational databases has expanded to include new database technologies such as NoSQL and graph databases. An interest in HPC and technical computing came about as a result of finding solutions to solving real time data analytics across distributed systems at web scale. Arno and his colleague, Ryan Quick, have received IDC Innovation Excellence Awards at both Super-Computing 2012 and 2104 as well as numerous HPC Wire Reader's Choice Awards. He's been invited to speak domestically and internationally on HPC and its deployment at PayPal. He is co-founder of independent consulting firm Providentia Worldwide.

Sessions and Abstracts

Monday, May 8th

08:30-10:00 Tutorial 1A

Migrating, Managing, and Booting Cray XC and CMC/eLogin

Longley, Keopp, Landsteiner

System management on Cray XC systems has improved since SMW 8.0/CLE 6.0 was introduced. This tutorial moves from introductory information to advanced topics in system management of XC series systems including CMC/eLogin. An overview of system management for the XC series system is provided: the Image Management and Provisioning System (IMPS), the Configuration Management Framework (CMF), and Node Image Mapping Service (NIMS) for XC series systems, and the CSMS, OpenStack, and eLogin software for external login nodes. New system management enhancements introduced since SMW 8.0.UP01/CLE 6.0.UP01 will be described including: node groups in config set; improved config set validation; best practices for using Ansible; an overview of the anatomy of a CLE boot; changes to improve boot performance and reliability; techniques for tuning boot performance; and techniques for troubleshooting CLE boots. The process of migrating an XC system from SMW 7.2.UP04/CLE 5.2.UP04 to SMW 8.0.UP03/CLE 6.0.UP03 and the external login nodes from CIMS/CDL software to the CMC/eLogin software will be described. Discussion will include comparison between the older generation of software and the newer generation of software.

08:30-10:00 Tutorial 1B

Getting the most of of Knight's Landing

Levesque

There are several large Knight Landing's systems in the community and there still is a lot of confusion on how best to utilize this new many-core system. KNL is a difficult system to utilize most effectively because there are so many confirgurations that can be employed and users do not have enough time or access to discern how best to configure the system for their particular application. This tutorial will be built upon the vast experience that the author has accumulated over the past twelve months using very large applications on the Trinity KNL at Los Alamos Scientific National Laboratory. Over that time a series of tests were conducted to determine the best clustering mode for different types of applications. The clustering modes will be discussed with results from a variety of applications. The next important aspect of KNL is how best to configured the High Bandwidth Memory - once again the decision depends upon the size of memory required on each node and memory access patterns. The most important aspect of using KNL is to vectorize as much as possible, as this is the only mode in which KNL can out-perform the state of the art Xeon node. While Threading on the node is not as critical as originally thought, there are still places when employing OpenMP threads can gain addition performance. This tutorial will cover all the aspects of getting the most of the Knight's Landing and when Knight's landing may not be better than the state-of-the-art Xeon.

08:30-10:00 Tutorial 1C

Shifter: Bringing Container Computing to HPC

Gerhardt, Canon, Jacobsen

This half-day tutorial will introduce researchers and developers to the basics of container computing and running those containers in a Cray environment using Shifter, a framework

Sessions and Abstracts

that delivers docker-like functionality to HPC by extracting images from native formats (such as a Docker image) and converting them to a common format that is optimally tuned for the HPC environment. The tutorial will also cover more advanced topics including how to set up a Shifter image Gateway and create images that run MPI applications that require high-performance networks. The tutorial will also cover ways that Docker images can be used in the scientific process including packaging images so they can be used to regenerate and confirm results and used in the publication process and will integrate hands-on exercises throughout the training. These exercises will include building Docker images on their own laptop and running those images on an Shifter-enabled HPC system via tutorial accounts. While attendees will not require advanced knowledge of Docker or Shifter, they should be familiar with basic Linux administration such as installing packages and building software. This tutorial will be presented by an experienced team including the authors of the Shifter tool and members of NERSCâ€™s Data and Analytics Services Group that focuses on training scientists to use HPC. This tutorial will present an updated version of the very popular tutorial presented earlier this year at the Supercomputing 2016 Conference.

13:00-14:30 Tutorial 2A
(continued)

13:00-14:30 Tutorial 2B
(continued)

13:00-14:30 Tutorial 2C
(continued)

16:40-18:20 BoF 3A
Systems Support SIG Meeting
Cardo

This BoF will be focused on topics related to the day-to-day operations of Cray supercomputers including operating system support, storage, and networking.

16:40-17:30 BoF 3B
New use cases and usage models for Cray DataWarp
Alam, Schulthess, Hadri, Bard, Martinasso

This BOF will be a collaborative effort between the sites that have deployed the Cray DataWarp technology and Cray DataWarp engineers and developers. The goal of this BOF is to explore use cases and usage models that could open up opportunities for data science workflows. Motivating examples will be presented, which will be followed by a technical discussion on the current implementation of DataWarp software stack. The BOF participants will have an opportunity to share experiences using the DataWarp technology and to contribute representative data science use cases that could benefit from the Cray DataWarp technology.

16:40-17:30 Tutorial 3C
(continued)

17:30-18:20 BoF 3B
Sharing Cray Solutions
Marquardt

You're invited to help brainstorm! Cray is considering ways to enhance the ecosystem of solutions for Cray platforms by building a strong community. Could include: Message boards Code sharing repositories Periodic meet ups Other ideas. Hasn't Cray always

Sessions and Abstracts

supported customer collaboration? Yes, Cray has long supported customers collaborating with each other through groups like CUG, Xtreme, etc. Why does Cray support customer collaboration? Cray systems are used in many different environments and need to interoperate with a variety of technologies that customers use to manage their data centers. The breadth of technologies that Cray systems need to interact with is extensive. When we are successful in helping customers collaborate it's a win-win. What has changed that would drive us to do more? The way that Cray systems are managed has been revolutionized with the introduction of CLE 6.0. New system management technologies offer far greater flexibility in configuring Cray systems, and do it in a way that facilitates sharing of solution recipes in ways that just weren't possible previously. Customer benefits? Solutions are readily available and easy to use Strong community to share successes and lessons learned Collaboration with Cray's very capable customer base Cray benefits? Greater breadth of solutions to customers' needs Reduced support burden Visibility into customer's needs Ability to pick off the most popular solutions and incorporate into future product releases How to organize? That's what this BOF is all about. We need your input!

Tuesday, May 9th

08:30-09:40 General Session 4

CUG Welcome

Hancock

Welcome from the CUG President.

08:30-09:40 General Session 4

Keynote: What are the Opportunities and Challenges for a new Class of Exascale Applications? What Challenge Problems can these Applications Address through Modeling and Simulation & Data Analytic Computing Solutions?

Kothe

The Department of Energy's (DOE) Exascale Computing Project (ECP) is a partnership between the DOE Office of Science and the National Nuclear Security Administration. Its mission is to transform today's high performance computing (HPC) ecosystem by executing a multi-faceted plan: developing mission critical applications of unprecedented complexity; supporting U.S. national security initiatives; partnering with the U.S. HPC industry to develop exascale computer architectures; collaborating with U.S. software vendors to develop a software stack that is both exascale-capable and usable on U.S. industrial and academic scale systems, and training the next-generation workforce of computer and computational scientists, engineers, mathematicians, and data scientists. The ECP aims to accelerate delivery of a capable exascale computing ecosystem that will enable breakthrough modeling and simulation (M&S) and data analytic computing (DAC) solutions to the most critical challenges in scientific research, energy assurance, economic competitiveness, and national security. The computer and computational science and engineering communities in the public, private, and government sectors have been arguably thinking about exascale-class modeling and simulation technologies and capabilities for almost a decade. With exascale platforms becoming more certain and finally within sight, application developers and users must "get real" now to adequately take advantage of this opportunity. The hardware and software technologies currently

Sessions and Abstracts

envisioned in exascale platforms will present new challenges for application developers that could be disruptive relative to current approaches. New algorithms, for example, that communicate infrequently and store very little, may be critical for applications to move forward or even "hold pace". Hybrid node architectures with hierarchical memory and compute technologies will likely be the norm, and applications may face comprehensive restructuring to exploit more appropriate task-based programming models and new data structures. Given these challenges, tremendous opportunity nevertheless exists for science-based computational applications that can deliver, through effective exploitation of exascale HPC technology, breakthrough M&S and DAC solutions that yield high-confidence insights and answers to the nation's most critical problems and challenges in scientific discovery, energy assurance, economic competitiveness, and national security. While reflecting on some of my own person R&D experiences, I will survey these application opportunities, where I will also touch upon challenges, decadal challenge problems, and prospective outcomes and impact.

09:40-09:50 Sponsor Talk 5

[DDN] **Flash-Native Caching for Predictable Job Completion in Data-Intensive Environments**

Coomer

In this Talk, Dr. James Coomer, Senior Technical Advisor for EMEA, will provide examples of testing and deployment results from the past year of DDN's Flash-Native caching solution: Infinite Memory Engine. He will also discuss the company's Lustre development and productization efforts and future plans.

09:50-10:00 Sponsor Talk 6

[ANSYS] Why Supercomputing Partnerships Matter for CFD Simulations

(Slagter)

This presentation will address how CFD scalability and capabilities for customization have evolved over the last decade, and how supercomputing partnerships are playing a crucial role. Examples of extreme scalability (on Cray systems) and application customization will be featured that are illustrative for all users - whether you're running on a 100,000+ core supercomputer or a 1,000-core cluster.

10:30-12:00 General Session 7

Cray Corporate Update

(Ungaro, Papaefstathiou, Scott)

Peter Ungaro – President & Chief Executive Officer – Corporate Update

Stathis Papaefstathiou, Sr. Vice President, Research & Development – R&D Update

Steve Scott, Sr. Vice President & Chief Technology Officer – Cray Futures

13:00-14:30 Technical Session 8A

Early Experience on Theta - the New Argonne's Intel Second Generation Xeon Phi-based Supercomputer

(Morozov, Chunduri, Harms, Parker, Knight, Kumaran)

Theta, the latest supercomputer at the ALCF, is based on Intel's second-generation Xeon Phi processor codenamed Knights Landing. This early production system will help ALCF users

transition to the new architecture, serving as a bridge between the current supercomputer, IBM Blue Gene/Q Mira, and the next leadership-class supercomputer, Intel-Cray Aurora. In this paper, we present the results of benchmarking Theta on various levels of the hardware hierarchy from the core level, vectorization and instruction scheduling, node level, memory hierarchy, the interaction of new multi-channel high-bandwidth memory and traditional DRAM. We also present the performance and scalability results of the system level benchmarking on widely used MPI operations, the effects of node allocation and routing schemes on performance and repeatability, and the influence of congestion versus congestion-free controlled environment. Finally, we present the results of porting, tuning, and scaling several ALCF applications, and discuss the lessons learned from this exercise.

Performance on Trinity Phase 2 (a Cray XC40 utilizing Intel Xeon Phi processors) with Acceptance Applications and Benchmarks

(Agelastos, Rajan, Wichmann, Baker, Domino, Draeger, Anderson, Balma, Behling, Berry, Carrier, Davis, Sandness, Thomas, Warren, Zhu)

Trinity is the first NNSA ASC Advanced Technology System (ATS-1) designed to provide the application scalability, performance, and system throughput required for the Nuclear Security Enterprise. Trinity Phase 1 with 9,436 dual-socket Haswell nodes is currently in production use. Phase 2 system, the focus of this paper, has close to 9,900 Intel Knights Landing (KNL) Xeon Phi nodes. This paper documents the performance of the selected acceptance applications, Sustained System Performance benchmark suite, and a number of micro-benchmarks. This paper discusses the experiences of the Tri-Lab (LANL, SNL, and LLNL) and Cray teams to extract the optimal performance with considerations to: to the choice of the KNL memory mode, hybrid MPI+OpenMP parallelization, vectorization, and HBM utilization.

Evaluating the Networking Characteristics of the Cray XC-40 Intel Knights Landing Based Cori Supercomputer at NERSC

(Doerfler, Austin, Cook, Deslippe, Kandalla, Mendygral)

There are many potential issues associated with deploying the Intel Knights Landing (KNL) manycore processor in a large-scale supercomputer. One in particular is the ability to fully utilize the high-speed communications network, given the serial performance of a Xeon Phi core is a fraction of a Xeon core. In this paper we take a look at the tradeoffs associated with allocating enough cores to fully utilize the Aries high-speed network versus cores dedicated to computation, e.g. the tradeoff between MPI and OpenMP. In addition, we evaluate new features of Cray MPI in support of KNL, such as inter-node optimizations and support for KNL's high-speed memory (MCDRAM). We also evaluate one-sided programming models such as Unified Parallel C. We quantify the impact of the above tradeoffs and features using a suite of NERSC applications.

13:00-14:30 Technical Session 8B

Toward Interactive Supercomputing at NERSC with Jupyter

(Thomas, Canon, Cholia, Gerhardt, Racah)

Extracting scientific insights from data increasingly demands a richer, more interactive

experience than traditional high-performance computing systems historically have provided. We present our efforts to leverage Jupyter for interactive data-intensive supercomputing on the Cray XC40 Cori system at the National Energy Research Scientific Computing Center (NERSC). Jupyter is a flexible, popular literate-computing web application for creating "notebooks" containing code, equations, visualization, and text. We explain the motivation for interactive supercomputing, describe our implementation strategy, and outline lessons learned along the way. Our deployment will allow users access to software packages and specialized kernels for scalable analytics with Spark, real-time data visualization with yt, complex analytics workflows with DASK and IPyParallel, and much more. We anticipate that many users may come to rely exclusively on Jupyter at NERSC, leaving behind the traditional login shell.

In-situ data analytics for highly scalable cloud modelling on Cray machines

(Brown, Hill, Shipway, Weiland)

MONC is a highly scalable modelling tool for the investigation of atmospheric flows, turbulence and cloud microphysics. Simulations produce very large amounts of raw data which must then be analysed for scientific investigation. For performance and scalability this analysis should be performed in-situ as the data is generated however one does not wish to pause the computation whilst analysis is performed. We present the in-situ analytics approach of MONC, where cores of a processor are shared between computation and data analytics. By asynchronously sending data to an analytics core, the computational cores can run continuously without having to pause for IO or analysis. We describe our framework and analytics pipeline, which is highly asynchronous, along with solutions to challenges encountered and the optimal configuration on Cray machines. The result of this work is a highly scalable analytics approach, with a minimal performance impact, illustrated on 36,045 cores of an XC30.

Precipitation Nowcasting: Leveraging Deep Recurrent Convolutional Neural Networks

(Heye)

Automating very short-term precipitation forecasts can prove a significant challenge in that traditional physics-based weather models are computationally expensive; by the time the forecast is made, it may already be irrelevant. Deep Learning offers a solution to this problem, as that a computationally dense machine can train a neural network ahead of time using historical data and deploy that trained network in real-time to produce a new output within seconds or minutes. Our team intends to prove the capabilities of Deep-Learning in short-term forecasting by leveraging a model built on Convolutional Long Short-Term Memory (convLSTM) networks. By designing a 3D sequence-to-sequence convLSTM model, we hope to offer accurate precipitation forecasts at minute level time resolution and comparable spatial resolution to the radar input data. Our work will be accelerated by the GPU-dense CS-Storm system for training and the Cray GX for real-time processing of radar data.

13:00-14:30 Technical Session 8C

Telemetry-enabled Customer Support using the Cray System Snapshot Analyzer (SSA)

(Duckworth, Coryell, McLeod, Blakeborough)

SSA is a Cray customer service application designed to support product issue diagnosis and reduce time to resolution. SSA is focused on the submission of product telemetry from a customer system to Cray, over a secure network channel. SSA provides value to the support process by automating 1) the collection, submission and analysis of product diagnostic information 2) the collection, submission and analysis of product health information and 3) key aspects of the customer support process. The focal topic for this paper is a discourse on the benefits of SSA use. For background, we will provide a general overview of SSA and references for further reading. Next, we will provide an update on SSA product release and operations history. Finally, we will discuss the anticipated product roadmap for SSA.

How-to write a plugin to export job, power, energy, and system environmental data from your Cray XC system

(Martin, Whitney, Rush, Kappel)

In this paper we take a deep dive into writing a plugin to export power, energy, and other system environmental data from a Cray XC system. With the release of SMW 8.0/CLE 6.0 software Cray has enabled customers to create site-specific plugins to export all of the data that can flow into the Cray Power Management Database (PMDB) into site-specific infrastructure. In this paper we give practical information on what data is available using the plugin, and how to write, test, and deploy a plugin. We also share and explain example plugin code, detail design considerations that need to be considered when architecting a plugin, and look at some practical use cases supported by exporting telemetry data off a Cray XC system. This paper is targeted at plugin developers, system administrators, data scientists, and site planners

Using Open XDMoD for accounting analytics on the Cray XC supercomputer

(Lorenzen, Kasacjak, Coverston)

As supercomputers grow in size, with more users and projects, the system utilization and accounting log files grow as well, often beyond easy native comprehension, and thus is a flexible graphical tool for accounting analytics required. This presentation will account for the joint effort of the Danish Meteorological Institute, DMI, and Cray to adapt Open XDMoD, http://open.xdmod.org, to the DMI Cray XC supercomputer. Extensions to the RUR framework to monitor metrics of particular relevance to the site and have them embedded into Open XDMoD will be presented as well. This will show Open XDMoD being a flexible tool for use with the Cray XC supercomputer with good possibility for extending metrics ingestion and graphical presentation in numerous ways.

15:00-16:30 Technical Session 9A

Using Spack to Manage Software on Cray Supercomputers

(Melara, Gamblin, Becker, French, Belhorn, Thompson, Hartman-Baker)

HPC software is becoming increasingly complex. A single application may have over 70 dependency libraries, and HPC centers deploy even more libraries for users. Users demand many different builds of packages with different compilers and options, but building them can be tedious. HPC support teams cannot keep up with user demand without better tools. Spack is an open-source package manager created at Lawrence Livermore National Laboratory (LLNL). It allows a build to be

Sessions and Abstracts

specified as a template for many combinatorial installations, and it gives users fine-grained control over a build's DAG. Spack is used by HPC centers, developers, and users to quickly deploy software. This paper describes the steps we have taken to integrate Spack with the Cray Programming Environment (CPE). We describe changes that we made to Spack's module and architecture support, and we describe preliminary results of the rollout of Spack on NERSC's Cray XC40 system (Cori).

Regression testing on Shaheen Cray XC40: Implementation and Lessons Learned

(Hadri, Kortas, Fiedler, Markomanolis)

Leadership-class supercomputers are becoming larger and more complex tightly systems integrated of different hardware components, counting tens of thousands of processors and memories, kilometers of networking cables, large amount of disks and supporting hundreds of applications and libraries. To increase scientists productivity and ensure that applications efficiently and effectively exploit the full potential of available resource, all the components must deliver a reliable, stable and performing service. Therefore, to deliver the best computing environment to our users the sanity assessment of the system is critical, especially after an unplanned downtime or any scheduled maintenance session. We present the design and the implementation improvements of the regression testing on Shahen2 XC40 to detect and track issues related to the performance and functionalities of compute nodes, storage, network as well as programming environment. We also share the analysis of the results over 24 months along with the lessons learned.

Python Usage Metrics on Blue Waters

(MacLean)

Blue Waters supports a large Python stack containing over 300 packages. As part of maintaining this support, logging functionality has been introduced to track the usage statistics of both National Center for Supercomputing Applications (NCSA) and user provided Python packages. Due to the number of NCSA supplied packages, it is rare to receive a request for packages which are not already installed, leading to a lack of information about which packages and their dependencies are being used. By tracking module imports, a detailed log of usage information has been used to focus support efforts on improving the usability and performance of popular usage patterns.

15:00-16:30 Technical Session 9B

libhio: Optimizing IO on Cray XC Systems With DataWarp

(Hjelm, Wright)

High performance systems are rapidly increasing in size and complexity. To keep up with the IO demands of applications and to provide improved functionality, performance and cost, IO subsystems are also increasing in complexity. To help applications to utilize and exploit increased functionality and improved performance in this more complex environment, we developed a new Hierarchical Input/Output (HIO) library: libhio. In this paper we present the motivation behind the development and the design of libhio.

How to Use Datawarp

(Overby)

Cray DataWarp is a set of technologies that accelerates application I/O in order to reduce job wall clock time. It creates a near storage layer between main memory and hard disk drives, with direct attached solid-state disk

(SSD) storage to provide more cost effective bandwidth than an external parallel file system (PFS) allowing DataWarp to be provisioned for bandwidth and the PFS to be provisioned for capacity and resiliency. This paper will discuss ways in which DataWarp can benefit applications and provide specific examples of using DataWarp with the Moab, PBS and Slurm workload managers. The detailed examples will include how to use DataWarp striped storage, how applications access that storage, how to stage data, how to use DataWarp per-node storage, how to request storage that persists across multiple jobs, and also how to use of DataWarp as a transparent cache.

Datawarp Accounting Metrics
(Barry)

Datawarp, a burst buffer package of fast flash storage and filesystem software, can provide a large improvement in I/O performance for jobs run on a Cray system, but the scale of this improvement depends on the configuration of the system as well as application optimization. Datawarp is a limited resource on Cray systems, with insufficient storage capacity for all jobs to completely replace their use of parallel filesystems. Thus, it is useful to know how various applications make use of the resource and to track total utilization. These statistics indicate which users are using Datawarp and for which applications. Cray Resource Utilization Reporting (RUR) plugins are available to collect Datawarp statistics and to archive it for future analysis. This paper describes the available Datawarp usage statistics, context for interpreting those metrics, and some case studies of applications using Datawarp in different ways.

15:00-16:30 Technical Session 9C

Theta: Rapid Installation and Acceptance of an XC40 KNL System
(Leggett, Fahey, Coghlan, Harms, Rich, Allen, Holohan, McPheeters)

In order to provide a stepping stone from the Argonne Leadership Computing Facility's (ALCF) world class production 10 petaFLOP IBM BlueGene/Q system, Mira, to its next generation 180 petaFLOPS 3rd generation Intel Xeon Phi system, Aurora, ALCF worked with Intel and Cray to acquire an 8.6 petaFLOPS 2nd generation Intel Xeon Phi based system named Theta. Theta was delivered, installed, integrated and accepted on an aggressive schedule in just over 3 months. We will detail how we were able to successfully meet the aggressive deadline as well as lessons learned during the process.

Extending CLE6 to a multi-supercomputer Operating System
(Jacobsen, Declerck)

NERSC operates multiple Cray supercomputing platforms using CLE6. In this paper we describe our methods, customizations, and additions to CLE6 to enable a coordinated management strategy for all four systems (two production systems and two development systems). Our methods use modern software engineering tools and practices to run precisely the same management software on all four systems, while still allowing version drift, customization, and some amount of acceptable feature drift between systems. In particular, we have devised procedures, software, and tools to allow the test and development systems to move rapidly between the exact production system configuration as well as various testing (future) configurations. We also discuss capabilities added to CLE6, such as customized ansible facts trees, integration of ansible-vault for

securing sensitive information, "branching" zypper repositories in a context sensitive way, replicating NIMS data structures across systems. These methods reduce administration and development cost and increase availability and reproduce-ability.

Updating the SPP Benchmark Suite for Extreme-Scale Systems

(Bauer, Anisimov, Arnold, Bode, Brunner, Cortese, Haas, Kot, Kramer, Kwack, Li, Mendes, Mokos, Steffen)

For the effective deployment of modern extreme-scale systems, it is critical to rely on meaningful benchmarks that provide an assessment of the achievable performance a system may yield on real applications. The Sustained Petascale Performance (SPP) benchmark suite was used very successfully in evaluating the Blue Waters system, deployed by Cray in 2012, and in ensuring that the system achieved sustained petascale performance on applications from several areas. However, some of the original SPP codes did not have significant use, or underwent continuous optimizations. Hence, NCSA prepared an updated SPP suite containing codes that more closely reflect the recent workload observed on Blue Waters. NCSA is also creating a public website with source codes, input data, build/run scripts and instructions, plus performance results of this updated SPP suite. This paper describes the characteristics of those codes and analyzes their observed performance, pointing to areas of potential enhancements on modern systems.

16:40-18:20 BoF 10A

Programming Environments, Applications, and Documentation SIG Meeting

(Hadri)

Programming Environments, Applications, and Documentation (PEAD) Special Interest Group meeting.

16:40-18:20 BoF 10B

A BoF - "Bursts of a Feather" - Burst Buffers from a Systems Perspective

(Paul, Bent, Kudryavtsev)

This BoF will bring interested parties together to discuss Burst Buffers- how they are integrated into today's supercomputers, what are their underpinnings and use cases, and what lies ahead for this exciting new high performance I/O technology. Their optimal integration into existing application workflows and supercomputer architectures continues to evolve. There are at least three possible burst buffer architectures: 1.) compute node-local burst buffers; 2.) dedicated shared burst buffers; and 3.) in-storage embedded burst buffers. We will attempt to compare these three architectures from the perspectives of performance, contention, resilience, and usability. This brief intro will be based off collaborative research with colleagues at Los Alamos National Labs. Burst Buffer resources are implemented on our systems and architected for simplified user access. Cray's DataWarp product combined with SchedMD's Slurm WLM will be the primary focus, participants may have differing solution experiences that would allow us to compare and contrast implementations. Details of SSD technology, system daemons, networking and resource management will be presented for discussion. Experiences with implementation, monitoring and problem identification/resolution will be solicited for the benefit of attendees. What is the future for Burst Buffers? One option is utilizing the NVMe over Fabric (NVMeF) protocol for low latency remote

access to NVMe SSDs as a building block for "/tmp on demand" solutions or burst buffer implementation. There are however some requirements for newer kernel features to support this integration.

16:40-17:30 BoF 10C

XC System Management Usability BOF

(Longley, Landsteiner)

This BOF will be a facilitated discussion around usability of system management software on an XC system with SMW 8.0/ CLE 6.0 software focusing on standard and advanced administrator use cases. The goal will be to gain an understanding of the best and worst parts of interacting with XC System Management software and to understand how customers would like to see the software evolve. The intended audience of this BOF will have some knowledge of Linux and system management of XC systems. The discussion will be broadly applicable to a wide audience. The examples will be particularly appropriate for site administrators and others with a more detailed knowledge of installing, configuring, booting, and upgrading a Cray XC system.

17:30-18:20 BoF 10C

HPC Storage Operations: from experience to new tools

(Chesi, Declerck, Treiber, Olchowik)

Managing storage systems at large scale is a challenging duty: Detecting incidents, correlating events, troubleshooting strange I/O behaviors, migrating data, planning maintenances, testing new technologies, dealing with users requests ... Starting from Storage Operations experiences we will discuss about the tools that are available today for Storage Administrators to monitor and control the I/O subsystem of HPC clusters and about the development of new features and tools that may satisfy current and future needs coming together with the data analytics and cloud computing trends.

16:40-17:30 BoF 10D

Open Discussion with CUG Board

(Hancock)

This session is designed as an open discussion with the CUG Board but there are a few high level topics that will also be on the agenda. The discussion will focus on corporation changes to achieve non-profit status (including bylaw changes), feedback on increasing CUG participation, and feedback on SIG structure and communication. An open floor question and answer period will follow these topics. Formal voting (on candidates and the bylaws) will open after this session, so any candidates or members with questions about the process are welcome to bring up those topics.

17:30-18:20 BoF 10D

Holistic Systems Monitoring and Analysis

(Showerman, Gentile, Brandt)

This BOF will be used to improve collaborations in the monitoring and analysis of Cray systems. It will include updates and future directions from many of the sites represented within the CUG monitoring working group. A goal is to improve the number and content of tool-and-technique quick start guides being developed by the Cray monitoring community. This will help the community to gather both lessons learned and requirements for future deployments of the full spectrum of Cray resources. This BoF is organized by the Cray System Monitoring Working Group (SMWG).

Sessions and Abstracts

Wednesday, May 10th

08:30-09:40 General Session 11

CUG Business
(Hancock)

CUG Business Meeting

Invited Talk: Perspectives on HPC and Enterprise High Performance Data Analytics. *(Kolster)*

Mr. Kolster will present his experience of blending HPC and enterprise architectures to solve real-time, web-scale analytics problems and discuss the need to bridge the gap between HPC and enterprise. His unique perspective illustrates the need for enterprise to embrace HPC technologies and vice versa.

09:40-09:50 Sponsor Talk 12

[Seagate] The Effects Fragmentation and Remaining Capacity has on File System Performance
(Fragalla)

After a Lustre file system is put in production, and the usage model increases with the variability of users deleting and creating files overtime, fragmentation and the available storage capacity has effect on overall performance throughput, compared to a pristine file system. In this presentation, Seagate will discuss the methodology of the benchmark setup how to fill up the storage capacity of a file system and introduce fragmentation at different capacity points to analyze the throughput performance impact. The presentation will illustrate how at various percentages of a capacity filled file system, the biggest impact is the amount of fragmentation that exists and not only the capacity filled.

09:50-10:00 Sponsor Talk 13

[SchedMD] Slurm Roadmap
(Jette)

Slurm is an open source, fault-tolerant, and highly scalable cluster management and job scheduling system used on many of the largest computers in the world including five of the top ten systems on the TOP500 supercomputer list. This presentation will briefly describe Slurm capabilities with respect to Cray systems and the Slurm roadmap.

10:30-11:50 General Session 14

CUG Best Paper Presentation
(tbd)

tbd

Plenary talk from Intel: Exascale Reborn
(Hazra)

Join Intel corporate vice president and general manager of the Enterprise and Government group, Dr. Rajeeb Hazra, for our plenary talk. Raj will discuss what must be done to address an ecosystem of ever changing, complex, and diversified applications and workloads to deliver real world performance at exascale scale.

11:50-12:00 Sponsor Talk 15

[PGI] OpenACC and Unified Memory
(Miles)

Optimizing data movement between host and device memories is an important step when porting applications to GPUs. This is true for any programming model (CUDA, OpenACC,

Sessions and Abstracts

OpenMP 4+, ...), and becomes even more challenging with complex aggregate data structures. While OpenACC data management directives are designed so they can be safely ignored on a shared memory system with a single address space, such as a multicore CPU, both the CUDA and OpenACC APIs require the programmer or compiler to explicitly manage data allocation and coherence on a system with separated memories. The OpenACC committee is designing directives to extend explicit data management for aggregate data structures. CUDA C++ has managed memory allocation routines and CUDA Fortran has the managed attribute for allocatable arrays, allowing the CUDA driver to manage data movement and coherence for dynamically allocated data. The latest NVIDIA GPUs include hardware support for fully unified memory, enabling operating system and driver support for sharing of the entire address space between the host CPU and the GPU. We will compare current and future explicit data movement with driver- and system-managed memory, and discuss the impact of these on application development, programmer productivity and performance.

13:00-13:10 Sponsor Talk 16

[Allinea] Tools and Methodology for Ensuring HPC Programs Correctness and Performance

(Paisley)

In this presentation we will discuss best practices and methodology for HPC software engineering. We will provide illustrations of how the Allinea debugging and performance analysis tools can be used to ensure that you obtain optimal performance from your codes and that your codes run correctly.

13:10 - 13:20 Sponsor Talk 17

[Altair] PBS Professional - Stronger, Faster, Better!

(Suchyta)

Not all workloads are created equal. They vary in size, duration, priority level, and are influenced by many other site-specific factors. Such requirements often lead administrators to compromise system utilization, service level agreements, or even limit the capabilities of the system itself. Altair continues to provide features and flexibility allowing administrators to control how jobs are scheduled without making dramatic compromises. PBS Professional v13 is architected for Exascale with increased speed, scale, resiliency, and much more. This presentation will provide a high-level overview of some of the key new PBS capabilities such as Multi-Scheduler, Preemption Targets, and flexibility & performance enhancements.

13:20-14:30 General Session 18

1 on 100 or More...

(Ungaro)

Open discussion with Cray President and CEO. No other Cray employees or Cray partners are permitted during this session.

15:00-17:00 Technical Session 19A

The Cray Programming Environment: Current Status and Future Directions

(DeRose)

The scale and complexity of current and future high end systems with wide nodes, many integrated core (MIC) architectures, multiple levels in the memory hierarchy, and heterogeneous processing elements brings a new set of challenges for application developers. These technology changes in the supercomputing industry are forcing

computational scientists to face new critical system characteristics that will significantly impact the performance and scalability of applications. Users must be supported by intelligent compilers, automatic performance analysis and porting tools, scalable debugging tools, and adaptive libraries. In this talk I will present the recent activities, new functionalities, roadmap, and future directions of the Cray Programming Environment, which are being developed and deployed on Cray Clusters and Cray Supercomputers for scalable performance with high programmability.

Current State of the Cray MPT Software Stacks on the Cray XC Series Supercomputers

(Kandalla, Mendygral, Ravichandrasekaran, Radcliffe, Cernohous, McMahon, Sadlo, Pagel)

HPC applications heavily rely on Message Passing Interface (MPI) and SHMEM programming models to develop distributed memory parallel applications. This paper describes a set of new features and optimizations that have been introduced in Cray MPT software libraries to optimize the performance of scientific parallel applications on modern Cray XC series supercomputers. For Cray XC systems based on the Intel KNL processor, Cray MPT libraries have been optimized to improve communication performance, memory utilization, while also facilitating better utilization of the MCDRAM technology. Cray MPT continues to improve the performance of hybrid MPI/OpenMP applications that perform communication operations within threaded regions. The latest Cray MPICH offers a new lock-ahead optimization for MPI I/O along with exposing internal timers and statistics for I/O performance profiling. Finally, this paper describes efforts involved in optimizing real-world applications such as WOMBAT and SNAP, along with Deep Learning applications on the latest Cray XC supercomputers.

Profiling and Analyzing Program Performance Using Cray Tools

(Poxon)

The Cray Performance Tools help the user obtain maximum computing performance on Cray systems with profiling and analysis that focuses on discovering key bottlenecks in programs that run across many nodes. The tools robust analysis capability helps users identify program hot spots, imbalance, communication patterns, and memory usage issues that impede scaling or optimal performance. As an example, CrayPAT and Cray Apprentice2 were recently used to scale the CNTK deep learning code to new levels for the system at the Swiss National Supercomputing Centre (CSCS). In addition to focusing on simple interfaces to make profiling and analysis accessible to more users, recent enhancements to CrayPAT, Cray Apprentice2 and Reveal include the new HBM memory analysis tool that identifies key arrays that can benefit from allocation in KNL's MCDRAM, a per-NUMA node memory high water mark, general Intel KNL and NVIDIA P100 support, as well as profiling support for Charm++.

Novel approaches to HPC user engagement

(Barrass, Henty)

EPCC operates the UK National HPC service ARCHER, a Cray XC30 with a diverse user community. A key challenge to any HPC provider is growing the user base, making new users aware of the potential benefits of the service and ensuring a low barrier to entry. To this end, we have explored a number of

approaches to user engagement that are novel within the context of UK HPC: - promoting the benefits of HPC via a network of ARCHER Champions; - giving easy access to ARCHER through an online driving test; - supporting novice users with screencast videos for common tasks; - regular long-term training impact surveys to quantify the benefits of the training programme. Here we choose to focus on those aspects of user support and engagement that we believe are novel and have proven particularly successful, rather than the more standard activities common to most HPC services.

15:00-17:00 Technical Session 19B

Improving I/O Bandwidth With Cray DVS Client-Side Caching

(Hicks)

Cray's Data Virtualization Service, DVS, is an I/O forwarder providing access to native parallel filesystems and to Cray Datawarp application I/O accelerator storage at the largest system scales while still maximizing data throughput. This paper introduces DVS Client-Side Caching, a new option for DVS to improve I/O bandwidth, reduce network latency costs, and decrease the load on both DVS servers and backing parallel filesystems. Client-side caching allows application writes to target local in-memory cache on compute nodes. This provides low latency and high throughput for write operations. It also allows aggregation of data to be written back to the filesystem so fewer network and parallel filesystem operations are required. Caching also enables applications to reuse data previously read or written without further network overhead. The paper will discuss motivations for this work, detailed design and architecture, acceptable use cases, benchmark testing results, and possibilities for future improvement.

Implementing a Hierarchical Storage Management system in a large-scale Lustre and HPSS environment

(Bode, Butler, Glasgow, Stevens, Schumann, Zago)

HSM functionality has been available with Lustre for several releases and is an important aspect for HPC systems to provide data protection, space savings, and cost efficiencies, and is especially important to the NCSA Blue Waters system. Very few operational HPC centers have deployed HSM with Lustre, and even fewer at the scale of Blue Waters. This paper will describe the goals for HSM in general and detail the goals for Blue Waters. The architecture in place for Blue Waters, the collaboration with Cray, priorities for production and existing challenges will be detailed as well in the paper.

Understanding the IO Performance Gap Between Cori KNL and Haswell

(Liu, Koziol, Tang, Tessier, Bhimji, Cook, Byna, Austin, Lockwood, Deslippe, Prabhat)

Cori system has a 2,004 nodes Haswell partition in phase one, and a 9,688 nodes KNL partition in phase two, which together forms the 5th most powerful and fastest supercomputer in the world by April 2017. Understanding the IO gap between the two partitions is important, not only to NERSC/LBNL users and other national labs (e.g., Argonne National Laboratory, which also observed the similar IO gap), but also to hardware (many core chips) and software (Lustre, MPI-IO) developers. In this paper, we have analyzed IO performance of single core and single node comprehensively on Haswell and KNL, through which, we discovered the major bottlenecks, which include CPU frequencies and memory copy. We have also

extended our performance tests to multi-nodes, and revealed the IO cost difference caused by network latency, buffer size, and communication cost, etc. Overall, we have a better understanding of the IO gap between Haswell and KNL, and the lessons learned in this exploration would guide us in designing optimal IO solutions in many-core era.

DXT: Darshan eXtended Tracing

(Xu, Snyder, Kulkarni, Venkatesan, Carns, Byna, Sisneros, Chadalavada)

As modern supercomputers evolve to exascale, their I/O subsystems are becoming increasingly complex, making optimization of I/O for scientific applications a daunting task. Although I/O profiling tools facilitate the process of optimizing application I/O performance, legacy profiling tools lack flexibility in their level of detail and ability to correlate traces with other sources of data. Additionally, a lack of robust trace analysis tools makes it difficult to derive actionable insights from large-scale I/O traces. Darshan is an HPC I/O characterization tool that records statistics in a lightweight manner that makes it appropriate for full-time production deployment. However, Darshan's default characterization mechanism records information at a fixed granularity. We augment Darshan by proposing Darshan eXtended Tracing (DXT) for more detailed profiling of I/O software stacks. DXT enables users and administrators to vary the level of fidelity captured by Darshan at run time without modifying or recompiling applications. This capability facilitates systematic analysis on the I/O behavior of applications and can provide useful application kernel I/O traces to help advance parallel I/O research. We have demonstrated the power of DXT by obtaining a wide range of useful statistics for multiple case studies, and we further show that DXT is able to do so same with negligible overhead.

15:00-17:00 Technical Session 19C

Scheduler Optimization for Current Generation Cray Systems

(Jette, Jacobsen, Paul)

The current generation of Cray systems introduces two major complications for workload scheduling: DataWarp burst buffers and Intel Knights Landing (KNL) processors. In a typical use case, DataWarp resources are allocated to a job and data is staged-in before compute resources are allocated to that job, then retained after computation for staging-out of data. KNL supports five different NUMA modes and three MCDRAM modes. Applications may require a specific KNL configuration or its performance may be highly configuration dependent. If KNL resources with the desired configuration are not available for pending work, the overhead of rebooting compute nodes must be weighed against running the application in a less than ideal configuration or waiting for processors already in the desired configuration to become available. This paper will present the algorithms used by the Slurm workload manager. a statistical analysis of NERSC's workload and experiences with Slurm's management of DataWarp and KNL.

Trust Separation on the Cray XC40 using PBS Pro

(Clarke)

As the UK's national weather agency, the Met Office has a requirement to produce regular, timely weather forecasts. As a major centre for climate and weather research, it has a need to provide access to large-scale supercomputing resources to users from within the organisation. It also provides a supercomputer

facility for academic partners inside the UK, and to international collaborators. Each of these user categories has a different set of availability requirements and requires a different level of access. This paper describes the steps taken to create an HPC facility that separates these different requirements using soft partitions created by the batch system. We detail our initial experiences with cgroup containers and our use of custom PBS hooks to partition the Lustre name space. We summarise some of the problems observed during implementation, comment on the scalibilty of the solution and outline possible future enhancements.

Experiences running different work load managers across Cray Platforms

(Ayyalasomayajula, West)

Workload management is a challenging problem both in Analytics and in High Performance Computing. The desire is to have efficient platform utilization while still meeting scalability and scheduling requirements. SLURM and Moab/Torque are two commonly used workload managers that serve both resource allocation and scheduling requirements on the Cray CS and XC series super computers. Analytics applications interact with a different set of workload managers such as YARN or more recently, Apache Mesos, which is the main resource manager for Urika-GX. In this paper, we describe our experiences using different workload managers across Cray platforms (Analytics and HPC). We describe the characteristics and functioning of each of the workload managers. We will compare the different workload managers and specifically discuss the pros and cons of the HPC schedulers vs. Mesos, and run a sample workflow on each of the Cray platforms and illustrate resource allocation and job scheduling.

An Operational Perspective on a Hybrid and Heterogeneous Cray XC50 System

(Alam, Bianchi, Cardo, Chesi, Gila, Gorini, Klein, Passerini, Ponti, Verzelloni)

The Swiss National Supercomputing Center (CSCS) has been in the process of upgrading its flagship system called Piz Daint in order to support a wide range of services. The upgraded system is a heterogeneous Cray XC50 and XC40 system with Nvidia GPU accelerated (Pascal) devices as well as multi-core nodes with diverse memory configuration. Despite the state-of-the-art hardware and the design complexity, the system was built in a matter of weeks and was returned to full service to the CSCS user communities in less than two months, while at the same time with providing an order of improvement in energy efficiency. This paper focuses on the innovative features of the Piz Daint system that not only resulted in an adaptive, scalable and stable platform but also offer a very high level of operational robustness for a complex ecosystem.

17:10-18:20 BoF 20A

Sonexion Monitoring and Metrics: data collection, data retention, user workflows

(Langer, Flaskerud)

This BOF will explore types of metrics useful in analyzing performance issues on Cray Sonexion storage systems, data retention and reduction concerns considering the volume of metrics that can be collected, and typical workflows administrators use to analyze and isolate Sonexion performance issues related to jobs launched from their Cray HPC systems.

Sessions and Abstracts

17:10-18:20 BoF 20B

eLogin Usability and Best Practices

(Keopp, Ahlstrom, Longley)

This BOF session is a facilitated discussion around usability and best practices for administrating and configuring eLogin nodes. eLogin nodes are the external login nodes for Cray XC systems running CLE 6.x. They replace the esLogin nodes used with CLE 5.x. The discussion will also include the new Cray Management Controller (CMC) and Cray System Management Software (CSMS) which replace the current CIMS and Bright Cluster Manager software for managing eLogin nodes. The goal will be to gain an understanding of the best and worst parts of administrating eLogin nodes and to understand how customers would like to see the software evolve. The intended audience of this BOF will have some knowledge of Linux and system management of eLogin nodes. The discussion will be broadly applicable to a wide audience. The examples will be particularly appropriate for site administrators and others with a more detailed knowledge of installing, configuring, booting and upgrading Cray CMC/eLogin systems.

17:10-18:20 BoF 20C

Building an Enterprise-Grade Deep Learning Environment with Bright and Cray

(Stober)

Enterprises are collecting increasing amounts of data. By leveraging deep and machine learning technologies, the analysis of corporate data can be taken to the next level, providing organizations with richer insight to their business, resulting in increased sales and / or significant competitive advantage. When business advantage is tied to the insights achieved via deep learning, it is essential for the underlying IT infrastructure to be deployed and managed as enterprise-grade, not as a lab experiment. However, building and managing an advanced cluster, installing the software that satisfies all of the library dependencies, and making it all work together, presents an enormous challenge. In this birds of a feature session, Bright Computing will lead an interactive discussion into how to simplify the build and deployment of an enterprise-grade deep learning environment, enabling organizations to focus on gaining actionable insights from rich, complex data. The group will look at how best to find, configure, and deploy all of the dependent pieces needed to run deep learning libraries and frameworks, in order to gain advantage from the deep learning evolution.

17:10-18:20 BoF 20D

Bringing "Shifter" to the Broader Community

(Cardo, Jacobsen)

The success and popularity of using "Shifter", developed by NERSC, for containers continues to grow. In order to keep pace with this success and growth of "Shifter", a community is forming behind it. During this BoF, we'll discuss the status, efforts, and opportunities of bringing "Shifter" to an Open Source Community. Topics will include, organisational structure, software management, and levels of participation.

Thursday, May 11th

08:30-09:25 General Session 21

Panel: Future Directions of Data Analytics and High Performance Computing

(He, Michael, Prabhat, Sukumar, Luraschi,

Sessions and Abstracts

Rangarajan)

Panel Discussion: Future Directions of Data Analytics and High Performance Computing The past several years have seen continued growth of Big Data software, such as the Apache Big Data Stack, and new interest in machine learning algorithms on high performance compute architectures. Much of the work in this area centers around cloud-like deployment models and platform architectures, while more traditional HPC simulation infrastructures have changed little. In this panel we will discuss the how the intersection of HPC architectures and operations continues to evolve to support Big Data and machine learning frameworks and approaches. We focus on data analytics applications and software and how those tools can best be adopted and adapted to productive use in today's HPC data center. We also look at how HPC infrastructure and operations might be altered to better accommodate data analytic practices. There is tremendous opportunity for both the data analytics and HPC communities in adopting HPC experience to port the analytics codes for scalable performance. We will discuss the adoption and benefits of HPC systems, software and programming best practices in emerging Big Data applications in graph analytics, deep learning and adversarial analytics. Moderator: Helen He, NERSC, Lawrence Berkeley National Laboratory. Panelists: Scott Michael, Indiana University; Mr Prabhat, NERSC, Lawrence Berkeley National Laboratory; Rangan Sukumar, Cray; Javier Luraschi, RStudio; Radhika Rangarajan, Intel.

09:25-09:40 General Session 22

CUG2018 Site Presentation

(Hancock)

This is the plenary session to hear about the presentation from CUG2018 site representative.

09:40-09:50 Sponsor Talk 23

[Adaptive Computing] Reporting and Analytics, Portal-based Job Submission, Remote Visualization, Accounting and High Throughput Task Processing on Torque and Slurm

(Ihli)

Adaptive Computing will present on Reporting & Analytics, Viewpoint (portal-based job submission), Remote Visualization, Accounting and Nitro (high throughput task processing) on Torque and Slurm as it covers how the "Open Platform" initiative helps bring Enterprise solutions to your choice of scheduler. Further, Adaptive Computing will present on significant new Torque advancements intended in its platform "Unification" initiative, as well as advancements in power management, datawarp integration and other product enhancements.

09:50-10:00 Sponsor Talk 24

[Bright Computing] Achieving a Dynamic Datacenter with Bright and Cray

(Stober)

IT teams are under pressure to manage an emerging range of computing-intensive and data-intensive workloads with very different characteristics from traditional IT. Executing these workloads means that companies need to master multiple technologies ranging from high performance computing and big data analytics to virtualization, containerization, and cloud. In this presentation, Bright Computing will explain how organizations

Sessions and Abstracts

can leverage Bright and Cray to benefit from connecting these seemingly diverse and disparate technologies into a dynamic and integrated clustered computing environment. Bright will examine the emergence of Linux clustering in the enterprise datacenter, and propose a set of criteria to consider when evaluating options for managing clustered IT infrastructure in a dynamic datacenter.

10:30-12:00 Technical Session 25A

Porting the microphysics model CASIM to GPU and KNL Cray machines

(Brown, Nigay, Shipway, Hill, Weiland)

CASIM is a microphysics model used to investigate interactions at the millimetre scale and study the formation and development of moisture. This is a crucial aspect of atmospheric modelling and as such is used as a sub model by other models but is computationally intensive and can severely impact the runtime of these models. We will discuss the porting of CASIM onto GPUs and optimisation for the KNL. Using Cray's OpenACC for the GPU port, the code complexity required that we offload over 120 functions found in various modules which raised specific challenges we had to solve. A significant increase in performance was found with the upgrade of Piz Daint from XC30 to XC50, the P100 making GPU CASIM an attractive approach. We will discuss profiling results from the GPU runs and a comparison will be made with running the model on KNL as well as optimisations for that platform.

An in-depth evaluation of GCC's OpenACC implementation on Cray systems

(Vergara Larrea, Elwasif, Hernandez, Philippidis, Allen)

In this study, we will perform and in-depth evaluation of GCC's OpenACC implementation on ORNL's Cray XK7 and compare it with other available implementations. The results presented will be useful for the larger community interested in using and evaluating new OpenACC implementations. Finally, a discussion on how an OpenACC implementation in GCC may help the interoperability of both OpenACC and OpenMP 4.5 (offload) specifications will be presented.

HPCG and HPGMG benchmark tests on Multiple Program, Multiple Data (MPMD) mode on Blue Waters - a Cray XE6/XK7 hybrid system

(Kwack, Bauer)

The High Performance Conjugate Gradients (HPCG) and High Performance Geometric Multi-Grid (HPGMG) benchmarks are alternatives to the traditional LINPACK benchmark (HPL) in measuring the performance of modern HPC platforms. We performed HPCG and HPGMG benchmark tests on a Cray XE6/XK7 hybrid supercomputer, Blue Waters at National Center for Supercomputing Applications (NCSA). The benchmarks were tested on CPU-based and GPU-enabled nodes separately, and then we analyzed characteristic parameters that affect their performance. Based on our analyses, we performed HPCG and HPGMG runs in Multiple Program, Multiple Data (MPMD) mode in Cray Linux Environment in order to measure their hybrid performance on both CPU-based and GPU-enabled nodes. We observed and analyzed several performance issues during those tests. Based on lessons learned from this study, we provide

recommendations about how to optimize science applications on modern hybrid HPC platforms.

10:30-12:00 Technical Session 25B

Project Caribou : Monitoring and Metrics for Sonexion

(Flaskerud)

The scale and number of subsystems in today's High Performance Computing system deployments make it difficult to monitor application performance and determine root causes when performance is not what is expected. System component failures, system resource oversubscription, or poorly written applications can all contribute to systems not running as expected and thus to poorly performing applications. This problem is exacerbated by the need to mine information from multiple sources across system subcomponents. Collecting the data may require privileged access and the data must be collected in a timely manner or critical information can be lost. Many of Cray's large customers have created their own solutions for monitoring Cray system resources to address these challenges. They have relied on available log files and created their own scripts and tools to collect, consolidate, and persist information. Each of these customers has similar needs; a toolset which will: collect relevant information from the different Cray systems, persist this information to allow for current and historical analysis, and present customized reports and alerts, all of which together allow administrators to proactively address failures or degradations in performance. Caribou is a new monitoring and metric software solution created by Cray to help customers address this problem. Caribou initially focuses on collecting and persisting performance and job metrics specific to the Cray Sonexion storage system, correlating this with job application information collected from the Cray computing systems. Caribou is installed on a customer's on-site standalone server that is connected to the customer's network. Caribou collects Lustre and jobstats metrics, system logs, and system events from each Sonexion storage system configured to be monitored. If Caribou is connected to the Sonexion high speed InfiniBand network, it will discover all HCAs and switches on the fabric and collect and persist port counters from the IB fabric. Caribou is integrated with the Cray System Management Workstation (SMW), collecting job information such as start/stop, job id for jobs that were launched using a workload manager. This information is persisted into a time-series database on the customer's Caribou server and persisted using a standardized data model. Administrators are able to view this information in "real-time" or look at information collected at different points in the past through user friendly dashboards and workflows.

Lustre Lockahead: Early Experience and Performance using Optimized Locking

(Moore, Farrell, Cernohous)

Recent Cray-authored Lustre modifications known as "Lustre lockahead" show significantly improved write performance for collective, shared-file I/O workloads. Initial tests show write performance improvements of more than 200% for small transfer sizes and over 100% for larger transfer sizes compared to traditional Lustre locking. Standard Lustre shared-file locking mechanisms limit scaling of shared file I/O performance on modern high performance Lustre servers. The new lockahead feature provides a mechanism for applications (or libraries) with knowledge of

their I/O patterns to overcome this limitation by explicitly requesting locks. MPI-IO is able to use this feature to dramatically improve shared file collective I/O performance, achieving more than 80% of file per process performance. This paper discusses our early experience using lockahead with applications. We also present application and synthetic performance results and discuss performance considerations for applications that benefit from lockahead.

A High Performance SVD Solver on Manycore Systems

(Sukkari, Ltaief, Esposito, Keyes)

We describe the high performance implementation of a new singular value decomposition (SVD) solver for dense matrices on distributed-memory manycore systems. Based on the iterative QR dynamically-weighted Halley algorithm (QDWH), the new SVD solver performs more floating-point operations than the bidiagonal reduction variant of the standard SVD, but exposes at the same time more parallelism, and therefore, runs closer to the theoretical peak performance of the system, thanks to more compute-bound matrix operations. The resulting distributed-memory QDWH-SVD solver is more numerically robust in presence of ill-conditioned large matrix sizes and achieves up to fourfold speedup on thousands of cores against current state-of-the-art SVD implementation from Cray Scientific Library.

10:30-12:00 Technical Session 25C

Cray XC40 System Diagnosability: Functionality, Performance, and Lessons Learned

(Schutkoske)

The Intel Xeon PhiTM CPU 7250 processor presents new opportunities for diagnosing the node in the Cray XC40TM system. The Intel Xeon PhiTM CPU 7250 processor supports a new high-bandwidth on-package MCDRAM memory and interfaces. It also supports the ability to support different Non-Uniform Memory Access (NUMA) configurations. The new Cray Processor Daughter Card (PDC) also supports an optional PCIe SSD card. Previous work [1] [2] [3] has outlined Cray system diagnosability for the Cray XCTM Series. The Intel Xeon PhiTM CPU 7250 processor requires new BIOS, administrative commands, power and thermal limits, as well as, new diagnostics to validate functionality and performance. The Hardware Supervisory System (HSS) supports the new high-bandwidth on-package MCDRAM memory and interfaces. It supports the On-Demand configuration of the MCDRAM and NUMA. The MCDRAM and NUMA configurations, as well as SSD enable/disable, are configurable from both the command line on the System Management Workstation (SMW) and by a workload manager (WLM) running under the Cray Linux Environment (CLE) utilizing the Cray Advanced Platform Monitoring and Control (CAPMC) interface. New Intel Xeon PhiTM CPU 7250 processor on-line diagnostics have been written to validate the node and MCDRAM functionality and performance. The diagnostics validate the node based on the MCDRAM and NUMA configurations. The Workload Test Suite (WTS) has also been updated to detect and diagnose Intel Xeon PhiTM CPU 7250 processor problems under CLE. There are new utilities and diagnostics to support the PCIe SSD card. There is a new diagnostic utility that executes in CLE on the Data Warp Node or the compute node to support the SSD. This diagnostic utility is periodically scheduled to check the health of the SSD. It reports the status to the SMW via RCA. The system administrator can query and display the current SSD health,

as well as, historical data. The results of the SSD diagnostic utility can also be viewed on the SMW. This paper describes the tool chain changes required to support the new blade with the Intel Xeon PhiTM CPU 7250 processor and optional PCIe SSD cards. It also provides detailed examples on how to diagnose Intel Xeon PhiTM CPU 7250 processor faults within the Cray XC40TM system.

KNL System Software

(Hill, Snyder, Sygulla)

Intel Xeon Phi "Knights Landing" (KNL) presents opportunities and challenges for system software. This paper starts with an overview of KNL architecture. We describe some of the key differences from traditional Xeon processors, such as processor (NUMA) and memory (MCDRAM) modes. We describe which KNL modes are most useful, and why. From there, we describe a day in the life of a KNL system, emphasizing unique features such as mode reconfiguration (selecting the processor and memory configuration for a job) and the zone sort feature (which optimizes performance of MCDRAM cache). As part of this coverage, we'll look at implementation, scaling and performance issues.

Runtime collection and analysis of system metrics for production monitoring of Trinity Phase II

(DeConinck, Nam, Bonnie, Morton, Lueninghoener, Brandt, Pedretti, Agelastos, Gentile, Allan, Repik)

We present the holistic approach taken by the ACES team in the design and implementation of a monitoring system tailored to the new Cray XC-40 KNL-based Trinity Phase II platform currently being deployed in an Open Science campaign. Our particular area of focus is enabling run-time analysis of system data combinations in order to improve application performance and system efficiency. We present analyses of both per-user (e.g., per node memory and CPU) and shared (e.g., network, storage, power, cooling) resource utilization from production and dedicated runs during Open Science. We present detail on what we are monitoring and why, including the hardware, software, and configuration deployed, to enable collection and appropriate retention of and access to the information of interest. We assess the level of adverse impact on application performance, and positive impact on system operations. Finally we discuss what has worked, lessons learned, and future work.

12:15-13:00 BoF 26

Deep Learning on Cray Platforms

(Sukumar, Prabhat, Martinasso)

The application of machine learning and deep learning has gained tremendous interest both in academic and commercial organizations. With increased adoption and faster rates of data creation and collection, the need for scale on these deep learning problems have arrived. In this Birds-of-a-feather session, we will engage in active discussion around running deep learning workloads on Cray platforms at scale. Researchers from leadership computing facilities along with Cray engineers will be sharing their experiences. More specifically, we will discuss: (i) how to run deep learning workloads on the XC, CS and Urika-GX platforms, (ii) popular DL toolkits (TensorFlow, MXNet, Caffe, CNTK, BigDL etc.), (iii) HPC best practices toward "strong-scaling" deep learning workloads on multi-node configurations, (iv) application of deep learning in the different science domains today and at exascale, (v) experiences with deep

learning on different HPC architectures (IB vs. Aries, CPU vs. GPU, etc.).

13:00-14:30 Technical Session 27A

Next Generation Science Applications for the Next Generation of Supercomputing

(Vaughan, Hammond, Dinge, Lin, Trott, Pase, Cook, Hughes, Hoekstra)

The Trinity supercomputer deployment by Los Alamos and Sandia National Laboratories represents the first Advanced Technology System deployment for the United States National Nuclear Security Administration. It will be one of the largest XC40 deployments in the world when installed during 2017. We present performance analysis of a suite of new applications that have been written from the ground up to be portable across computing architectures, parallel in terms of multi-node and on-node threading and to feature more flexible component-based code design. These applications leverage Kokkos, Sandia's C++ Performance Portability programming mode, the Trilinos linear solver library and our broader performance analysis capabilities in a close knit codesign program. Driven by the NNSA's Advanced Technology Development and Mitigation ("ATDM") program, the new codes represent prototypes of fully-capable production science codes that will execute with high levels of efficiency on the next-generation of supercomputing platforms including Trinity and beyond.

Fusion PIC Code Performance Analysis on The Cori KNL System

(Koskela, Deslippe, Friesen, Raman)

We study the attainable performance of Particle-In-Cell codes on the Cori KNL system by analyzing a miniature particle push application based on the fusion PIC code XGC1. We start from the most basic building blocks of a PIC code and build up the complexity to identify the kernels that cost the most in performance and focus optimization efforts there. Particle push kernels operate at high AI and are not likely to be memory bandwidth or even cache bandwidth bound on KNL. Therefore, we see only minor benefits from the high bandwidth memory available on KNL, and achieving good vectorization is the most beneficial optimization path and can theoretically yield up to 8x speedup on KNL, but is in practice limited by the data layout to 4x.

Performance of MPI/OpenMP Hybrid VASP on Cray XC40 Based on Intel Knights Landing Many Integrated Core Architecture

(Zhao, Marsman, Wende, Kim)

With the recent installation of Cori, a Cray XC40 system with Intel Xeon Phi Knights Landing (KNL) many integrated core (MIC) architecture, NERSC is transitioning from the multi-core to the more energy-efficient many-core era. The developers of VASP, a widely used materials science code, have adopted MPI/OpenMP parallelism to better exploit the increased on-node parallelism, wider vector units, and the high bandwidth on-package memory (MCDRAM) of KNL. To achieve optimal performance, KNL specifics relevant for the build and runtime setup must be explored. In this paper, we present the performance analysis of representative VASP workloads on Cori and focus on compilers, libraries, and runtime options like the NUMA/MCDRAM modes, Hyper-Threading, hugepages, task/thread/memory affinity, core specialization, and thread scaling. The paper will serve as a KNL performance guide for

VASP users and others.

13:00-14:30 Technical Session 27B

Toward a Scalable Bank of Filters for High Throughput Image Analysis on the Cray Urika-GX System

(Shilpika, Ferrier, Vishwanath)

High throughput image analysis is critical for experimental sciences facilities and enables one to glean timely insights of the various experiments and to better understand the physical phenomena being imaged. We present the design and evaluation of the bank of filters, the core building blocks for high throughput image analysis, on the Cray Urika-GX system. We describe our infrastructure developed with Apache Spark. We scaled this to 500 cores of the Urika-GX system for analysis of a Combustion engine dataset imaged at the Advanced Photon Source at Argonne National Laboratory and observe significant speedups. This scalable infrastructure now opens the doors to the application of a wide range of image processing algorithms and filters to the large-scale datasets being imaged at various light sources.

Towards Seamless Integration of Data Analytics into Existing HPC Infrastructures

(Hoppe, Gienger, Boenisch, Moise, Shcherbakov)

Customers of the High Performance Computing Center (HLRS) tend to execute more complex and data-driven applications, often resulting in large amounts of data of up to 1 Petabyte. The majority of our customers, however, is currently lacking the ability and knowledge to process this amount of data in a timely manner in order to extract meaningful information. We have therefore established a new project in order to support our users with the task of knowledge discovery by means of data analytics. We put the high performance data analytics system, a Cray Urika-GX, into operation to cope with this challenge. In this paper, we give an overview about our project and discuss immanent challenges in bridging the gap between HPC and data analytics in a production environment. The paper concludes with a case study about analyzing log files of a Cray XC40 to detect variations in system performance. We were able to identify successfully so-called aggressor jobs, which reduce significantly the performance of other simultaneously running jobs.

Quantifying Performance of CGE: A Unified Scalable Pattern Mining and Search System

(Maschhoff, Vesse, Sukumar, Ringenburg, Maltby)

CGE was developed as one of the first applications to embody our vision of an analytics ecosystem that can be run on multiple Cray platforms. This paper presents Cray Graph Engine (CGE) as a solution that addresses the need for a unified ad-hoc subject-matter driven graph-pattern search and linear-algebraic graph analysis system. We demonstrate that the CGE implemented using the PGAS parallel programming model performs better than most off-the-shelf graph query engines with ad-hoc pattern search while also enabling the study of graph-theoretic spectral properties in runtimes comparable to optimized graph-analysis libraries. Currently CGE is provided with the Cray Urika-GX and can also run on Cray XC systems. Through experiments, we show that compared to other state-of-the-art tools, CGE offers strong scaling and can often handle graphs three orders of magnitude larger, more complex datasets (long diameters, hypergraphs, etc.), and more computationally intensive complex pattern

searches.

13:00-14:30 Technical Session 27C

Application-Level Regression Testing Framework using Jenkins

(Bouvet, Budiardja, Arnold)

This paper will explore the challenges of regression testing and monitoring of large scale systems such as NCSA's Blue Waters. Our goal was to come up with an automated solution for running user-level regression tests to evaluate system usability and performance. Jenkins was chosen for its versatility, large user base, and multitudes of plugins including plotting test results over time. We utilize these plots to track trends and alert us to system-level issues before they are reported by our partners (users). Not only does Jenkins have the ability to store historical data, but it can also send email or text messages based on results of a test. Other requirements we had include two-factor authentication for accessing the Jenkins GUI with administrator privileges and account management through LDAP. In this paper we describe our experience in deploying Jenkins as a user-level system-wide regression testing and monitoring framework for Blue Waters.

Data-Driven Understanding of Fault Scenarios and Impacts Through Fault Injection: Experimental Campaign in Cielo

(Formicola, Jha, Deng, Chen, Bonnie, Mason, Greiner, Gentile, Brandt, Kaplan, Repik, Enos, Showerman, Kalbarczyk, Kramer, Iyer)

We present a set of fault injection experiments on the ACES (LANL/SNL) Cray XE supercomputer Cielo. We use this experimental campaign to improve the understanding of failure causes and propagation that we observed in the field failure data analysis of NCSA's Blue Waters. We use the data collected from the logs and from network performance counter data to 1) characterize the fault-error-failure sequence and recovery mechanisms in the Gemini network and in the Cray compute nodes, 2) understand the impact of failures on the system and the user applications at different scales, 3) identify and recreate fault scenarios that induce unrecoverable failures, in order to create new tests for system and application design. The faults were injected through special input commands to bring down network links, directional connections, nodes, and mezzanines. We present expected extensions necessary to apply our methodologies of injection and analysis to the Cray XC (Aries) systems.

A regression framework for checking the health of large HPC systems

(Karakasis, Jocksch, Holanda Rusu, Peretti-Pezzi)

We present a framework to facilitate the sanity checking of Cray HPC systems. It allows the user to write their own regression tests in a compact way. The tests are simple Python classes and the framework takes care of the low level interactions with the system, i.e., managing the jobs and parsing the results. Thus the tests can be maintained and adapted to new systems easily, they require a basic knowledge of python only. The regression suite comprises currently 109 regression checks resulting in more than 250 test cases for the different programming environments. It is run before and after maintenance of Piz Daint with tests checking the whole system for sanity including performance. In addition it is run daily with a series of application checks. This procedure improves the reliability of the system and reduces the downtime.

Sessions and Abstracts

15:00-16:30 Technical Session 28A

HPC Containers in use

(Sparks)

Linux containers in the commercial world are changing the landscape for application development and deployments. Container technologies are also making inroads into HPC environments, as exemplified by NERSC's Shifter and LBL's Singularity. While the first generation of HPC containers offers some of the same benefits as the existing open container frameworks, like CoreOS or Docker, they do not address the cloud/commercial feature sets such as virtualized networks, full isolation, and orchestration. This paper will explore the use of containers in the HPC environment and summarize our study to determine how best to use these technologies in the HPC environment at scale.

Fast and consistent HPC workflows using containers

(Benedicic, Cruz, Schulthess)

In this work we describe the experiences of building and deploying containers using Docker and Shifter, respectively. We will present basic benchmarking tests that show the performance portability of certain workflows as well as performance results from the deployment of widely used non-trivial scientific applications. Furthermore, we discuss the resulting workflows through use cases that cover from the container creation on a laptop to their deployment at scale while taking advantage of specialized hardware: Cray Aries interconnect and NVIDIA Tesla P100 GPU accelerators.

ExPBB: A framework to explore the performance of Burst Buffer

(Markomanolis)

ShaheenII supercomputer provides 268 Burst Buffer nodes based on Cray DataWarp (DW) technology. Thus, there is an extra layer between the compute nodes and the parallel filesystem by using SSDs. However, this technology is new and many scientists try to understand and gain the maximum performance. We present a framework called Explore the Performance of Burst Buffer (ExPBB). The purpose of this project is to determine the optimum parameters to acquire the maximum performance of the executed applications on the Burst Buffer. We study the number of the used DW nodes, MPI aggregators, striping unit of files, and MPI/OpenMP processes. The framework aggregates I/O performance data from the Darshan tool and MPI I/O statistics provided by Cray MPICH, then it proceeds to the study of the parameters, depending on many criteria, till it concludes to the maximum performance, and finally a detailed report is created. Keywords: DataWarp, Performance, I/O

15:00-16:30 Technical Session 28B

Tuning Sub-filing Performance on Parallel File Systems

(Byna, Chaarawi, Koziol, Choi)

Sub-filing is a technique used on parallel file systems to reduce locking and contention issues when multiple compute nodes interact with the same storage target node. Sub-filing provides a compromise between the single shared file approach that instigates the lock contention problems on parallel file systems and having one file per process, which results in generating a massive and unmanageable number of files. In this paper, we evaluate and

tune the performance of sub-filing feature implemented in HDF5. In specific, we will explain the implementation strategy of sub-filing feature in HDF5, provide examples of using the feature, and evaluate and tune parallel I/O performance of this feature with parallel file systems of the Cray XC40 system at NERSC (Cori) that include a burst buffer storage and a Lustre disk-based storage. Our initial results on Lustre show up to 4X performance advantage with sub-filing compared to writing a single file.

Enabling Portable I/O Analysis of Commercially Sensitive HPC Applications Through Workload Replication

(Dickson, Wright, Maheswaran, Herdman, Harris, Miller, Jarvis)

Benchmarking and analyzing I/O performance across high performance computing (HPC) platforms is necessary to identify performance bottlenecks and guide effective use of new and existing storage systems. Doing this with large production applications, which can often be commercially sensitive and lack portability, is not a straightforward task and the availability of a representative proxy for I/O workloads can help to provide a solution. We use Darshan I/O characterization and the MACSio proxy application to replicate five production workloads, showing how these can be used effectively to investigate I/O performance when migrating between HPC systems ranging from small local clusters to leadership scale machines. Preliminary results indicate that it is possible to generate datasets that match the target application with a good degree of accuracy. This enables a predictive performance analysis study of a representative workload to be conducted on five different systems. The results of this analysis are used to identify how workloads exhibit different I/O footprints on a file system and what effect file system configuration can have to performance.

An Exploration into Object Storage for Exascale Supercomputers

(Raja Chandrasekar, Evans, Wespetal)

The need for scalable, resilient, high performance storage is greater now than ever, in high performance computing. Exploratory research at Cray studies aspects of emerging storage hardware and software design for exascale-class supercomputers, analytics frameworks, and commodity clusters. Our outlook toward object storage and scalable database technologies is improving as trends, opportunities, and challenges of transitioning to them, also evolve. Cray's prototype SAROJA (Scalable And Resilient ObJect storAge) library is presented as one example of our exploration, highlighting design principles guided by the I/O semantics of HPC codes and the characteristics of up-and-coming storage media. SAROJA is extensible I/O middleware that has been designed ground-up with object semantics exposed via APIs to applications, while supporting a variety of pluggable file and object back-ends. It decouples the metadata and data paths, allowing for independent implementation, management, and scaling of each. Initial functional and performance evaluations indicate there is both promise and plenty of opportunity for advancement.

15:00-16:30 Technical Session 28C

Enabling the Super Facility with Software Defined Networking

(Canon, Draney, Lee, Paul, Skinner, Declerck)

Experimental and Observational facilities are increasingly turning to high-performance computing centers to meet their growing

analysis requirements. The combination of experimental facilities with HPC Centers has been termed the Super Facility. This vision requires a new level of connectivity and bandwidth between these remote instruments and the HPC systems. NERSC in collaboration with Cray has been exploring a new model of networking that builds on the principles of Software Defined Networking. We envision an architecture that allows the wide-area network to extend into the Cray system and enables external facilities to stream data directly to compute resources inside the system at over a 100 Gbs in the near-future and eventually reach beyond a 1 Tbs. In this paper will expand on this vision, describe some of the motivating use cases in more detail, layout our proposed architecture and implementation, describe our progress to date, and outline future plans.

Advanced Risk Mitigation of Software Vulnerabilities at Research Computing Centers

(Kaila)

Software security vulnerabilities have caused research computing centers concern, excess work, service breaks, and data leakages since the time of the great Morris Internet worm in 1988. Despite evolving awareness, testing and patching procedures, vulnerabilities and vulnerability patching still cause too much trouble both for users and for the sites. In this paper we will analyze actual operational risks and in published vulnerabilities, measure operational costs for deploying released security patches, and identify the genuine benefits of deploying the updates. As sources for our analysis we use vulnerability and incident metrics, results of penetration test and interviews with site representatives and other stakeholders. Our paper will results in recommendations for sites and providers on how to improve technology, procedures and best security practices, such as tuning rolling reboots, improving site specific risk identification, and implementing more fine grained access controls and intrusion detection measures.

Comparing Spark GraphX and Cray Graph Engine using large-scale client data

(Dull, Sacash)

Graph analytics are useful for overcoming real-world analytic challenges such as detecting cyber threats. Our Urika GX system is configured to use both the Cray Graph Engine (CGE) and Apache Spark for developing and executing hybrid workflows utilizing both Spark and Graph analytic engines. Spark allows us to quickly process data, stored in HDFS, powering flexible analytics in addition to graph analytics for cyber threat detection. Spark's GraphX library offers an alternative graph engine to CGE. In this presentation, we will compare the available algorithms, challenges, and performance of both CGE and GraphX engines in the context of a real-world client use case utilizing 40 billion RDF triples.

16:40-16:50 New Site Talk 29

New CUG member site talk from SSC

*(*Sulliva*)*

New CUG member site talk from SSC

16:50-17:10 General Session 30

CUG 2017 Conference Close

(Hancock)

+ Announcements + Special Thanks

Sessions and Abstracts

Note**

Technical paper "Preparing NERSC users for Cori, a Cray XC40 system with Intel Many Integrated Cores *(He, Cook, Deslippe, Friesen, Gerber, Hartman-Baker, Koniges, Kurth, Leak, Yang, Zhao, Baron, Hauschildt)* has not been assigned to a session as a courtesy. It is intended to take the slot emptied by the to-be-selected Best Paper when that paper is moved to the General Session 14 on Wed morning at 10:30 am.

Abstract: The newest NERSC supercomputer Cori is a Cray XC40 system consisting of 2,000+ Intel Xeon Haswell nodes and 9,600+ Intel Xeon-Phi "Knights Landing" (KNL) nodes. Compared to the Xeon-based clusters NERSC users are familiar with, optimal performance on Cori requires consideration of KNL mode settings; process, thread, and memory affinity; fine-grain parallelization; vectorization; and use of the high-bandwidth MCDRAM memory. This paper describes our efforts preparing NERSC users for KNL through the NERSC Exascale Science Application Program (NESAP), web documentation, and user training. We discuss how we configured the Cori system for usability and productivity, addressing programming concerns, batch system configurations, and default KNL cluster and memory modes. System usage data, job completion analysis, programming and running jobs issues, and a few successful user stories on KNL are presented.

Social Events

Sunday, 7th May 2017

18:00 – 19:30 Welcome Reception

All attendees and their guests are invited to a reception in the SEAR Fire-Inspired Restaurant at the Seattle Marriott Redmond, to renew existing acquaintances and establish new ones.

Monday, 8th May 2017

19:00 – 22:00 Farmers Market Special Event Sponsored by DDN Storage

All attendees and their guests are invited to the Redmond Saturday Market for a sampling of food trucks: Dante's Inferno Dogs, The Picnic and The Ultimate Melt. Enjoy Molly Moon's ice cream for dessert, brews from Mac & Jack's Brewery and a selection of local wines. Live music will be provided by Acoustic Transitions. The Farmers Market is adjacent to the Redmond Marriott Seattle.

Tuesday, 9th May 2017

19:00 – 22:00 Cray Networking Event at Living Computer Museum
Sponsored by CRAY
http://www.livingcomputers.org/

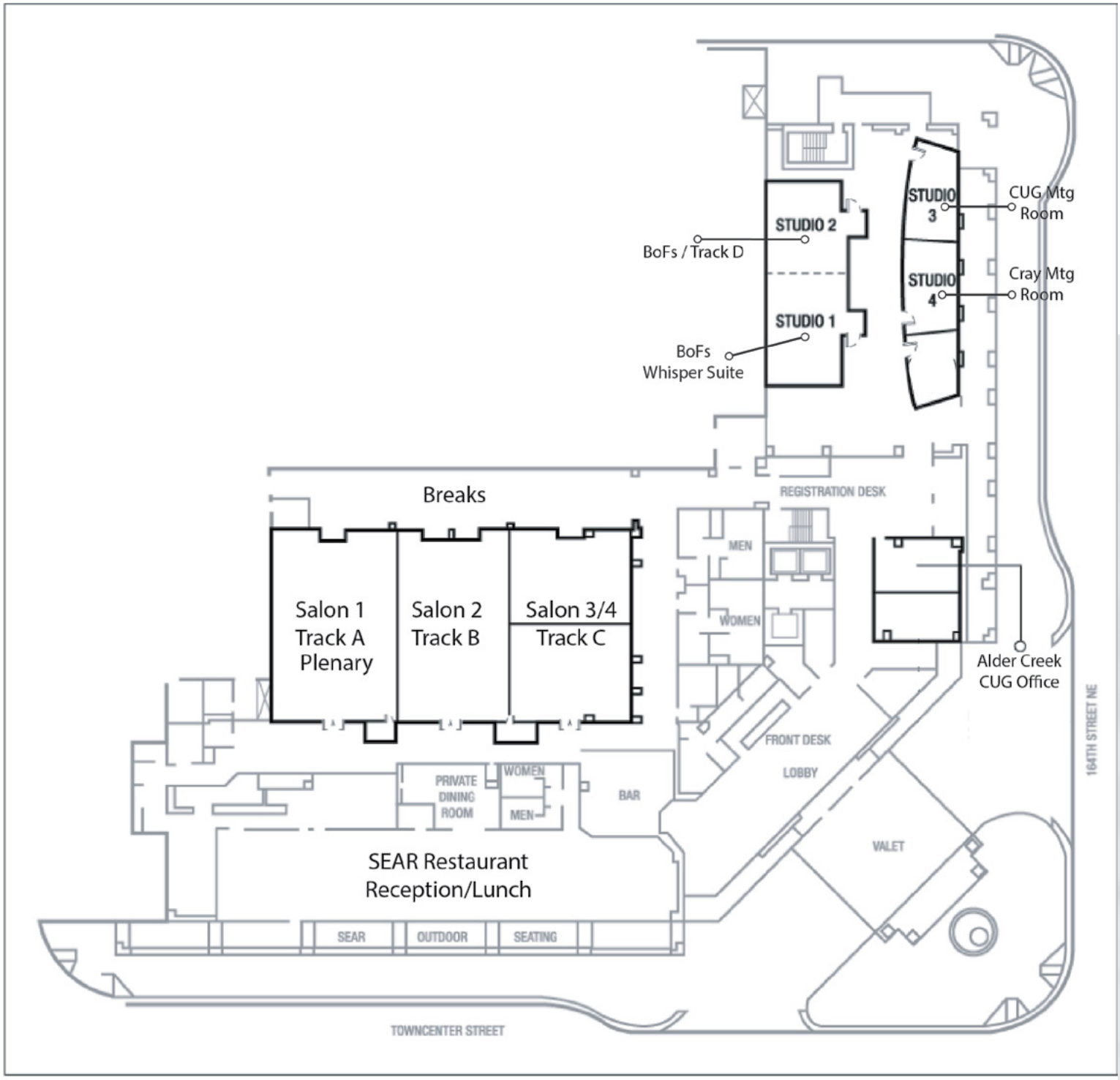
The Living Computer Museum provides a one-of-a-kind, hand-on experience with computer technology from the 1960's to the present. The museum honors the history of computing with the world's largest collection of fully restored – and usable – supercomputers, mainframes, minicomputers and microcomputers. A new main gallery offers direct experiences with robotics, virtual reality, artificial intelligence, self-driving cars, big data, the Internet of Things, video-game making, and digital art. Multiple buses have been scheduled to transport attendees between the Marriott and the Living Computer Museum. Pick up at the Marriott starting at 18:30 with last bus leaving at 19:00. For the return, buses will be available from 21:00. All attendees and their guests are invited.

Wednesday, 10th May 2017

19:00 – 22:00 CUG Night out at Chateau Ste. Michelle Sponsored by INTEL
https://www.ste-michelle.com/

All registered attendees and their paid guests are invited to Chateau Ste. Michelle, Washington State's oldest winery, for a gourmet dinner. Appetizers, entrées and dessert will be paired with the perfect Chateau Ste. Michelle wine. Transportation from The Marriott to the Chateau will be provided. Buses will load at The Marriott between at 18:15 and 18:30. The event starts at 18:45 at the museum, with a welcome drink held in the Barrel Room. From the Barrel Room, we will move through to the Banquet room where a 3-course meal will be served. Return transport will then be available from 21:30.

The Marriott Floor Plan



Local Arrangements

How to Contact Us

After the conference:

Oak Ridge National Laboratory
Attn: Jim Rogers
1 Bethel Valley Road P.O. Box 2008; MS 6008
Oak Ridge, TN 37831-6008
cug2017@cug.org

During the conference:

You can find us in The Alder Creek Boardroom (aka CUG Office) or at the registration desk

Conference Registration

Jim Rogers
Oak Ridge National Laboratory
1 Bethel Valley Road
Oak Ridge, TN 37831-6008
USA (1-865)-576-2978 Fax: (1-865)-241-9578
jrogers@ornl.gov

Attendance and Registration

Badges and registration materials will be available:

Sunday: 15:00 to 18:00 Registration desk
Monday: 07:30 to 17:30 Registration desk
Tuesday: 07:30 to 10:30 Registration desk

To register after Tuesday morning visit the CUG office (Alder Creek Boardroom).

All attendees must wear badges during CUG Conference activities.

Smoking Policy

There is no smoking allowed at the Conference.

Special Assistance

Any requests for special assistance during the conference should be noted on the "SpecialRequirements" area of the registration form

Conference Registration Fees

Your registration fee includes

- Admission to all program sessions, meetings, and tutorials
- Arrival, morning and afternoon breaks, and lunch Monday through Thursday
- All Social Events Sunday through to Wednesday

Proceedings

Proceedings details will be announced at the conference. Sites can use their member login or contact board@cug.org for general access

Contacts

CUG Board

President
David Hancock
Indiana University

Vice-President / Program Chair
Yun (Helen) He
National Energy Research Scientific Computing Center

Secretary
Tina Declerck
National Energy Research Scientific Computing Center

Treasurer
Jim Rogers
Oak Ridge National Laboratory

Director-at-Large
Richard Barrett
Sandia National Laboratory

Director-at-Large / Sponsor Liaison
William (Trey) Breckenridge III
Mississippi State University

Director-at-Large / Treasurer in Training
Scott Michael
Indiana University

Past President++
Nicholas Cardo
Swiss National Supercomputer Centre

Cray Advisor to the CUG Board **
Christy Adkinson
Cray Inc.

** Note: This is not a CUG Board position.
++ Note: Appointed Position
EMAIL board@cug.org for general CUG inquiries or cug2017@cug.orgfor specific inquiries.

Special Interest Groups (SIGs)

Programming Environments, Applications and Documentation

Chair: Biele Hadri (KAUST)
bilel.hadri@kaust.edu.sa

Deputy Chair: Ashley Barker (ORNL)
ashley@ornl.gov

Deputy Chair: Greg Bauer (NCSA)
gbauer@illinois.edu

Deputy Chair: Suzanne Parete-Koon (ORNL)
paretekoonst@ornl.gov

Deputy Chair: Zhengji Zhao (NERSC)
zzhao@lbl.gov

Cray Inc. SIG Liaison Applications: Jef Dawson
jef@cray.com

Programming Environments: Luiz DeRose
ldr@cray.com

Documentation: Kevin Stelljes stelljes@cray.com

Systems Support

Chair: Hans-Hermann Frese (ZIB)
frese@zib.de

Cray Inc. SIG Liaison Systems & Integration, Operating Systems, and Operations:
Kelly Marquardt kmarquardt@cray.com

XTreme Systems

Chair: Tina Butler (NERSC)
tbutler@nersc.gov

Deputy Chair: Frank Indiviglio (NCRC)
Frank.Indiviglio@noaa.gov
Cray Inc. SIG Liaison Kelly Marquardt
kmarquardt@cray.com

Sponsors

Diamond Sponsor

CRAY
THE SUPERCOMPUTER COMPANY

Platinum Sponsor

intel

Special Event Sponsor

DDN STORAGE

Welcome Reception Sponsor

SEAGATE

Gold Sponsors

ANSYS

PGI

Silver Sponsors

Adaptive COMPUTING

allinea

Altair

SchedMD

Bright Computing

CUG 2018

VISIONARY COMPUTING

CRAY USER GROUP 2018 I Stockholm, Sweden

The PDC Center for High Performance Computing at the KTH Royal Institute of Technology invites you to join us for CUG 2018 in Stockholm, Sweden from 20-24 May 2018.

"Visionary Computing" is the theme of CUG 2018. The many challenges in our field need visionary approaches to push our boundaries and our infrastructure enables visionary research. Sweden and the Stockholm region is the home of great and visionary minds such as Carl Linnaeus, Anders Celsius and Ander Jonas Ångström, to name a few. Stockholm is also the home town of Alfred Nobel, whose legacy of Nobel prizes gather the brightest minds every year in Stockholm. Leading IT businesses like Ericsson and Spotify make Stockholm a hotspot for the IT industry and with its focus on life sciences and engineering, which includes companies like Scania and Saab nearby, Stockholm is also a prime place for industrial HPC. In this exciting environment, CUG 2018 will explore new frontiers in HPC but also increasingly focus on data, analytics and the convergence of HPC and Big Data.

The PDC Center for High Performance Computing, which is hosting CUG 2018, is a prime Swedish HPC provider. PDC was established in 1990 and has a long tradition of providing the largest Swedish supercomputing resources for research. PDC's current flagship system is a 2 petaflops Cray XC40, Beskow, which is the largest academic system in the Nordic countries. As a national HPC centre, PDC supports a wide variety of Swedish and European researchers and also engages in industrial collaborations.

We look forward to welcoming you to Stockholm, where you can enjoy long days of Nordic sunlight, a vibrant city with a medieval centre and many cultural highlights, like Fotografiska (one of the world's largest contemporary photography museums), the Royal Castle, the Vasa Museum, and, for those who still remember them, the ABBA museum. Perhaps you would also like to take a ferry trip and explore the Stockholm archipelago which has over 30,000 islands.

Prof. Erwin Laure Director PDC – Center for High Performance Computing KTH

SEAGATE

Top 5 New Products or Technologies to Watch
HPC wire
Readers' Choice Awards 2016

Best HPC Storage Product or Technology
HPC wire
Readers' Choice Awards 2016

Performance, Speed and Efficiency Delivered

Seagate ClusterStor™ L-Series With Lustre Powers Cray® Sonexion®

Cray Sonexion 3000

- Reduce TCO by 25%
- Scale efficiently
- Performance-optimized End to End
- Achieve bandwidth of 112GB/s per rack. Fewer than 10 racks can perform at speeds in excess of 1TB/s
- GridRAID provides up to 400% faster rebuild times, meaning less management and downtime

CRAY SONEXION

For more information please contact:
Nick Ehrman, Business Development HPC: US East, Canada, & Latin America | (612) 618-2742
Patrick Ho, Business Development HPC: US West Coast | (408) 318-7775
Janet Duncombe, Business Development HPC: EMEA, India, & Pakistan | +44 (0) 7775 765688

www.seagate.com/enterprise-storage/systems/clustered-file-systems/#products | www.cray.com/products/storage/sonexion

Make HPC Productivity Easy

Request a Demo Today

- Job Submission and Management Portal
  Expand your user base to include even non-IT-skilled workers
- Reporting & Analytics Framework
  Gain insight into your system through reports and custom dashboards
- Remote Visualization
  Reduce file transfer time and save on expensive licenses and GPU's
- DataWarp Integration
  Speed time-to-solution for I/O-intensive jobs with tight integration

info@adaptivecomputing.com

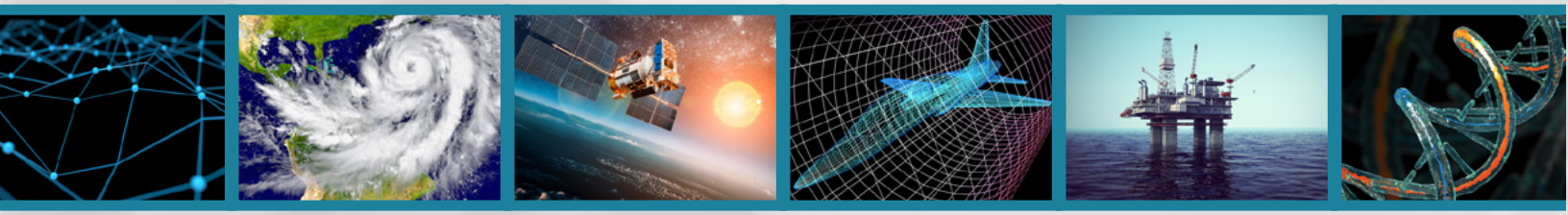
Adaptive COMPUTING

THE TOOLS BEHIND GROUNDBREAKING SOFTWARE

allinea
Now part of ARM

Debug, **profile** and **optimize** high performance applications

Allinea Software, now part of ARM, is the trusted provider of software development tools and code performance analytics for High Performance Computing (HPC). Currently, 80 percent of the world's top 25 supercomputers use Allinea tools, with key customers including the US Department of Energy, NASA, a range of supercomputing national labs and universities, and private companies using HPC systems for their own scientific computation. Leaders in HPC depend on the extremely scalable, capable and intuitive tools from Allinea for transforming the efficiency of their HPC investment, reducing development times and application run times.



www.allinea.com | T: 408 576 1141

Leave your competition in the dust with a solution from Cray and Bright

Achieve a dynamic data center

Bright connects compute-intensive and data-intensive workloads into a dynamic and integrated clustered computing environment

CRAY
THE SUPERCOMPUTER COMPANY

Bright Computing
brightcomputing.com