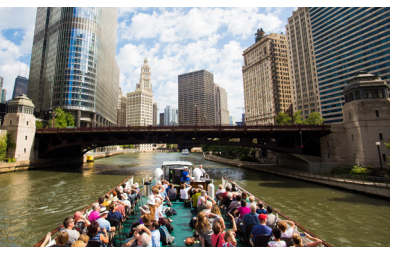
CUG 2015

CUG2015 Magnificent Computing

CHICAGO, ILLINOIS APRIL 26-30



Photos courtesy of Choose Chicago

Welcome To CUG 2015

Dear Friends and Colleagues,

The National Center for Supercomputing Applications at the University of Illinois invites you to CUG 2015 in Chicago. The meeting will provide Cray system users, developers, and administrators with an opportunity to discuss issues and innovations regarding large-scale, high-productivity computing, data analysis, and collaboration.

Established in 1986 at the University of Illinois as one of the original National Science Foundation supercomputing centers, NCSA has built an international reputation for accelerating computational science and engineering, deploying innovative machine architectures at extreme scale, and advancing the fields of data-intensive computing, cybersecurity, and visualization. NCSA has deployed and operates Blue Waters, one of the most powerful supercomputers in the world and the largest system Cray has ever built. Blue Waters, with a sustained performance of more than 1 PFLOPS, consists of 288 cabinets with 22,640 XE6 and 4,224 XK7 nodes. Additionally, its file system is the largest and fastest Lustre file system in the world. Similarly, the tape nearline storage system is the largest HPSS deployment worldwide. Blue Waters enables transformational research in a multitude of areas from astronomy to molecular biology to quantum physics and weather and climate modeling, to name only a few. Blue Waters also makes unparalleled computational resources available to the partnerships between academia and the private sector in fields such as health care, manufacturing, and independent software vendors.

We look forward to seeing you in Chicago, where we hope you will have the opportunity to visit the city's great parks and museums, sample its tremendous variety of cuisines and activities, and explore the rich and diverse cultural heritage of many of Chicago's neighborhoods, which provides the city with so much of its vibrant energy. We hope to make your visit to Chicago as rewarding as possible and would like to extend our invitation beyond the conference room to a series of social events that will highlight the best Chicago has to offer, from breathtaking skyscrapers, to beloved sport teams and soulful blues music.

Sincerely,

Dr. William Kramer,
Blue Waters Director

Dr. Cristina Beldica
Blue Waters Executive Director

Invited Speakers

Edward Seidel is the director of the National Center for Supercomputing Applications, a distinguished researcher in high-performance computing and relativity and astrophysics, and a Founder Professor in the University of Illinois Department of Physics and a professor in the Department of Astronomy. His previous leadership roles include serving as the senior vice president for research and innovation at the Skolkovo Institute of Science and Technology in Moscow, directing the Office of Cyberinfrastructure and serving as assistant director for Mathematical and Physical Sciences at the U.S. National Science Foundation, and leading the Center for Computation & Technology at Louisiana State University. His research has been recognized by a number of awards, including the 2006 IEEE Sidney Fernbach Award, the 2001 Gordon Bell Prize, and 1998 Heinz-Billing-Award.

Supercomputing in an Era of Big Data and Big Collaboration

Supercomputing has reached a level of maturity and capability where many areas of science and engineering are not only advancing rapidly due to computing power, they cannot progress without it. Detailed simulations of complex astrophysical phenomena, HIV, earthquake events, and industrial engineering processes are being done, leading to major scientific breakthroughs or new products that cannot be achieved any other way. These simulations typically require larger and larger teams, with more and more complex software environments to support them, as well as real world data. But as experiments and observation systems are now generating unprecedented amounts of data, which also must be analyzed via large-scale computation and compared with simulation, a new type of highly integrated environment must be developed where computing, experiment, and data services will need to be developed together. I will illustrate examples from NCSA's Blue Waters supercomputer, and from major data-intensive projects including the Large Synoptic Survey Telescope, and give thoughts on what will be needed going forward.

Invited Speakers

Paul Fischer is a Blue Waters Professor at the University of Illinois at Urbana-Champaign in the departments of Computer Science and Mechanical Science & Engineering. He received his Ph.D. in mechanical engineering from MIT and was a post-doc in applied mathematics at Caltech, where he was the first Center for Research in Parallel Computation fellow. His work is in the area of high-order numerical methods for partial differential equations and scalable linear solvers. He is the architect of the open-source fluid dynamics/heat transfer code Nek5000, which is based on the spectral element method. Nek5000 has scaled beyond a million ranks and has been awarded the Gordon Bell Prize in high-performance computing. It is used by more than 200 researchers for a variety of applications in turbulent and transitional flows.

Scalability Limits for Scientific Simulation

Current high-performance computing platforms feature millions of processing units, and it is anticipated that exascale architectures featuring billion-way concurrency will be in place in the early 2020s. The extreme levels of parallelism in these architectures influence many design choices in the development of next-generation algorithms and software for scientific simulation. This talk explores some of the challenges faced by the scientific computing community in the post-frequency-scaling era. To set the stage, we first describe our experiences in the development of scalable codes for computational fluid dynamics that have been deployed on over a million processors. We then explore fundamental computational complexity considerations that are technology drivers for the future of PDE-based simulation. We present performance data from leading-edge platforms over the past three decades and couple this with communication and work models to predict the performance of domain decomposition methods on model exascale architectures. We identify the key performance bottlenecks and expected performance limits at these scales and note a particular need for design considerations that will support strong scaling in the future.

Monday, 27th

8:30-10:00 Tutorial 1A

Next Generation Cray Management System for XC Systems

Lonelgy, Hesterberg, Navitsky

New major versions of CLE and SMW are being developed that include the next generation Cray Management System (CMS) for Cray XC systems. This next generation of CMS is bringing more common and easy to use system management tools and processes to the Cray XC systems, while at the same time preserving the system reliability and scalability upon which you depend. The next generation CMS includes a new common installation process for SMW and CLE, and more tightly integrates external Cray Development and Login (CDL) nodes as part of the Cray XC system. It includes the Image Management and Provisioning System (IMPS) that provides prescriptive image creation and centralized configuration. Finally, it integrates with the next major Linux distribution version from SUSE, SUSE Linux Enterprise Server 12. The tutoral will first cover an overview of the overall concepts of the next generation CMS, followed by examples of different system management activities.

08:30-10:00 Tutorial 1B

Cray XC Power Monitoring and Control

Martin, Kappel, Rush

This tutorial will focus on the setup, usage and use cases for Cray XC power monitoring and management features. The tutorial will cover power and energy monitoring and control from three perspectives: site and system administrators working from the SMW command line, users who run jobs on the system, and third party software development partners integrating with Cray's RUR and CAPMC features.

08:30-10:00 Tutorial 1C

Preparing for a smooth landing: Intel's Knights Landing and Modern Applications

Sewall

Knights Landing, the 2nd generation Intel® Xeon Phi™ processor, utilizes many breakthrough technologies to combine breakthrough's in power performance with standard, portable, and familiar programming models. This presentation will provide an overview of new technologies delivered by Knights Landing microarchitecture. Additionally, Dr. Sewall will provide studies of how applications have been developed using the first generation Intel® Xeon Phi™ coprocessor to be ready for Knights Landing.

13:00-14:30 Tutorial 2A

Next Generation Cray Management System for XC Systems

Lonelgy, Hesterberg, Navitsky

New major versions of CLE and SMW are being developed that include the next generation Cray Management System (CMS) for Cray XC systems. This next generation of CMS is bringing more common and easy to use system management tools and processes to the Cray XC systems, while at the same time preserving the system reliability and scalability upon which you depend. The next generation CMS includes a new common installation process for SMW and CLE, and more tightly integrates external Cray Development and Login (CDL) nodes as part of the Cray XC system. It includes the Image Management and Provisioning System (IMPS) that provides prescriptive image creation and centralized configuration. Finally, it integrates with the next major Linux distribution version from SUSE, SUSE Linux Enterprise Server 12. The tutoral will first cover an overview of the overall concepts of the next generation CMS, followed by examples of different system management activities.

Sessions and Abstracts

13:00-14:30 Tutorial 2B

Job-Level Tracking with XALT: A Tutorial for System Administrators and Data Analysts

Fahey, McLay, Budiardja

Let's talk real, no-kiddin' supercomputer analytics, aimed at moving beyond monitoring the machine as a whole or even its individual hardware components. We're interested in drilling down to the level of individual batch submissions, users, and binaries. And we're not just targeting performance: we're after ready answers to the "what, where, how, when and why" that stakeholders are clamoring for – everything from which libraries (or individual functions!) are in demand to preventing the problems that get in the way of successful research. This tutorial will show how to install and set up the XALT tool that can provide this type of job-level insight. The XALT tool can provide a wide range of metrics and measures of job-level activity. There are benefits to stakeholders beyond just end users: sponsoring institutions interested in strategic priorities and measurable impact; support organizations and development teams concerned about meeting users' needs and expectations; and those seeking to study user activity to improve value and effectiveness. We will show how to install and configure XALT for a variety of machines and usage modes. The proposers have experience installing and running their tool on a variety of machines in production from Crays to SGIs to clusters each with different compilers, job launchers, and batch systems. We will also show how this tool provides high value to centers and their users as it can provide documentation on how an application was built and when it was run well beyond what is usually been tracked by center.

13:00-14:30 Tutorial 2C

Debugging, Profiling and Tuning Applications on Cray CS and XC Systems

Paisley

The debugger Allinea DDT and profiler Allinea MAP are widely available to users of Cray systems - this tutorial, aimed at scientists and developers that are involved in writing or maintaining code, will introduce debugging and profiling using the tools. No prior experience with the tools is required. Attendees can use the tutorial to apply the tools to their own bugs or performance problems, or to educate their colleagues on the use of the tools. We will show how to get started on Cray systems - using the tool suite, Allinea Forge, which combines DDT and MAP in a single interface. We start by debugging simple MPI problems - exploring tactics for application crashes and unexpected behaviour, and move through topics such as debugging beyond-bound array accesses or memory leaks, into extreme scale debugging. We present tips that will aid when exposed to the most challenging environments. Moving on to performance, we show how to prepare applications for profiling and then how to interpret performance in Allinea MAP. The examples will show common performance issues and how to narrow down on the cause and remove bottlenecks - including I/O, MPI communication, processor extensions, memory performance and OpenMP. Key outcomes: being able to debug, profile and tune code running on Cray systems.

16:45-18:00 Interactive 3A

Systems Support

16:45-18:00 Interactive 3B

Programming Environments, Applications and Documentation

Tuesday, 28th

07:30-08:15 Interactive 4A

System Testing and Resiliency in HPC
Barker

As supercomputing system offerings from Cray become increasingly larger, more heterogeneous, and more tightly integrated with storage and data analytics, the verification of hardware and software components becomes an ever more difficult and important aspect of system management. Whether HPC resources are dedicated to local users or are shared with an international user community, regression testing is necessary to ensure that centers are providing usable and trustworthy resources for scientific discovery. However, unlike regression testing in software development projects, where there exists a range of well-established continuous integration tools, regression testing in HPC production environments is typically carried out in a more ad hoc fashion, using custom scripts or tools developed independently by individual HPC centers and with little or no collaboration between centers. The aim of this session is to bring together those with experience and interest in regression testing theory and practice with the aim of fostering collaboration and coordination across CUG member sites. We will assess the state-of-the-art in regression testing at member sites and determine the needs of the community moving forward. We will discuss the testing of components in terms of both functionality and performance, including best practices for operating system, driver and programming environment updates. The session will provide an open forum to share ideas and concerns in order to produce a more concerted effort towards the treatment of system testing and resilience across HPC centers. The session will be used to kick off a cross-site working group dedicated to sharing ideas and frameworks.

General Session 5

08:30-8:45

Welcome from the CUG President David Hancock

8:45-10:00

Supercomputing in an Era of Big Data and Big Collaboration
Seidel

Supercomputing has reached a level of maturity and capability where many areas of science and engineering are not only advancing rapidly due to computing power, they cannot progress without it. Detailed simulations of complex astrophysical phenomena, HIV, earthquake events, and industrial engineering processes are being done, leading to major scientific breakthroughs or new products that cannot be achieved any other way. These simulations typically require larger and larger teams, with more and more complex software environments to support them, as well as real world data. But as experiments and observation systems are now generating unprecedented amounts of data, which also must be analyzed via large-scale computation and compared with simulation, a new type of highly integrated environment must be developed where computing, experiment, and data services will need to be developed together. I will illustrate examples from NCSA's Blue Waters supercomputer, and from major data-intensive projects including the Large Synoptic Survey Telescope, and give thoughts on what will be needed going forward.

General Session 6

10:30-12:00

Cray Corporate Update
Ungaro

13:00-14:30 Technical Session 7A

Driving More Efficient Workload Management on Cray Systems with PBS Professional

Suchyta

The year 2014 brought an increase in adoption of key HPC technologies, from data analytics solutions to power-efficient scheduling. The HPC user landscape is changing, and it is now critical for workload management vendors to provide not only foundational scheduling functionality but also the adjacent capabilities that truly optimize system performance. In this presentation, Altair will provide a look at key advances in PBS Professional for improved performance on Cray systems. Topics include new Cray-specific features like Suspend/Resume, Xeon Phi support, HyperThreading, Power-aware Scheduling, and Exclusive/Non-exclusive ALPS reservations. The presentation will also preview the upcoming capabilities of cgroups and DataWarp integration.

13:00-14:30 Technical Session 7A

Innovations for The Cray

Beer, Brown

Moab and Torque have been efficiently managing the workload on Cray supercomputers for years. Aside from the policy-rich scheduling Moab provides, several new advancements have been made and are being developed specifically for Cray supercomputers. Recent releases include improvements to scheduling jobs at scale, additional power management controls, job-based energy accounting, and the ability to place jobs according to the topology of a 3D Torus network. Future releases will include NUMA-aware job task scheduling and placement, administrator portal, and full-integration with Data Warp technology. This session will discuss the challenges that have motivated these developments, the impacts that the released features have had in production, and the expected impacts of the features which are to be released.

13:00-14:30 Technical Session 7A

Slurm Road Map 15.08

Jenson

Slurm is an open source workload manager used on six of the world's top 10 most powerful computers and provides a rich set of features including topology aware optimized resource allocation, the ability to expand and shrink jobs on demand, failure management support for applications, hierarchical bank accounts with fair-share job prioritization, job profiling, and a multitude of plugins for easy customization. This presentation will provide a road map for Slurm 15.08. Topics will include: data warp (burst buffers), power management, improved job array scalability, support for different generic resource types, new API statistics from sdiag, database performance enhancements, expanded charging options and support for PMI Exascale (PMIx).

13:00-14:30 Technical Session 7B

Toward Understanding Life-Long Performance of a Sonexion File System

Swan, Petesch

Many of Cray's customers will be using their systems for several years to come. The one resource that is most affected by long-term use is storage. Files, both big and small, both striped and unstriped, are continually created and deleted, leaving behind free space of different sizes and in difference places on the spinning media. This paper will explore the effects of continual reuse of a Sonexion file system and a method of tuning the allocation parameters of the OSTs to minimize these effects.

13:00-14:30 Technical Session 7B

How Distributed Namespace Boosts Lustre Metadata Performance

Dilger

The Lustre Distributed Namespace Environment (DNE) feature allows Lustre metadata performance to scale upward with the addition of metadata servers to a single file system. Under development by Intel and others for several years, DNE functionality is a vital part of the latest production releases of Lustre software. During this technical session you'll learn how DNE works today, an update on continued improvements, and how DNE allows Lustre metadata performance to scale to meet the demands of applications having many thousands of threads.

13:00-14:30 Technical Session 7C

Porting the Urika-GD Graph Analytic Database to the XC30/40 Platform

Maschhoff, Vesse, Maltby

The Urika-GD appliance is a state of the art graph analytics database that provides high performance on complex SPARQL queries. This performance is due to a combination of custom multithreaded processors, a shared memory programming model and a unique network. We will present our work on porting the database and graph algorithms to the XC30/40 platform. Co-array C++ was used to provide a PGAS environment with a global address space, a Cray-developed soft threading library was used to provide additional concurrency for remote memory accesses and the Aries network provides RDMA and synchronization features. We describe the changes necessary to refactor these algorithms and data structures for the new platform. Having this type of analytics database available on a general-purpose HPC platform enables new use cases, and several will be discussed. Finally we will compare the performance of the new XC30/40 implementation to the Urika-GD appliance.

13:00-14:30 Technical Session 7C

A Graph Mining "App-Store" for Urika-GD

Sukumar, Lee, Brown, Hong, Roberts, Ainsworth, Lim

Researchers at Oak Ridge National Lab have created a suite of tools called EAGLE that will be made available for users of the Urika-GD installation. EAGLE is the acronym for "EAGLE 'Is A' Algorithmic Graph Library for Exploratory-Analysis" and includes an emulator environment for code development and testing, graph conversion and creation from heterogeneous data sources, interactive visualization along with implementation of traditional graph mining algorithms. We will present benchmark results of EAGLE on real world datasets across 5 seminal graph-theoretic algorithms (Degree distribution, PageRank, connected component analysis, node eccentricity, and triangle count). We compare EAGLE on Urika-GD with graph-mining on other architectures (e.g. distributed-memory GraphX, distributed-storage Pegasus) and programming models (Map-reduce, Pregel, SQL). We will conclude by demonstrating how EAGLE is serving as the building block of knowledge discovery using semantic reasoning and its application to biology, medicine and national security.

13:00-14:30 Technical Session 7C

Implementing a social-network analytics pipeline using Spark on Urika XA

Hinchey

We intend to discuss and demonstrate the use of new generation analytic techniques to find communities of users that discuss certain topics (consumer electronics, sports) and identify key users that play a role in or between those

communities (originators, rebroadcasters, connectors). The analytics execution is performed on a Cray Urika XA, a cluster of 48 nodes with 4T of RAM, 38T of SSD, and Lustre storage. The software framework used is Apache Spark and HDFS with the Java programming language. The Spark framework is similar to Hadoop/MapReduce, written in a functional style, allowing the engine to make efficient use of the full cluster, lazily, in parallel, with failure recovery, but without the user having to code for such complexity. The entire pipeline includes ETL, numerous aggregations and joins, and a graph algorithm. Spark-Streaming is used for complex event processing on real-time data, to identify patterns and changing trends.

15:00-17:00 Technical Session 8A

Cray DataWarp: Administration & SLURM integration

Declerck, Sakrejda

The National Energy Research Scientific Computing center (NERSC) is one of the Department of Energy's (DOE) primary centers for high performance computing needed for research. One of the areas that large compute centers have worked to find a solution is the ability to efficiently move data to and from compute nodes. Cray is addressing this with their Data Warp technology. As new technologies are being developed and used, new tools are needed to address the administration and troubleshooting. NERSC is collaborating with Cray to develop the capabilities needed by NERSC to provide functionality for our user base. In addition, integration into the workload manager will be needed to allow access for jobs. To address this, NERSC is collaborating with SchedMD to implement the key features needed to integrate Cray's Data Warp solution into SLURM. This paper will concentrate on the administrative interface and the integration with SLURM.

15:00-17:00 Technical Session 8A

A Converged Management Solution for HPC AND Big Data Analytics

Hunneyman

Bright Cluster Manager has been provisioning, monitoring and managing HPC clusters for over a decade. Last year, an add-on for Hadoop clusters became generally available. From bare-metal servers to the application stack, a single instance of the Bright GUI or command-line interface actually delivers a converged administration experience for HPC and Big Data Analytics. In practice, this means HPC admins can rapidly introduce Hadoop clusters alongside their existing HPC environments. And with Bright, HPC admins do not need to be experts on emerging technologies (e.g., Apache Hadoop including HDFS and YARN, Apache Spark) to deploy and maintain environments for pilot or production purposes. Summarized case studies, involving Cray systems, will be shared to illustrate how customers are making use of Bright to accelerate the introduction of Hadoop clusters for Big Data Analytics, alongside their existing HPC environments.

15:00-17:00 Technical Session 8A

Realtime process monitoring on the Cray XC30

Jacobsen, Canon, Srinivasan

Increasingly complex workflows of data-intensive calculations are extremely challenging to characterize. In preparation for the increasing prevalence of this new style of workload, we describe a recent effort to implement the "procmon" system on the Cray XC30 system. The procmon system was developed to characterize data-instensive workflows of the mid-range clusters at NERSC, enabling efficient whole system monitoring of all running processes on the system with live real-time analysis of the data.

procmon's resource-conscious implementation results in a scalable monitoring system that is minimally disruptive to both user and system processes, thereby providing useful monitoring opportunities on the large-scale Cray systems deployed at NERSC. Use of AMQP messaging enables flexible and fault-tolerant delivery of messages, while HDF5 storage of data allows efficient analysis using standardized tools. This approach results in an open monitoring system that provides users and operators detailed, realtime feedback about the state of the system.

15:00-17:00 Technical Session 8A

Cray System Snapshot Analyzer (SSA
Duckworth

The Cray System Snapshot Analyzer (SSA) represents a new support technology offering. SSA is a managed technology program designed to collect and analyze key customer system information. With SSA, we are targeting three areas of improvement. These areas are (1) reducing turn-around time for the collection of data in response to customer inquiries and issues (2) improving detection of and resolution time for customer system issues (3) improving Cray's knowledge of the product configurations in the field throughout their life-cycle. In this paper, we will first provide an overview of SSA. Next we will discuss anticipated benefits to Cray and, most importantly, to our customers. We will then discuss the architecture, including measures to ensure transparency in the operation of SSA and its security features. Finally, we will discuss the anticipated release and feature schedules for SSA.

15:00-17:00 Technical Session 8B

Data Transfer Study for HPSS Archiving
Wynne, Parete-Koon, Mitchell

The movement of large data produced by codes run in a High Performance Computing (HPC) environment can be a bottleneck for project workflows. To balance filesystem capacity and performance requirements, HPC centers enforce data management policies to purge old files to make room for new user project data. Users at Oak Ridge Leadership Computing Facility (OLCF) and other HPC user facilities must archive data to avoid the purge, therefore the time associated with data movement is something that all users must consider. This study observed the difference in transfer speed from the Lustre filesystem to the High Performance Storage System (HPSS) using a number of different transfer agents. The study tested files that spanned a variety of sizes and compositions that reflect OLCF user data. This will be used to help users of Titan plan their workflow and archival data transfers to increase their project's efficiency.

15:00-17:00 Technical Session 8B

A Storm (Lake) is Coming to Fast Fabrics: The Next-Generation Intel® Omni-Path Architecture
Davis

The Intel® Omni-Path Architecture, Intel's next-generation fabric product line, is designed around industry-leading technologies developed as a result of Intel's multi-year fabric development program. The Intel Omni-Path Architecture will deliver new levels of performance, resiliency, and scalability, overcoming InfiniBand limitations and paving the path to Exascale. Learn how the Intel Omni-Path Architecture will deliver significant enhancements and optimization for HPC at both the host and fabric levels, providing huge benefits to HPC applications over standard Infiniband-based designs.

15:00-17:00 Technical Session 8B

Applying Advanced IO Architectures to Improve Efficiency in Single and Multi-Cluster Environments

Vildibill

For 15 years DDN has been working with the majority of leading supercomputing facilities, pushing the limits of storage IO to improve the productivity of the world¹s largest systems. Storage technology advancement toward Excascale have not progressed as quickly as computing technology. The gap cannot be bridged by improving today's technologies - drive interface speeds are not increasing fast enough, parallel file systems need optimization to accomplish Exascale concurrency, and scientists will always want to model more data than is financially reasonable to hold in memory. Discontinuous innovation is called for. In this talk DDN will present the background collaboration, key technologies and community work - that has led to the development of an entirely new class of IO caching and new approaches to optimize file systems in order to divorce storage performance from storage capacity and allow for Exascale IO to transpire even on storage systems in use today.

15:00-17:00 Technical Session 8B

The Accelerated Road to Exascale

Southard

Why Moore's law started letting us down; what that means for accelerators and new ISAs; how GPU accelerators will maintain exponential gains in efficiency all the way to exascale.

15:00-17:00 Technical Session 8C

Use of Continuous Integration Tools for Application Performance Monitoring

Vergara Larrea, Joubert, Fuson

High performance computing systems are becoming increasingly complex, both in node architecture and in the multiple layers of software stack required to compile and run applications. As a consequence, the likelihood is increasing for application performance regressions to occur as a result of routine upgrades of system software components which interact in complex ways. The purpose of this study is to evaluate the effectiveness of continuous integration tools for application performance monitoring on HPC systems. In addition, this paper also describes a prototype system for application performance monitoring based on Jenkins, a Java-based continuous integration tool. The monitoring system described leverages several features in Jenkins to track application performance results over time. Preliminary results and lessons learned from monitoring applications on Cray systems at the Oak Ridge Leadership Computing Facility are presented.

15:00-17:00 Technical Session 8C

Parallel Software usage on UK National HPC Facilities 2009-2015: How well have applications kept up with increasingly parallel hardware?

Turner

One of the largest challenges facing the HPC user community on moving from terascale, through petascale, towards exascale HPC is the ability of parallel software to meet the scaling demands placed on it by modern HPC architectures. In this paper we analyse the usage of parallel software across two UK national HPC facilities: HECToR and ARCHER to understand how well applications have kept pace with hardware advances. These systems have spanned the rise of multicore architectures: from 2 to 24 cores per compute node. We analyse and comment on: trends in usage over time; trends in parallel programming models; trends in the calculation

size; and changes in research areas on the systems. The in-house Python tool that is used to collect and analyse the application usage statistics is also described. We conclude by using this analysis to look forward to how particular parallel applications may fare on future HPC systems.

15:00-17:00 Technical Session 8C

Sorting at Scale on BlueWaters in a Cosmological Simulation

Feng, Straka, Di Matteo, Croft

We implement and investigate a parallel sorting algorithm (MP-sort) on Blue Waters. MP-sort sorts distributed array items with non-unique integer keys into a new distributed array. The sorting algorithm belongs to the family of partition sorting algorithms: the target storage space of a parallel computing unit is represented by histogram bin whose edges are determined by partitioning the input keys, requiring exactly one global shuffling of the input data. The algorithm is used in a cosmology simulation (BlueTides) that utilizes 90\% of the computing nodes of Blue Waters, the Cray XE6 supercomputer at the National Center for Supercomputing Applications. MP-sort is optimal in communication: any array item is exchanged over the network at most once. We analyze a series of tests on Blue Waters with up to 160,000 MPI ranks. At scale, the single global shuffling of items takes up to 90\% of total sorting time, and overhead time added by other steps becomes negligible. MP-sort demonstrates expected performance on Blue Waters and served its purpose in the BlueTides simulation. We make the source code of MP-sort freely available to the public.

17:15-18:15 Interactive 9A

Systems monitoring of Cray systems

Showerman

This session is intended to present some of the challenges and solutions to monitoring Cray systems. The range of topics include data collection methods, application impact analysis for large scale systems, data storage strategies and visualization. While solutions for monitoring compute node statistics are beginning to mature, there remain many challenges in integrating data across subsystems that produce the insights necessary for effective administration. We will seek to gather current best practices as well as approaches to produce cross cutting data that incorporates job, system, and filesystem information to maximize the end to end performance of Cray systems. The session will include a few short presentations from sites and move to an open discussion format. The product will be a summary of the findings and a report made available to CUG sites.

17:15-18:15 Interactive 9B

Getting the Most Out of HPC User Groups

Parete-Koon

User groups can provide HPC user facilities with valuable feedback about current and future center resources, services, and policies. User groups serve as hub to regularly allow users and HPC facility staff to connect and identify user needs for training, software, and hardware. They also provide a forum where facility staff and vendors, such as Cray, can make users aware of beneficial new resources and services. User groups can be formally organized with a charter, elections, and regular meetings or informally organized by a simple mailing list. The main function is regular effective communication between users and between users and HPC Facility staff and vendors. This BoF will draw on the experiences from Oak Ridge Leadership Computing Facility, National Energy Research Scientific Computing Center, and National Center for Supercomputing Applications for organizing and developing an

effective and ongoing dialog with and between their users. We will discuss our best practices for organization and communication with short presentations from each of our centers and then invite the participants to share their experiences. The outcome of our discussion will be documented in an HPC User Group Best Practices.

17:15-18:15 Interactive 9C

Experiences with OpenACC

Poole, Foertter

The OpenACC API has been earning praise for leadership in directives programming models which accelerate code in a performance portable manner. This BOF will discuss the recent developer experiences with the latest OpenACC Compilers available from Cray and PGI. Several teams were brought together in just prior to CUG, which were composed of developers, compiler vendors, and other OpenACC supporters in a week long effort to make significant progress porting their code to use accelerators. Attendees of this BOF will get an opportunity to understand what obstacles were faced, how they were overcome, and what results could be achieved in short order with good support. Attendees will also come away with knowledge about the strengths and weaknesses of the approaches they took, of the current implementations, and what to expect in the future.

Wednesday, 29th

08:30-10:00 Technical Sessions 10A

Enabling Advanced Operational Analysis Through Multi-Subsystem Data Integration on Trinity

Brandt, DeBonis, Gentile, Lujan, Martin, Martinez, Olivier, Pedretti, Taerat, Velarde

Operations management of the ACES Trinity platform will rely on data from a variety of sources including System Environment Data Collections (SEDC); node level information, including high speed network (HSN) performance counters and high fidelity energy measurements; scheduler/resource manager; and plant environmental facilities. The water-cooled Cray XC platform requires a cohesive way to manage both the facility infrastructure and the platform due to several critical dependencies. We present preliminary results from analysis of integrated data on the Trinity test systems as it pertains to enabling advanced operational analysis through the understanding of operational behaviors, relationships, and outliers.

08:30-10:00 Technical Sessions 10C

An Investigation of Compiler Vectorization on Current and Next-generation Intel processors using Benchmarks and Sandia's Sierra Application

Rajan, Doerfler, Tupek, Hammond

Motivated by the need for effective vectorization in order to take full advantage of the dual AVX-512 vector units in Intel's Knights Landing (KNL) processor, to be used in the NNSA's Cray XC Trinity supercomputer, we carry out a systematic study of vectorization effectiveness using GNU, Intel and Cray compilers using current-generation Intel processors. The study analyzes micro-benchmarks, mini-applications and a set of kernel operations from Sandia's SIERRA mechanics application suite. Performance is measured with and without vectorization/optimizations and the effectiveness of the compiler generated performance improvement is measured. We also present an approach using C++ templates, data structure layout modifications and the direct use of Intel vector intrinsics to systematically improve vector performance of important Sandia Sierra application kernels, like eigenvalue/eigenvector

computations and nonlinear material model evaluations, for which the current generation of compilers cannot effectively auto-vectorize.

08:30-10:00 Technical Sessions 10A

Experience with GPUs on the Titan Supercomputer from a Reliability, Performance and Power Perspective

Tiwari, Gupta, Rogers

The performance efficiency and advantage of GPUs are well-studied, however the reliability aspects are relatively less well understood. In this paper, we provide a detailed understanding of GPU failures on the Titan supercomputer. We analyze how GPU failures affect the stability of the whole system. We will build a model and demonstrate when is it more beneficial to run on GPUs both in terms of performance and reliability. We point out several inconsistencies in the vendor error logging software and how it may affect the management of future GPU-enabled supercomputers. We also discuss the energy-efficiency benefits, scalability results and lessons learned while porting our codes to the Titan supercomputer. As we approach exascale, the resilience challenge will become even more critical due to increase in system-scale. Therefore, we believe that this large-scale field study on GPU error characterization, quantification, and impact would be useful to the community.

08:30-10:00 Technical Sessions 10A

Detecting and Managing GPU Failures

Cardo

GPUs have been found to have a variety of failure modes. The easiest to detect and correct is a clear hardware failure of the device. However, there are a number of not so obvious failures that can be more difficult to detect. With the objective to provide a stable and reliable GPU computing platform, it is imperative to identify issues with the GPUs and remove them from service. At the Swiss National Supercomputing Centre (CSCS), a significant amount of effort has been invested in the detection and isolation of suspect GPUs. Techniques have been developed to identify suspect GPUs and automated testing put into practice, resulting in a more stable and reliable GPU computing platform. This paper will discuss these GPU failures and the techniques used identify suspect nodes.

08:30-10:00 Technical Sessions 10B

Evaluation of Parallel I/O Performance and Energy Consumption with Frequency Scaling on Cray XC30

Byna, Austin

Large-scale simulations produce massive data that needs to be stored on parallel file systems. The simulations use parallel I/O to write data into file systems, such as Lustre. Since writing data to disks is often a synchronous operation, the application-level computing workload on CPU cores is minimal during I/O and hence it is considered energy may be saved by keeping the cores in lower power states. To examine this postulation, we have conducted a thorough evaluation of energy consumption and performance of various I/O kernels from real simulations on a Cray XC30 supercomputer, named Edison, at the National Energy Research Supercomputing Center (NERSC). To adjust CPU power consumption, we use the frequency scaling capabilities provided by the Cray power management and monitoring tools. In this paper, we present our initial observations that when the I/O load is high enough to saturate the capability of the filesystem, down-scaling the CPU frequency on compute nodes reduces energy consumption without diminishing I/O performance.

08:30-10:00 Technical Sessions 10B

A More Realistic Way of Stressing the End-to-end I/O System

Vergara Larrea, Oral, Leverman, Nam, Wang, Simmons

Synthetic I/O benchmarks and tests are insufficient by themselves in realistically stressing a complex end-to-end I/O path. Evaluations built solely around these benchmarks can help establish a high-level understanding of the system and save resources and time, however, they fail to identify subtle bugs and error conditions that can occur only when running at large-scale. The Oak Ridge Leadership Computing Facility recently started an effort to assess the I/O path more realistically and improve the evaluation methodology used for major and minor file system software upgrades. To this end, an I/O test harness was built using a combination of real-world scientific applications and synthetic benchmarks. The experience with the harness and the testing methodology introduced are presented in this paper. The more systematic testing performed with the harness resulted in a successful upgrade of Lustre on OLCF systems and a more stable computational and analysis environment.

08:30-10:00 Technical Sessions 10B

Tuning Parallel I/O on Blue Waters for Writing 10 Trillion Particles

Byna, Sisneros, Chadalavada, Koziol

Large-scale simulations running on hundreds of thousands of processors produce hundreds of terabytes of data that need to be written to files for analysis. One such application is VPIC code that simulates plasma behavior such as magnetic reconnection and turbulence in solar weather. The number of particles VPIC simulates is in the range of trillions and the size of data files to store is in the range of hundreds of terabytes. To test and optimize parallel I/O performance at this scale on Blue Waters, we used the I/O kernel extracted from a VPIC magnetic reconnection simulation. Blue Waters is a supercomputer at National Center for Supercomputing Applications (NCSA) that contains Cray XE6 and XK7 nodes with Lustre parallel file systems. In this paper, we will present optimizations used in tuning the VPIC-IO kernel to write a 5TB file with 5120 MPI processes and a 290TB file with ~300,000 MPI processes.

08:30-10:00 Technical Sessions 10C

The Cray Programming Environment: Current Status and Future Directions

DeRose

In order to achieve high performance on large-scale systems, application developers need a programming environment that can address and hide the issues of scale and complexity of high end HPC systems. In this talk I will present the recent activities 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. I will discuss some of the new functionality in the Cray compilers, tools, and libraries, such as support for GNU intrinsics, our C++11, and OpenMP plans, and will highlight the Cray's activities to help porting and hybridization of applications to support the emerging MIC architectures (Intel PHI), such as the scoping tool Reveal and the recently released Cray Comparative Debugger. Finally, I will discuss our roadmap for all areas of the Cray Programming Environment.

08:30-10:00 Technical Sessions 10C

Using Reveal to Automate Parallelization for Many-Core Systems
Poxon

Reveal, an application parallelization assistant, helps users add deeper levels of parallelism to an MPI program by analyzing loops, identifying issues with parallelization, and by automating tedious and error-prone tasks for the user. In preparation for Intel KNL many-core systems, Cray is extending Reveal with a new automatic parallelization mechanism that can be used in both "Build & Go" and "Tune & Go" user environments. With this functionality, the user follows a simple recipe to collect performance data that is typically done prior to application tuning. Instead of the user analyzing the data to determine where parallelism should be applied, Reveal and the Cray compiling environment analyze the data, and focus automated parallelization efforts on best-candidate loops. With a single step that requires no source code modifications, Reveal and the Cray compiling environment parallelize select loops and rebuild the program with the applied parallelism.

General Session 11

10:30-12:00

Taking HPC to New Heights
Hazra

Relentless focus on system performance continues to be the mantra for HPC, driving fundamental changes in memory, fabric, power efficiency and storage, and the need for new architectural frameworks for future HPC systems. Big data analytics coupled with HPC will enable accessing broad data sets for real-time simulation, further increasing demand for HPC and storage as well as Cloud based capabilities. Join Raj as he discusses significant trends in technology and how Intel is working with key partners to innovate in HPC system architecture.

General Session 12

13:00-13:45

1 on 100 or More
Ungaro

Open discussion with Cray President and CEO.

General Session 13

14:00-15:30

Scalability Limits for Scientific Simulation
Fischer

Current high-performance computing platforms feature millions of processing units, and it is anticipated that exascale architectures featuring billion-way concurrency will be in place in the early 2020s. The extreme levels of parallelism in these architectures influence many design choices in the development of next-generation algorithms and software for scientific simulation. This talk explores some of the challenges faced by the scientific computing community in the post-frequency-scaling era. To set the stage, we first describe our experiences in the development of scalable codes for computational fluid dynamics that have been deployed on over a million processors. We then explore fundamental computational complexity considerations that are technology drivers for the future of PDE-based simulation. We present performance data from leading-edge platforms over the past three decades and couple this with communication and work models to predict the performance of domain decomposition methods on model exascale architectures. We identify the key performance bottlenecks and expected performance limits at these scales and note a particular need for design considerations that will support strong scaling in the future.

15:45-17:15 Technical Session 14C

Contain This, Unleashing Docker for HPC

Canon, Pezzaglia, Jacobsen, Choila

Container-based computing is revolutionizing the way applications are developed and deployed and a new ecosystem has emerged around Docker to enable container based computing. However, this revolution has yet to reach the HPC community. In this paper, we will provide an overview of container computing and the potential value to the HPC community. We describe early work in using Docker to support scientific computing workloads. We will also discuss investigations in how Docker could be deployed in large-scale HPC systems.

15:45-17:15 Technical Session 14C

Using Maali to Efficiently Recompile Software Post-CLE Updates on the Cray XC Systems

Bording, Harris, Schibeci

One of the main operational challenges of High Performance Computing centers is the maintaining numerous scientific applications to support a large and diverse user community. At the Pawsey Supercomputing Centre we have developed "Maali", is a lightweight automated system for managing a diverse set of optimized scientific libraries and applications on our HPC resources. Maali is a set of BASH scripts that reads a template file that contains the all the information to necessary to download a specific version of an application or library, configure and compile it. This paper will present how we recently used Maali after the latest CLE update and the hardware changes of Magnus a Cray XC40 to recompile a large portion of our scientific software stack. Including what changes to Maali were needed for both the CLE and hardware updates to differentiate between Magnus and our Cray XC30 system Galaxy.

15:45-17:15 Technical Session 14A

LNet RAS Best Practices

Horn

Cray systems are engineered to withstand the loss of components, however, Lustre, historically, has not been as resilient in some cases. In this paper we discuss recent enhancements made to Lustre to improve resiliency and best practices for realizing Lustre RAS on Cray systems including how to tune timeouts and configure certain Lustre features for resiliency.

15:45-17:15 Technical Session 14A

Cray XC System Level Diagnosability Roadmap Update

Schutkoske

This paper highlights the current capabilities and the technical direction of Cray XC System level diagnosability. Cray has made a number of enhancements to existing diagnostics, commands, and utilities as well as providing new diagnostics, commands and utilities. This paper reviews the new capabilities that are available now, such as the Simple Event Correlator (SEC), Workload Test Suite (WTS), Node diagnostics and HSS diagnostic utilities. It will also look at what is planned for in the upcoming releases, including the initial integration of new technologies from the OpenStack projects.

15:45-17:15 Technical Session 14A

Cray XC System Node Level Diagnosability

Schutkoske

Cray XC System node level diagnosability is not just about diagnostics. Diagnostics are just one aspect of the tool chain that includes BIOS, user commands, power and thermal data and event logs. There are component level tests that are used to checkout the individual components, but quite often issues do not appear until full scale is

reached. From experience over the last few years, we have seen that no single tool or diagnostic can be used to identify problems, but rather multiple tools and multiple sources of data must be analyzed to provide proper identification, isolation, and notification of hardware and software problems. This paper provides detailed examples using the existing tool chain to diagnose node faults within the Cray XC system.

15:45-17:15 Technical Session 14B

The time is now. Unleash your CPU cores with Intel® SSDs

Kudryavtsev

When trying to solve humankind's most difficult and important challenges, time is critical. Whether it's mapping population flows to thwart the spread of Ebola, identifying in real-time potential terrorists or analyzing big data to find a promising cure for cancer, data scientists, government leaders, researchers, engineers, all of us can't wait. Yet, today most super computing platforms require users to do just this. CPUs remain idle while waiting for data to arrive for analysis or waiting for data to be written back. In this session, Bill Leszinske and Andrey Kudryavtsev will discuss advancements in Intel SSD technology that are unleashing the power of the CPU and Moore's Law. They'll dive into NVMe, a new standard specification interface for SSDs that can greatly benefit the HPC community, talk about the results early adopters are experiencing, and how adoption sets the foundation for consumption of disruptive NVM technology on the horizon.

15:45-17:15 Technical Session 14B

DataWarp: First Experiences

Andersson, Sachs, Tuma, Schuett

In this paper, we'll talk about our first experiences using the new Cray® XC™ DataWarp™ applications I/O accelerator technology on both I/O benchmarks and real world applications. The Cray® XC™ series DataWarp™ applications I/O accelerator technology is based on Flash SSD I/O blades being directly connected to the same Cray Aries interconnect as the compute nodes used by the user application. The DataWarp accelerator allocates storage dynamically in either private (dedicated) or shared modes. Storage performance quality of service can be provided to individual applications, based on the user's policies. This work compares the performance of standard I/O Benchmarks like IOR using the DataWarp file system against the same runs on Lustre. We also compare the performance of real applications like BQCD, a quantum chromodynamics program, and the Molpro and Turbomole quantum chemistry packages as well as the OpenFOAM CFD solver.

15:45-17:15 Technical Session 14C

PGI C++ with OpenACC

Leback, Colgrove, Wolfe, Trott

The last year, PGI has moved OpenACC for C++ from a place where it could only offload code and data structures that looked like C, to providing support for many C++-specific language features. Working closely with Sandia National Labs, we continue to push into new areas of the language. In this paper and talk, we will use examples to illustrate accelerating code using class member functions, inheritance, templates, containers, handling the implicit 'this' pointer, lambda functions, private data and deep copies. OpenACC 2.0 features such as unstructured data regions and the "routine" directive are highlighted, as well as a PGI feature to auto-detect and generate class member functions which are called from compute regions as ³routine seq². Results using the beta Unified Memory functionality in PGI 15.x, which can simplify data management, will also presented.

Finally, we'll discuss current limitations and the future directions of OpenACC with respect to C++.

17:30-18:15 Interactive Session 15A

Customer Support Modernisation
Kugel

Open discussion on service modernisation.

Wednesday, 29th

General Session 16

08:30-10:00
Panel Discussion - Does Big Data Imply Big Compute?
Cardo

Moderated panel discussion

08:30-10:00
New Member Lightning Talks
Cardo

Invited lightning talks by new CUG members featuring: Hong Kong Sanatorium & Hospital, Thomas Leung Argonne National Laboratory, Mark Fahey

10:30-12:00 Technical Session 17A

Resource Utilization Reporting Two Year Update
Barry

In the two years since CUG 2013 the Cray RUR feature has gone from powerpoint to the forth release of software, running on a variety of Cray systems. The most basic features of RUR have proven the most interesting to the widest spread of users: Cpu usage, memory usage, and energy usage are enduring concerns for site planning. Functionality added since the first release of RUR has largely focused on providing greater fidelity of measurement, and support for a full range of hardware. This paper briefly reviews the architecture of the RUR software, describes new functionality added since the initial implementation, and solicits user input on future designs. Also included are a sampling of statistics gathered from Cray datacenter machines contrasted with production machines at Cray customer sites.

10:30-12:00 Technical Session 17A

Cray Advanced Platform Monitoring and Control (CAPMC)
Martin, Kappel, Rush

With SMW 7.2.UP02 and CLE 5.2.UP02, Cray released its platform monitoring and management API called CAPMC (Cray Advanced Platform Monitoring and Control). This API is primarily directed toward workload manager vendors to enable power-aware scheduling and resource management on Cray XC-series systems and beyond. In this paper, we give an overview of CAPMC features, applets, and their driving use cases. We further describe the RESTful architecture of CAPMC, its security model, and discuss tradeoffs made in design and development. Finally, we preview future enhancements to CAPMC in support of in-band control and additional use cases.

10:30-12:00 Technical Session 17A

Overview of the KAUST's Cray X40 System – Shaheen II
Hadri, Kortas, Feki, Khurram, Newby

In November 2014, King Abdullah University of Science and Technology (KAUST) acquired a Cray XC40 supercomputer along with DataWarp technology, a Cray Sonexion 2000 storage system,

a Cray Tiered Adaptive Storage (TAS) system and a Cray Urika-GD graph analytics appliance. This new Cray XC40 system scheduled to be installed in March 2015, named Shaheen II, will deliver 25 times the sustained computing capability of KAUSTs current system. Shaheen II is composed of 6198 nodes representing a total of 198144 processors cores tightly integrated with a richly layered memory hierarchy and dragonfly interconnection network. Total storage space is of 17 PB with additional 1.5 PB dedicated to burst-buffer. An overview of the systems specifications, the challenges raised in term of power capping, Datawarp use, data migration as well as very early results on benchmarking will be presented and discussed.

10:30-12:00 Technical Session 17B

Illuminating and Electrifying OpenMP + MPI Performance

Paisley

The "one size fits all" MPI age has passed: ahead complex MPI and many-core OpenMP or MPI and GPUs. Increased core counts per CPU mean that performance will increasingly come from optimization within each node and this calls out for developer tools that point to the root causes of underwhelming performance or of bugs that prevent successful completion. With Allinea's debugging and profiling tools now on the majority of Cray systems and used regularly at extreme scale, we explore new capabilities for scientists and developers aiming to improve performance, scalability or correctness. We focus on the Allinea MAP profiler with its newly added OpenMP profiling support. We present our approach to the significant and important challenge of determining and exposing multi-threaded performance whilst maintaining low overhead measurement and extreme scalability.

We present examples of common performance patterns and key optimization steps from MPI and OpenMP through to vectorization and I/O.

10:30-12:00 Technical Session 17B

Performance and Extension of a Particle Transport Code using Hybrid MPI/Open-MP Programming Models

Pringle, Barrett, Turland, Weiland, Parsons

We describe AWE's HPC benchmark particle transport code, which employs a wavefront sweep algorithm. After almost 4 years collaboration between EPCC and AWE, we present Chimaera-2_3D: a Fortran90 and MPI/OpenMP code which scales well to thousands of cores for large problem sizes. Significant restructuring has increased the degrees of parallelism available to efficiently exploit future many-core exascale systems. For OpenMP, we have introduced slices through the cuboid mesh which present a set of cells which may be computed independently; and computation over the angles within each cell can also be parallelized using OpenMP. Previously, the initial form of Chimaera computed a coupled, inter-dependent iteration over 'Energy Groups'. Our new code now decouples these iterations which, whilst increasing the computational time, permits a new task level of efficient parallelism encoded using MPI. This paper will present results from the extensive benchmarking exercise using a Cray XT4/5 (HECToR) and a Cray XC30 (ARCHER).

10:30-12:00 Technical Session 17B

Optimizing Cray MPI and Cray SHMEM for Current and Next Generation Cray-XC Supercomputers

Kandalla, Knaak, Pagel

Modern compute architectures such as the Intel Many Integrated Core (MIC) and the NVIDIA GPUs are shaping the landscape of

supercomputing systems. Current generation interconnect technologies, such as the Cray Aries, are further fueling the design and development of extreme scale systems. Message Passing Interface (MPI) and SHMEM programming models are strongly entrenched in High Performance Computing. However, it is critical to carefully design and optimize communication libraries on emerging computing and networking architectures to facilitate the development of next generation science. In this talk, I will present the primary research and development thrust areas in Cray MPI and SHMEM software products targeting the current and next generation Cray XC series systems. Next, we will discuss some of the MPI-I/O enhancements and our experiences with optimizing I/O intensive applications on the Cray XC. Finally, we will discuss the design and development of MPI-4 Fault Tolerance capabilities for Cray XC systems.

10:30-12:00 Technical Session 17C

Application Performance on a Cray XC30 Evaluation System with Xeon Phi Coprocessors at HLRN-III

Wende, Noack, Schütt, Sachs, Steinke

We report experiences in using the Cray XC30 Test and Development System (TDS) at the HLRN-III site at ZIB for many-core computing on the Intel Xeon Phi coprocessors. The TDS comprises 16 compute nodes, each of which with one Intel Xeon Phi 5120D coprocessor installed. We present performance data for selected workloads including BQCD, VASP, GLAT, and Ising-Swendsen-Wang. For the GLAT application, we use the HAM-Offload framework (developed at ZIB) to offload computations to remote Xeon Phis using Heterogeneous Active Messages. By means of micro- benchmarks, we determined the characteristics of the different communication paths between the host(s) and the Xeon Phi(s) involving the Aries interconnect and the PCIe link(s), and compare the respective measurements against those taken on the InfiniBand cluster. Based on these results, we discuss their impact on the performance of the applications considered.

10:30-12:00 Technical Session 17C

Climate Science Performance, Data and Productivity on Titan

Mayer, da Silva

Climate Science models are flagship codes for the largest of HPC resources both in visibility, with the newly launched DOE ACME effort, and in terms of significant fractions of system usage. The performance of the DOE ACME model is captured with application level timers and examined through a sizeable run archive. Performance and variability of compute, MPI communication and disk I/O will be discussed. As Climate Science advances in their use of HPC resources there has been an increase in the required human and data systems to achieve our programs goals. A description of current workflow processes (hardware, software, human), planned automation of the workflow, along with historical and projected data in motion and at rest data usage will be detailed. The combination of these two topics will lead to description of future systems requirements for DOE Climate Modeling efforts.

10:30-12:00 Technical Session 17C

Memory Scalability and Efficiency Analysis of Parallel Codes

Janjusic, Kartsaklis

Memory scalability is an enduring problem and bottleneck that plagues many parallel codes. Parallel codes designed for High Performance Systems are typically designed over the span of several, and in some instances 10+, years. As a result, optimization practices which were appropriate for earlier systems may no longer

be valid and thus require careful optimization consideration. Specifically, parallel codes whose memory footprint is a function of their scalability must be carefully considered for future exa-scale systems. In this paper we present a methodology and tool to study the memory scalability of parallel codes. Using our methodology we evaluate an application's memory footprint as a function of scalability, which we coined memory efficiency, and describe our results. In particular, using our in-house tools we can pinpoint the specific application components which contribute to the application's overall memory foot-print (application data-structures, libraries, etc.).

13:00-14:30 Technical Session 18A

Custom Product Integration and the Cray Programming Environment

Byland, Ward

With Cray's increasing customer base and product portfolio a faster, more scalable, and flexible software access solution for the Cray Programming Environment became required. The xt-asyncpe product-offering required manual updates to add new product and platform support, took a significant amount of time to evaluate the environment when building applications, and didn't harness useful standards used by the Linux community. CrayPE 2.x, by incorporating the flexibility of modules, the power of pkg-config and a programmatic design, offers a stronger solution going forward with simplified extensibility, a more robust solution for adding products to a system, and a significant reduction in application build time for users. This paper discusses the issues addressed and the improved functionality available to support Cray, customers and 3rd-party software access.

13:00-14:30 Technical Session 18A

Cray Storm Programming

Race

The Cray Cluster Storm is a dense, but highly power efficient computing platform for both current and next generation scientific applications. This product combine the latest Intel processors (Haswell), eight NVIDIA K40s or K80s and single/dual Mellanox IB connections into a hardware package that delivers performance to applications. The ability to access this computing capability relies on the different programming options available to the users and their applications. At the end of this presentation, the user will have a basic understanding of the programming options available on the storm and some basic performance information of some of these options. The basic programming options will include - Compilers, OpenACC, MPI and MPI+X.

13:00-14:30 Technical Session 18A

HPC Workforce Preparation

Lathrop

Achieving the full potential of today's HPC systems, with all of their advanced technology components, requires well-educated and knowledgeable computational scientists and engineers. Blue Waters is committed to working closely with the community to train and educate current and future generations of scientists and engineers to enable them to make effective use of the extraordinary capabilities provided by Blue Waters and other petascale computing systems. This session will provide a presentation on efforts to address the preparation of the HPC workforce including: Providing training webinars, workshops and summer schools, Providing web-based graduate credit courses, Graduate fellowships and internships focused on extreme scale computing,

Strategies for engaging women, minorities and people with disabilities, and Providing a repository of quality reviewed training and education materials. The session will include a discussion among the participants to foster sharing of information and provide groundwork for collaborations among the participants.

13:00-14:30 Technical Session 18B

Utilizing Unused Resources To Improve Checkpoint Performance

Miller, Atchley

Titan, the Cray XK7 at Oak Ridge National Laboratory, has 18,688 compute nodes. Each node consists of a 16-core AMD CPU, an NVIDIA GPU and 32GB ram. In addition, there is another 6GB of ram on each GPU card. Not all the applications that run on Titan make use of all a node's resources. For applications that are not otherwise using the GPU, this paper discusses a technique for using the GPU's ram as a large write-back cache to improve the application's file write performance.

13:00-14:30 Technical Session 18B

Lustre Metadata DNE Performance on Seagate Lustre System

Fragalla

Alongside the high demands of streaming bandwidth in High Performance Computing (HPC) storage, there is a growing need for increased metadata performance associated with various applications and workloads. The Lustre parallel filesystem provides a distributed namespace feature, which divided across multiple metadata servers, allows the metadata throughput to scale with increasing numbers of servers. This presentation explains how Seagate's solution addresses DNE Phase 1 in terms of performance, scalability, and high availability, including details on the DNE configuration and MDTEST performance benchmark results.

13:00-14:30 Technical Session 18B

Sonexion - SW versions/roadmap

Friesen

New Sonexion software releases for the Sonexion product line will include significant improvements including changes to Reliability, Availability, Serviceability as well as support for Lustre 2.5. The paper will explain the incremental changes, the planned timeline, and the targeted products (i.e. 900, 1600, 2000) for each software release.

13:00-14:30 Technical Session 18C

Large-Scale Modeling of Epileptic Seizures: Scaling Properties of Two Parallel Neuronal Network Simulation Algorithms

Pesce, Wildeman, Suresh, Eissa, Eijkhout, Hereld, Van Dongelen

Our limited understanding of the relationship between the behavior of individual neurons and large neuronal networks is an important limitation in current epilepsy research and may be one of the main causes of our inadequate ability to treat it. Addressing this problem directly via experiments is impossibly complex, thus, we have been developing and studying medium-large scale simulations of detailed neuronal networks to guide us. Flexibility in the connection schemas and a complete description of the cortical tissue seem necessary for this purpose. In this paper we examine some of the basic issues encountered in these multi-scale simulations. The observed memory and computation-time scaling behavior for a distributed memory implementation was very good over the range studied, both in terms of network sizes and processor pool sizes. We believe that these simulations proved that modeling of epileptic seizures on networks with millions of cells should be feasible on Cray supercomputers.

13:00-14:30 Technical Session 18C

The Impact of High-Performance Computing Best Practice Applied to Next-Generation Sequencing Workflows

Sosa, Carrier, Long, Walsh, Haas, Tickle, William, Dawson, Long

High Performance Computing (HPC) Technology and Best Practice (performance analysis and optimization) are enabling scientists in many disciplines to achieve progressively more demanding and valuable results. In this talk we will illustrate how the same technology and best practice can be used to dramatically accelerate next-generation sequencing (NGS) workflows. We illustrate how the XC family of systems is well suited for NGS workflows.

13:00-14:30 Technical Session 18C

Parallelization of whole genome analysis on a Cray XE6

Puckelwartz, Pesce, McNally, Foster

The declining cost of generating DNA sequence is promoting an increase in whole genome sequencing, especially as applied to the human genome. Whole genome analysis requires the alignment and comparison of raw sequence data. Given that the human genome is made of approximately 3 billion base pairs, each of which can be sequenced 30 to 50 times, this generates large amounts of data that have to be processed by complex, computationally expensive, and quickly evolving workflows. On the University of Chicago Cray XE6, Beagle, we implemented a scalable concurrent multiple genome analysis that also increased usable sequence per genome. Relying on publicly available software, the Cray XE6 has the capacity to align and call variants on hundreds of whole genomes in 50 h. The workflow displayed very good scalability and utilization of the computational resources when applied to 80 whole genomes. Multisample variant calling is also accelerated.

15:00-16:30 Technical Session 19A

Implementing "Pliris-C/R" Resiliency Features Into the EIGER Application

Davis, Kotulski, Tucker

EIGER is a frequency-domain electromagnetics simulation code based on the boundary element method. This results in a linear equation whose matrix is complex valued and dense. To solve this equation the Pliris direct solver package from the Trilinos library is used to factor and solve this matrix. This code has been used on the Cielo XE6 platform to solve matrix equations of order 2 million requiring 5000 nodes for 24 hours. This paper describes recent work to implement "Pliris-C/R", a set of checkpoint/restart and other resilience features for Pliris. These include: targeting multiple file systems in parallel; striping controls; checkpoint period controls; turnstiling; open-file-descriptor sharing across processes; checkpointing on imminent job termination; application relaunch within the job; and scripts to monitor application progress. Timing data for runs using Pliris-C/R will also be presented.

15:00-16:30 Technical Session 19A

Monitoring and Analyzing Job Performance Using Resource Utilization Reporting (RUR) on A Cray XE6 System

Su, Baer, Rogers, McNally, Whitten, Crosby

This paper describes the collection and analysis of job performance metrics using the Cray Resource Utilization Reporting (RUR) software on Mars, a Cray XE6 system at the National Institute for Computational Sciences (NICS). Cray offers users a new feature RUR in the second half of 2013. We can collect an easily expanded set of utilization data about each user[1]s applications with RUR. The overhead and scalability of

RUR will be measured using an assortment of benchmarks that covers a wide range of typical cases in realistic user environment, including computational-bound, memory-bound, and communication-bound applications. A number of the Cray-supplied data and output RUR plugins will be investigated. Possible integration with the XSEDE Metrics on Demand (XDMoD) projects will also be discussed.

15:00-16:30 Technical Session 19A

Molecular Modelling and the Cray XC30 Power Management Counters

Bareford

This paper explores the usefulness of the data provided by the power management (PM) hardware counters available on the Cray XC30 platform. PM data are collected for two molecular modelling codes, DL POLY and CP2K, both of which are run over multiple compute nodes. The first application is built for three programming environments (Cray, Intel and gnu): hence, the data collected is used to test the hypothesis that the choice of compiler should not impact energy use significantly. The second code, CP2K, is run in a mixed OpenMP/MPI mode, allowing us to explore the relationship between energy usage and thread count. The Cray-compiled DL POLY code had the lowest energy usage on average, 3-4% lower than the Intel and gnu results. In general, energy usage follows execution time. For the CP2K code, an energy-usage sweet spot of three threads per MPI process was revealed; for higher thread counts, execution times increase monotonically with energy usage.

15:00-16:30 Technical Session 19B

Staying Out of the Wind Tunnel with Virtual Aerodynamics

Clifford, Suchyta

In this presentation, Altair will present results from recent benchmark testing for both small and large simulations using HyperWorks Virtual Wind Tunnel on Cray XC30 systems. The tests focused on two problems of different sizes: a relatively small (22 million element) analysis of a benchmark car model used frequently in auto manufacturing, and a large (1 billion finite element cells) problem involving the drafting simulation of two Formula-1 cars. Result highlights included: virtually ideal efficiency when scaling to 300 cores; and for the larger problem, excellent performance up to 1600 cores with very good performance at 3000+ cores.

15:00-16:30 Technical Session 19B

Experiences Running and Optimizing the Berkeley Data Analytics Stack on Cray Platforms

Maschhoff, Ringenburg

The Berkeley Data Analytics Stack (BDAS) is an emerging framework for big data analytics. It consists of the Spark analytics framework, the Tachyon in-memory filesystem, and the Mesos cluster manager. Spark was designed as an in-memory replacement for Hadoop that can in some cases improve performance by up to 100X. In this paper, we describe our experiences running BDAS on the new Cray Urika-XA extreme analytics platform, on Cray XC systems, and on a prototype Aries-based system with node-local SSDs. We discuss how we configured and optimized the BDAS stack, and describe the execution environment used on each platform. BDAS applications differ significantly from traditional HPC applications: they run in the Java Virtual Machine, and communicate via TCP/IP.

We explore how Cray system capabilities, such as the Aries interconnect and SSDs, can be better leveraged to improve performance of these types of applications.

15:00-16:30 Technical Session 19B

Cyber-threat analytics using graph techniques

Dull

Computer network analysis can be very challenging due to the volumes and varieties of data. Organizations struggle with analyzing their network data, merging it against contextual information, and using that information. Graph analysis is an analytic approach that overcomes these challenges. Urika-GD powered graph analytics have been demonstrated at SC14 while Cray participated in the Network Security team on SCinet, the SC14 conference network. Cray participates on the network security team because of the scale of data (18 billion triples from 5 days of data), time-to-first solution (analytics need to be developed in minutes to an hour or two), and time-to-solution (answers need to be generated in seconds to minutes to be useful) requirements. This talk describe computer network information, computer network analysis problems, graph algorithm applications to these problems, and successes using Urika-GD to perform graph analytics during SC14.

15:00-16:30 Technical Session 19C

The Impact of Failures on the Workload of Modern Supercomputers

Meneses, Ni, Jones, Maxwell

The unprecedented computational power of current supercomputers has made possible the exploration of complex problems in many scientific fields, from genomic analysis to computational fluid dynamics. Modern machines are powerful because they are massive: they assemble millions of cores and a huge quantity of disks, cards, routers, and other components. But, it is precisely the size of these machines what glooms the future of supercomputing. A system that comprises many components has a high chance to fail, and fail often. Therefore, to make the next generation of supercomputers usable, it is imperative to use some type of fault tolerance platform to run applications on large machines. Most of fault tolerance strategies can be optimized for the peculiarities of each system and boost its efficacy in keeping the system productive. In this paper, we aim to understand how failure characterization can improve resilience in several layers of the software stack: applications, runtime systems, and job schedulers. We examine Titan supercomputer, one of the fastest systems in the world. We analyze a full year of Titan in production and distill the failure patterns of the machine. By looking into Titan's log files and using the criteria of experts, we provide a detailed description of the types of failures. In addition, we inspect the job submission files and describe how the system is used. Using those two sources, we cross correlate failures in the machine to executing jobs and provide a picture of how failures affect the user experience. We believe such characterization is fundamental in developing appropriate fault tolerance solutions for Cray systems similar to Titan. We also investigate how failures impact long-running jobs. We provide a series of recommendations for developing resilient software on supercomputers.

15:00-16:30 Technical Session 19C

Preparation of codes for Trinity

Vaughan, Rajan, Dinge, Dohrmann, Glass, Franko, Pierson, Tupek

Sandia and Los Alamos National Laboratories are acquiring Trinity, a Cray XC40, with half of the nodes having Haswell processors and the other half having Knight's Landing processors.

As part of our Center of Excellence with Cray, we are working on porting three codes, a Solid Mechanics code, a Solid Dynamics code, and an Aero code, to effectively use this machine. In this paper, we will detail the work that we have done in porting the codes in preparation of getting the machine. We have started by profiling the codes using tools including CrayPat, which showed that a large portion of the time is being spent in the solvers. We will describe the work we are doing on the solvers such as ongoing work on Haswell processors and Knight's Corner machines.

15:00-16:30 Technical Session 19C

Reducing Cluster Compatibility Mode (CCM) Complexity

Kohnke, Barry

Cluster Compatibility Mode (CCM) provides a suitable environment for running out of the box ISV and third party MPI applications, serial workloads, X11, and doing compilation on Cray XE/XC compute nodes. At times, customers have experienced CCM issues related to setting up or tearing down that environment. The tight coupling of CCM to workload manager prologue and epilogue services has been a primary source of issues. A new configurable ALPS prologue and epilogue service specific to CCM will be provided. Removing this workload manager dependency will reduce the CCM complexity. Other problem areas have been identified, and solutions will be implemented to avoid or correct those issues. This paper will describe problems and the changes made to CCM to reduce the CCM complexity and provide a more robust, workload manager independent product.

General Session 20

16:45

Closing Session - CUG 2016 Preview

Monday, 27th

18:30-21:30 Blues & Brews Special Event Sponsored by DDN

Chair: Jim Rogers (Oak Ridge National Laboratory)

Head two blocks west on Grand Ave to Rock Bottom Restaurant & Brewery (1 West Grand Avenue, Chicago, IL) for a delightful evening of Blues & Brews sponsored by DDN Storage. The event will be on the 2nd floor of the restaurant, includes a buffet dinner, and will feature Chicago native blues performer, artist, playwright, and educator Fernando Jones. http://www.fernandojon.com

Tuesday, 28th

18:30-21:30 Cray Networking Event at Spiaggia

Chair: Christy Adkinson (Cray, Inc.)

Attendees should head north on Michigan Avenue into the Gold Coast Area to Spiaggia at 980 North Michigan Avenue, Chicago, Illinois 60611. It's 0.7 mi from the Conrad, approximately a 15 minute walk. Take the escalator to the 2nd floor and the elevator from there to Spiaggia private event space. A coat and bag check will be available. Spiaggia is the only four-star Italian restaurant in Chicago. Spiaggia has earned international praise and numerous awards, including a 2014 nomination for Outstanding Restaurant by the James Beard Foundation. The private dining room offers floor to ceiling panoramic views of Lake Michigan and the Magnificent Mile. http://www.spiaggiarestaurant.com/

Wednesday, 29th

18:30-22:00 CUG Night Out at the Museum

Chair: Jim Rogers (Oak Ridge National Laboratory)

CUG Night Out at the Museum of Science and Industry. Transportation from the Conrad to the museum and back will be provided. Buses will load at the Conrad at 6:15 p.m. The event starts at 7:00 PM at the museum. Coat and bag check will be available at the museum. A variety of delectable food prepared by renowned Chicago chefs will be served in the beautiful rotunda of the museum. Three fantastic museum displays will be open for CUG guests: * Science Storms: Feel the physics and consider the chemistry of natural phenomena like tornados and avalanches. * Transportation Gallery/The Great Train Story: In the Transportation Gallery, you can explore climb aboard the engine of the Empire State Express 999, the first machine to break the 100 mph barrier; walk through a real United 727, hanging from the Museum's balcony; or take in the Spirit of America, the car that drove more than 530 miles per hour in 1965. The Great Train Story model railroad allows you to witness more than 20 trains running on 1,400 feet of track, completing the winding journey between Chicago and Seattle. * Numbers in Nature: A Mirror Maze: From the delicate nested spirals of a sunflower's seeds, to the ridges of a majestic mountain range, to the layout of the universe, mathematical patterns abound in the natural world. Numbers in Nature: A Mirror Maze exposes and explains the patterns that surround us. http://www.msichicago.org/

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
cug2015@cug.org

During the conference:

You can find us at The Boardroom (aka CUG Office) on the 8th floor or at the registration desk on the 6th floor (Magnolia pre-function space on the map) on Sunday, Monday, Tue AM.

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: 3:00 p.m. to 6:00 p.m.
Registration desk - 6th floor
Monday: 7:30 a.m. to 5:30 p.m.
Registration desk - 6th floor
Tuesday: 7:30 a.m. to 10:30 a.m.
Registration desk - 6th floor

To register after Tuesday morning visit the CUG office on the 8th floor.

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 "Special Requirements" area of the registration form

Conference Registration Fees

Your registration fee includes

- Admission to all program sessions, meetings, and tutorials
- Morning and afternoon breaks, and lunch Monday through Thursday
- CUG Night Out on Tuesday night
- Continental breakfast at the hotel

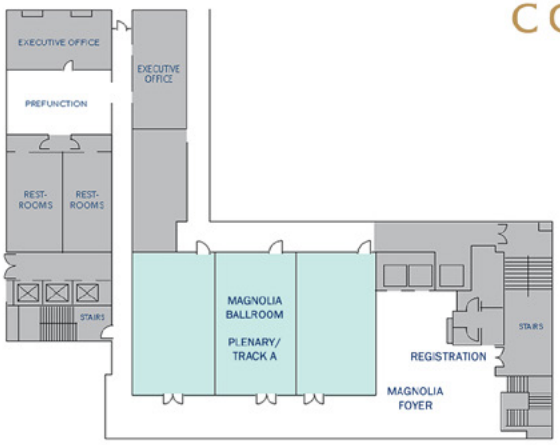
Proceedings

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

Conrad Floor Plan

6TH FLOOR

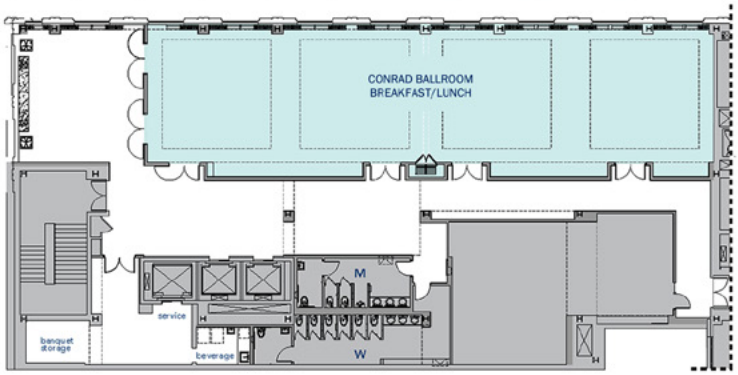
CONRAD
CHICAGO



8TH FLOOR

10TH FLOOR

11TH FLOOR



Sponsors

2015 Host

NCSA

Diamond Sponsor

CRAY
THE SUPERCOMPUTER COMPANY

Platinum Sponsor

intel

Special Event Sponsor

DataDirect NETWORKS

Gold Sponsors

Altair

SEAGATE

NVIDIA

PGI

Silver Sponsors

allinea

SPECTRA
DEEP STORAGE EXPERTS

SchedMD

Bronze Sponsors

Adaptive COMPUTING

Bright Computing

Contacts

CUG Board

President
David Hancock
Indiana University

Vice-President
open

Secretary
Robert Ballance
Sandia National Laboratories

Treasurer
Jim Rogers
Oak Ridge National Laboratory

Director-at-Large
Tina Declerck
National Energy Research Scientific Computing Center

Director-at-Large
Jeremy Enos
National Center for Supercomputing Applications

Director-at-Large
Liam Forbes
Arctic Region Supercomputing Center

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 cug2015@cug.orgfor specific inquiries.

Special Interest Groups

Programming Environments, Applications and Documentation

Chair: Tim Robinson (CSCS)

Deputy Chair: Ashley Barker (ORNL)

Deputy Chair: Helen He (NERSC)

Deputy Chair: Rolf Rabenseifner (HLRS)

Deputy Chair: Greg Bauer (NCSA)

Deputy Chair: Suzanne Parete-Koon (ORNL)

Deputy Chair: Zhengji Zhao (NERSC)

Systems Support

Chair: Jason Hill (ORNL)

Deputy Chair: Hans-Hermann Frese

Deputy Chair: Sharif Islam

Cray Inc. SIG Liaison Systems & Integration, Operating Systems, and Operations: Kelly Marquardt

XTreme Systems

Chair: Tina Butler (NERSC)

Deputy Chair: Frank Indiviglio (NCRC)

Cray Inc. SIG Liaison: Vito Bongiorno

CUG 2016

Dear Cray User Group colleagues,

The European Centre for Medium-Range Weather Forecasts (ECMWF) would like to invite you to CUG 2016 in London in May 2016. Our theme for this meeting is "Scalability". Exploiting parallelism on all architectural levels and improving the scalability of all codes is both vital and challenging for Numerical Weather Prediction.

ECMWF's first operational forecast in 1979, at 210 km global resolution, took five hours to run on a single-processor Cray-1A. Today, two Cray XC30 systems, each with more than 84,000 cores, give us more than twenty million times the peak performance of that first Cray allowing a much larger and more advanced forecasting system featuring a high-resolution global model resolution of 16 km and a 51-member ensemble with a model resolution of 32 km.

Clearly, HPC technology developments are influencing the directions our research will take. This has never been truer as we face the challenges of ever-larger heterogeneous systems and exponential growth in data. Founded on the principal of international collaboration, ECMWF is delighted to have the opportunity to host a meeting that will be an opportunity for users, developers and administrators from all over the world to exchange ideas, solve problems and discuss the future of high performance computing.

Established in 1975 as a major initiative in European scientific and technical co-operation in meteorology, ECMWF is an independent intergovernmental organisation supported by 34 states and is both a research institute and a 24/7 operational service, producing and disseminating numerical weather predictions to its Member States. The supercomputer facility (and associated data archive) at ECMWF is one of the largest of its type in Europe and Member States can use 25% of its capacity for their own purposes.

We look forward to seeing you in London, a metropolis steeped in history with a rich cultural life and many culinary delights to enjoy. From Buckingham Palace and St Paul's Cathedral to the more recent London Eye and the Shard skyscraper, there is no shortage of landmarks to visit. Top museums, such as the British Museum, the Natural History Museum and the Science Museum, are free of charge. We are sure you will enjoy your stay.

Yours sincerely
Mike Hawkins
Head of High Performance Computing and Storage Section