Center for High-Performance Computing
NEWS

THE UNIVERSITY OF UTAH
SPRING 2016

Research The USTAR Center for Genetic Discovery and CHPC Partner to Tackle Large Medical Genomics Datasets

By Barry Moore, USTAR Center for Genetic Discovery

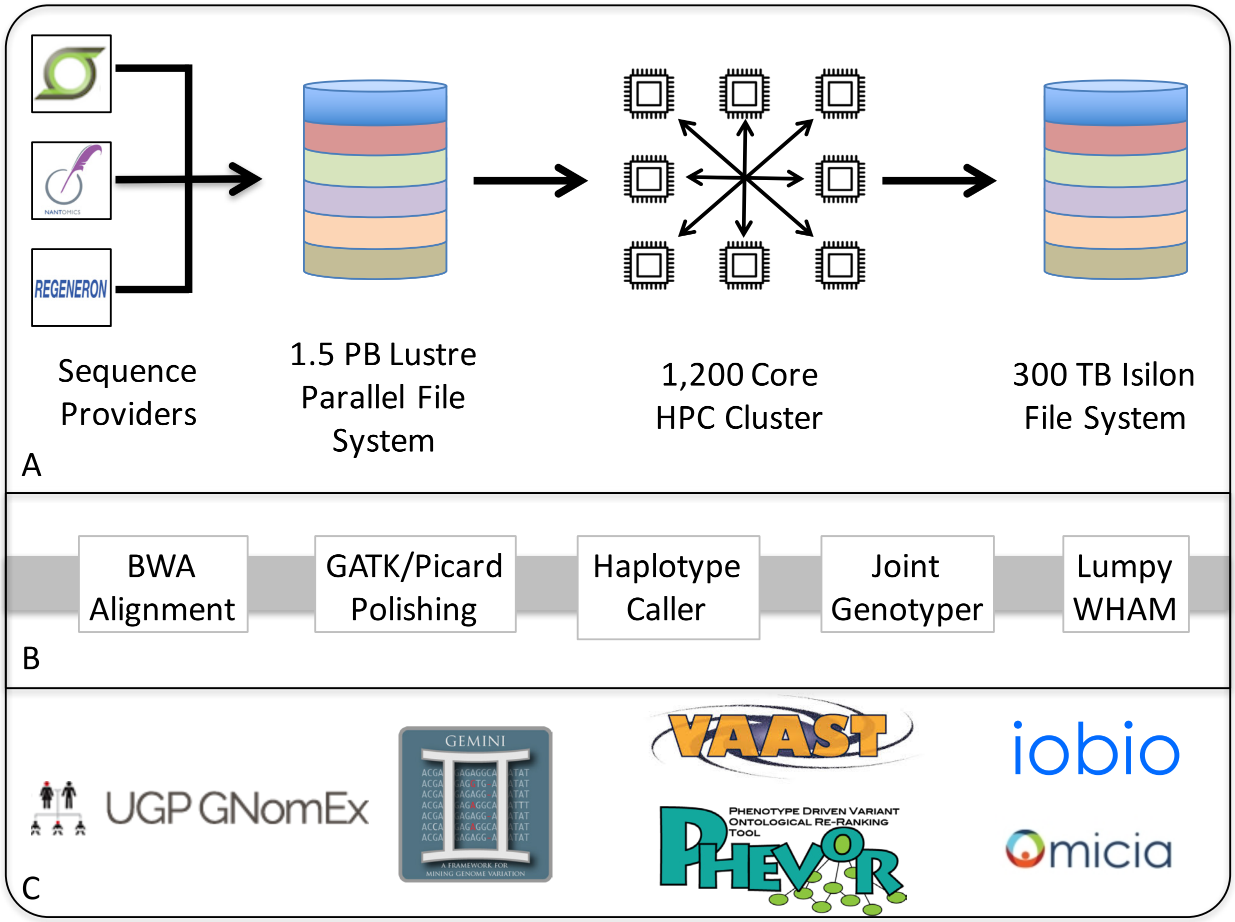
The University of Utah has a long and rich history of genetic research that spans decades and has led to the discovery of over 30 genes linked to genetic disease. These Utah discoveries range from relatively common and well-known heritable disease, such as breast cancer linked to BRCA1/BRCA2 genes, to the truly obscure Ogden syndrome, which in 2010 became the first new genetic disease to be described based on genome sequencing. The Utah Genome Project (UGP), together with clinical investigators across the University of Utah, is continuing this tradition of cutting edge genetic research in Utah by launching several large medical genomics projects over the last year. The USTAR Center for Genetic Discovery (UCGD) – the computational engine for the UGP – has partnered with the University's Center for High Performance Computing (CHPC) to tackle the massive datasets and the large scale computing requirements associated with these projects.

Early efforts to unravel the genetic basis of disease were limited by our inability to 'see' the molecular description of the human genetic blueprint – our DNA. It took teams of scientists and years of intense molecular biology research to discover the genes for Huntington's disease, cystic fibrosis and other well-characterized genetic diseases in the latter part of the twentieth century. Completion of the Human Genome project in 2001 foreshadowed a new era of medical genomics, but with a cost approaching $3 billion for the first human genome sequence, the ability to routinely sequence human genomes for medical diagnosis and discovery would have to wait almost 10 years until advances in sequencing technology caused the cost of DNA sequencing to plummet. Today your genome could be sequenced for medical research for about $1,000 making it possible to sequence the genomes of thousands of people, including both those with genetic disease and healthy individuals as controls.

The ability to read the genome – letter for letter – of thousands of individuals has revolutionized our ability to interrogate genetic disease. But at 3.2 billion 'letters' in length, sifting through thousands of copies of the human genome searching for clues about genetic disease is a truly daunting task. Fortunately dramatic advances in computer technology and rapidly falling costs have played to another strength at the University of Utah – high performance computing.

The process of uncovering the gene responsible for genetic disease is conceptually straightforward. Look at the 3.2 billion 'letters' in a genome from someone with genetic disease and compare them to the 'letters' in a genome of someone healthy with respect to that disease and find where

FastQForward: high-throughput variant calling pipeline



FastQForward is a highly parallel genomic analysis pipeline that fully exploits the power of CHPC's cluster based resources by splitting big jobs into smaller chunks and continually load balancing across multiple cluster nodes. A) A high level view of the UCGD resources at CHPC and how data flows across them. B) A view of the FastQForward pipeline from an algorithmic point of view. Each of these tools have been optimized by UCGD to maximize the I/O and CPU resource saturation to make sure the hardware is used as efficiently as possible. C) Once the initial FastQ-Foward variant calling pipeline has run to discover the positions in each genome that differ from all of the other genomes in the study a variety of tools developed by the UCGD are used for visualization and disease gene prioritization.

they differ. However, conceptually straightforward this comparison quickly becomes computationally overwhelming when we consider the nature of genomic data. To read the 3.2 billion 'letters' in your genome, your DNA is broken into fragments a few hundred nucleotides long and the sequencer reads 150 nucleotides from each end. Think of sequencing your genome as purchasing a jigsaw puzzle and these sequenced fragments or reads are the puzzle pieces that together make up your genome. But this puzzle is very different from a typical jigsaw puzzle. Your genomic puzzle box contains about a billion pieces…the pieces contain only letters…many of the pieces have mistakes on them compared to the picture on the front of the box…there are 30-60 different copies of the same puzzle all mixed together in the box…the different copies of the puzzle are cut in different shapes…some parts of the puzzle are missing entirely from the box…and many parts of the picture on the front of the box look almost exactly alike. However, there are computational tools designed to put the puzzle together to get the complete genome. Once this has been done for hundreds or thousands of genomes, comparisons can be made. By sifting through the differences in genomes shared by those who have genetic disease but not shared with those who don't, clues to 'genetic errors' can be identified. The massive datasets, the errors inherent in the sequencing process and the complex nature of genetic inheritance require robust, statistically rigorous algorithms for analysis, extremely fast network technologies, enormous disk arrays for storage and very large cluster-based supercomputers.

The UGP was launched in 2012 to coordinate efforts to solve the medical genomics puzzles being studied across the University of Utah. Two years later the State of Utah Science Technology and Research (USTAR) initiative and the University of Utah Health Sciences Center established the USTAR Center for Genetic Discovery (UCGD). The UCGD consists of three research teams – led by Mark Yandell, Gabor Marth and Aaron Quinlan – that develop algorithms, software tools, analysis pipelines and data management systems for the interpretation and visualization of large genomic datasets. The goals of the UGP are to standardize genomic analysis techniques, discover genes that promote health and the ones that contribute to disease, utilize Utah's unique genealogical records to accelerate these discoveries and develop tests, diagnostics and public education that keep pace with genomic medicine's rapid advances. The UCGD research groups and the tools that they develop support these goals of the UGP.

In pursuit of these goals, the UGP, in collaboration with numerous clinical investigators across the University, has launched several large medical genomics projects in the last 12 months and the members of the UCGD and CHPC have been working closely to provide the algorithms and the computational muscle necessary for these projects to succeed. Three of these projects are described below.

The Heritage 1K Project

The UGP launched the Heritage 1K Project to carry out deep-coverage whole-genome sequencing of 1,000 people in Utah who have a history of genetic disease in their respective families. The project focuses on discovering the genetic causes of more than 25 conditions, including chronic lymphocytic leukemia, breast cancer, multiple myeloma, amyotrophic lateral sclerosis (ALS), autistim, suicide, early infantile epileptic encephalopathy, hereditary hemorrhagic telangiectasia, tuberous sclerosis, congenital diaphragmatic hernia, primary ovarian insufficiency, ataxia, familial neuropathy, immune deficiency, Treacher-Collins syndrome and other hereditary conditions. The Heritage 1K project is made possible by a $12 million gift from Patrick Soon-Shiong, a Los Angeles based philanthropist, surgeon and entrepreneur. Each of the H1K samples will receive 60X (the number of copies in the puzzle box) whole-genome sequencing. By the end of March 2016 over 800 of these genomes have been sequenced and more than 600 of them have been processed through the UGP variant calling pipeline. While analysis is just getting underway for many of these studies, the H1K project has already involved transfer of over 200 TB of raw data, data storage of close to a petabyte of total data and millions of hours of CPU compute time.

Pediatric Cardiac Genomics Consortium

The Pediatric Cardiac Genomics Consortium (PCGC) was formed as part of the National Heart Lung and Blood Institute's efforts to elucidate the genetic underpinnings of congenital heart disease (CHD). Utah was selected as one of five research centers to participate in the PCGC program, with Drs. Martin Tristani-Firouzi, Mark Yandell and Joe Yost serving as Principal Investigators. A genome can be divided into regions known as exons and one or more exons combine to make up a gene. The cellular machinery translates a gene's exonic regions to build the enzymes and other proteins that comprise the human body. Most (but not all) known deleterious mutations that cause disease are located in these exon regions. Sequencing all known exons in a genome (the exome) is often the first step in trying to unravel the genetic mechanisms of disease. In the past few months the UCGD informatics team has processed over 10,000 exome sequences of patients, family members and healthy controls, to support the Utah PCGP goals of identifying new causative mutations in known CHD-causing genes, identifying novel CHD loci and developing a better understanding of the complex inheritance of CHD.

Regeneron Genetics Center Collaboration

The UGP and the Regeneron Genetics Center (RGC) have joined forces for a large-scale family-based sequencing effort focusing on diseases that have an autoimmune component including interstial lung disease, psoriasis, and inflammatory bowel disease. RGC has sequenced the exomes of over 3,600 individuals again including individuals affected by these diseases, their families and healthy control individuals. Members of the UCGD and others in Utah's medical genomics community are working closely with the RGC to analyze this sequence data, searching for clues about the genetic underpinnings of these diseases. Collaborations such as the one with RGC have the potential to not only to further our understanding of genetic mechanisms causing these diseases, but to also immediately translate into research towards diagnostics and pharmaceuticals that may treat the disease.

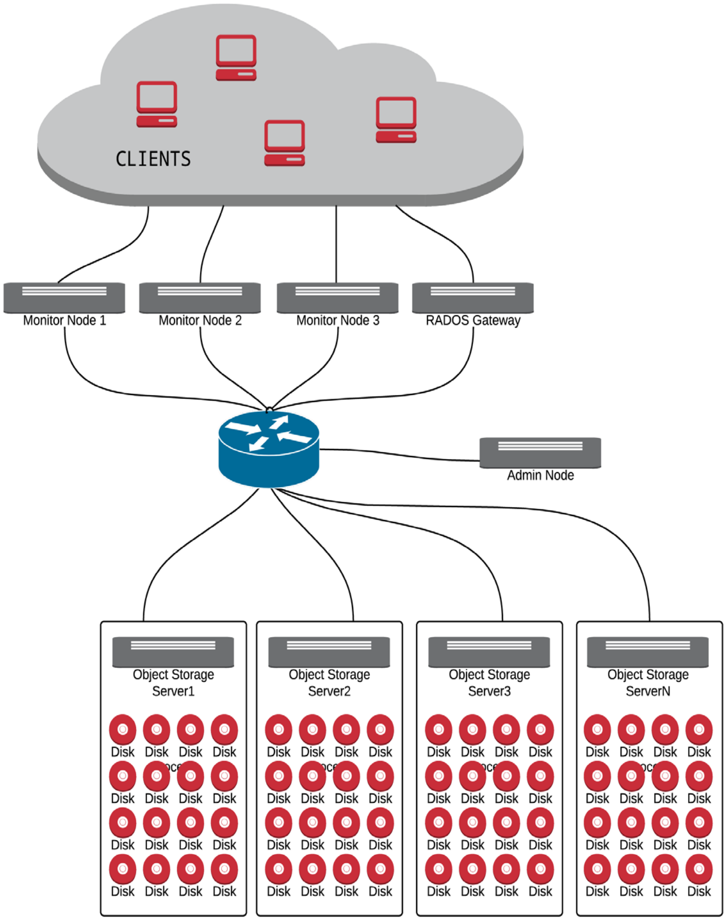
CHPC Contributions

Utah's Center for High Performance Computing has supported the analysis efforts of the UCGD and the goals of the UGP in several key areas. UCGD has transferred over 500 TB of sequence data from external sequencing centers to local servers. The CHPC network team set up dedicated data transfer nodes and supported high-throughput threaded networking protocols to make these massive data transfers possible. Working with these large datasets once they arrive requires large amounts of very fast storage. CHPC's data storage experts assisted the UCGD in the purchase, provisioning and maintenance of a 1.5 PB Lustre distributed file system that allows incredibly high bandwith I/O for reading and writing data. This fast I/O becomes critically important in high performance cluster environments when thousands of CPUs need to simultaneous process different portions of the same file. Finally, millions of CPU hours are needed to process these raw genomic datasets into a useable form for analysis. Again, experts from CHPC provided the UCGD with support in all stages of purchasing, deploying and maintaining a 1,200 core compute cluster dedicated to processing data for UGP projects and other UCGD efforts. It is fair to say that the analysis efforts of the UCGD rest heavily on the support provided by CHPC and without that support fully achieving the goals of the UGP would not be possible.

Object Storage

By Sam Liston, CHPC

Over the past five years CHPC has seen tremendous growth in requests for storage, in particular our group storage space offering. This space has grown from 650TB in 2011 to 6.5PB as of March 2016. Part of the reason for this growth is that this space is inexpensive while being reasonably resilient against failure. In addition, CHPC offers, as a service, the ability to archive this space to tape quarterly for the price of the required tapes. Though some groups owning group storage have elected to use this archival service, the majority has no real backup. In addition, much of this data is data at rest – data that is not actively being changed. Many groups have chosen to expand their group space, often due to the lack of alternative storage options. In an effort to offer a more scalable archive solution, one that will complement the existing group space and possibly slow its growth, CHPC has designed a new storage offering – archive storage.



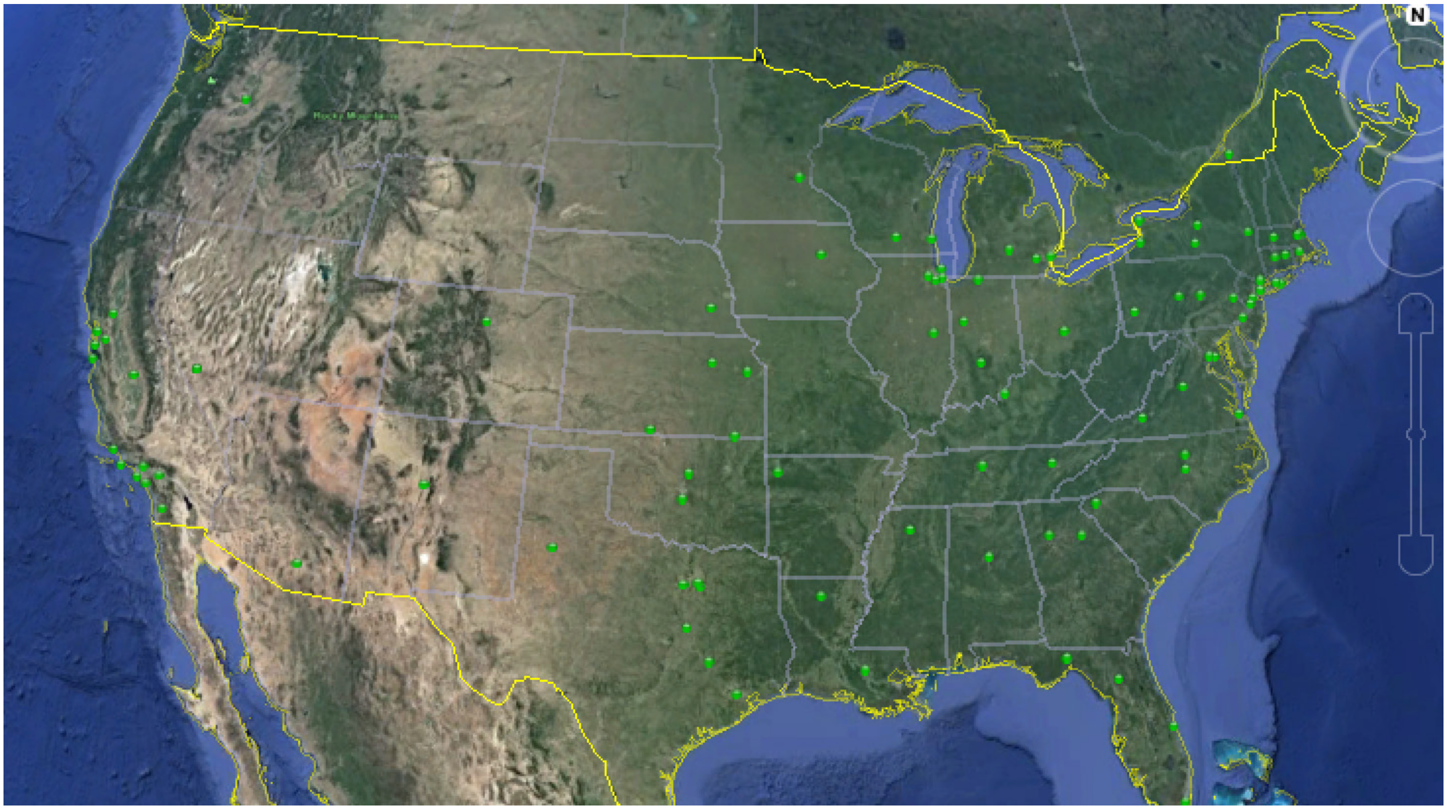
Proposed CHPC Archive Storage Architecture

The archive solution is based around object storage, specifically CEPH, a distributed object store suite developed at UC Santa Cruz. Unlike a traditional file system that organizes files and directories into a hierarchical tree structure, object storage uses a flat structure. A user puts a file into the system and the object store turns it into an object, which is a discrete, addressable unit of storage. Objects are organized into buckets, and buckets are contained within pools. Behind the scenes, the system controls where and how the data is placed, allowing for the data to be efficiently organized, as compared to a user organized file and directory structure of a traditional file system. Allowing the system to optimize data placement also permits object stores to scale to larger capacities than traditional file systems.

When a pool is defined it is assigned a level of resiliency. This level can be 1, 2, 3, or N times replication, or it can be configured to use erasure coding, which is defined as N fragments + M redundant chunks. CHPC in its initial offering will use 6+3 erasure coding; this choice is based partially on the size of the initial purchase and as replication configurations make more sense when there are multiple systems that are on different hardware and possibly geographically separated. With 6+3 erasure coding, each file put into the system will be broken into 6 chunks and have 3 redundant chunks calculated and written to disparate object storage targets (OSTs) and servers. A pool configured in this way can lose three drives from a server, or three servers from the cluster, before data is compromised, providing an increase of data protection in comparison to a traditional RAID6 configuration while keeping the price per TB lower than that of the current group space. In the future, as demand warrants, CHPC will look at offering additional pools with 2-way or 3-way replication, providing even greater data security, but at a higher price per TB.

One of the key features of the archive system is that it is user driven. Users can move data in and out of the archive storage as needed -- they can archive milestone moments in their research, store an additional copy of crucial instrument data, or retrieve data as needed. The system was intentionally made as a stand-alone entity. It will not be mounted on other CHPC resources. This archive storage solution will be accessible via applications that use Amazon's S3 API. GUI tools such as transmit as well as command-line tools such as s3cmd and rclone can be used to move the data.

The hardware to implement this solution is being purchased. We will have an initial raw capacity of 1.15PB, which gives a usable capacity of 768TB in the 6+3 erasure coding configuration with a price of usable space in the $100-110/TB range for the lifetime of the hardware. As we currently do with our group space, we will operate this space in a condominium model by reselling this space in TB chunks. For more information about the new CHPC archive solution please contact issues@chpc.utah.edu.



OSG Sites - Soon UofU will be added

Open Science Grid

By Guy Adams and Wim Cardoen, CHPC

The Open Science Grid (OSG), http://www.opensciencegrid.org/, facilitates access to distributed High Throughput Computing (HTC) resources for research communities within the US. While CHPC's focus has been and will continue to be High Performance Computing (HPC), we are starting to explore ways that we can participate in the OSG network in order to better serve our user base. HPC workloads are geared to utilize many CPUs intercommunicating with each other during execution, whereas HTC workloads target a single CPU during execution, running serially one after the other with no intercommunication during execution. Therefore, typical HTC simulations consists of a large number of single processor calculations.

The OSG framework is a federation of computing and storage elements at over 125 individual sites spanning the US. The sites involved are primarily universities, national labs, and cloud providers. Each of those sites provide access to a few hundred to thousands of CPU cores when they might otherwise sit idle to researchers outside of their organization via the OSG network. In this manner, OSG can offer massive computing resources. In 2015, over 900 million CPU hours were provided through the OSG framework. OSG continues to provide large amounts of CPU cycles to process the data sets harvested from the Large Hadron Collider (LHC) at CERN. The OSG framework also provides substantial resources to other projects such as the Compact Muon Solenoid (CMS), Laser Interferometer Gravitational wave Observatory (LIGO), Deep Underground Neutrino Experiment (DUNE), South Pole Neutrino Observatory (IceCube), as well as to individual researchers in fields such as computational biology, genetics and medicine, to name a few.

So how does the OSG work? The OSG framework is set up in such a way that local demand of the computational resources prevails over the demands from other locations. At CHPC, similar to the other OSG participants, the OSG simulations will be preempted as soon as there is a local need for those CPU resources. The developers behind the OSG framework have developed a software stack (HTCCondor) that manages the preempted jobs and automatically reschedules them on other resources.

At CHPC, we are currently setting up the infrastructure to provide cycles to the OSG consortium. However, researchers at the UofU already have the possibility to compute on the OSG framework. In order to start you need to sign up for an account at https://osgconnect.net/signup. As soon as your account has been activated you are ready to go. We will collaborate with the national OSG helpdesk to assist you.

If you are interested in the OSG framework, especially if your work load fits well with the OSG model (see http://www.opensciencegrid.org/about/what-kind-of-computational-problems-fit-well-on-osg/), please contact us at issues@chpc.utah.edu. We look forward to working with you.

What is Research Computing & CHPC?

Research Computing & CHPC serves as an expert team to broadly support the increasingly diverse research computing needs on campus. These needsinclude not only high performance computational resources and advanced user support and training, but also support for big data, big data movement, data analytics, security, virtual machines, Windows science application servers and advanced networking.

CHPC also operates a protected environment (PE) for researchers who work with data that is sensitive in nature. These resources have been reviewed and vetted by the Information Security Office and the Compliance Office as being an appropriate place to work with Protected Health Information (PHI).

These computing resources are available to all faculty at the University of Utah, their students and research staff.

CHPC Presentations: Summer 2016

For the first time CHPC will offer a series of presentations during the Summer semester. All presentations will be held in the INSCC Auditorium (Room 110).

- Overview of CHPC, Thu, June 2, 1-2pm
- Hands on Intro to Linux (part 1), Tue, June 7, 1-3 pm
- Hands on Intro to Linux (part 2), Thu, June 7, 1-3 pm
- XSEDE boot camp, Tue-Fri, Jun 14-17, 9am-3pm (break 11am-noon)
- Module Basics, Tue, Jun 21, 1-2pm
- Slurm Basics, Thu, Jun 23, 1-2pm

https://www.chpc.utah.edu/presentations/index.php

SC16

Salt Lake City, Utah | hpc matters.

Salt Lake City is hosting the 2016 Supercomputing Conference, the premier international conference on high performance computing, networking, storage and analysis. The event will be held at the Salt Palace November 13 - 18.

SC16 brings together scientists, engineers, researchers, educators, programmers, system administrators and managers from across the country to showcase how developments in these areas are driving new ideas, discoveries and industries. CHPC will have a booth that highlights our activities and the research being done with CHPC resources. If you would like your research highlighted at CHPC's booth, please contact Sam Liston at sam.liston@utah.edu.

For nearly 20 years, Janet Ellingson has worked and taught at the University of Utah. She holds a PhD in History, and taught in the History department from 1998 - 2009. Before joining the CHPC staff, she worked as an Administrative Assistant in the Internal Medicine department.

Janet joined the CHPC staff in March 2006, and has worked with us for over 10 years. As CHPC's Administrative Manager Janet took care of CHPC office administration and edited the newsletter. In addition, she learned about our systems and helped with user problem tickets. She will retire from the University this Spring and will be missed by not only the CHPC staff, but the many users with whom she interacted.

CHPC thanks Janet for her dedication and hard work. On behalf of the entire staff, we wish her well as she embarks on a new chapter of her life.

Name	Primary Focus	email address
Administration		
Tom Cheatham	Director	tec3@utah.edu
Julia Harrison	Associate Director	julia.harrison@utah.edu
Colette Durrant	Administrative Assistant	colette.durrant@utah.edu
Amanda Allen	Help Desk	amanda.allen@utah.edu
User Services		
Wim Cardoen	Scientific Computing; C; C++; Fortran; Python; Java; MPI; Hadoop	wim.cardoen@utah.edu
Martin Cuma	Parallel Code Development; Matlab; Geophysics	martin.cuma@utah.edu
Anita Orendt	Computational Chemistry; Gaussian;Quantum Espresso; XSEDE Campus Champion	anita.orendt@utah.edu
Chonghuan Xia	Software Developer	chonghuan.xia@utah.edu
Systems		
Guy Adams	Hardware; Open Science Grid	guy.adams@utah.edu
Wayne Bradford	Architecture - Security	wayne.bradford@utah.edu
Irvin Allen	Systems Administration; Backups	irvin.allen@utah.edu
Paul Fischer	Intern - HPC	u0770441@utah.edu
Steve Harper	Virtualization	s.harper@utah.edu
Brian Haymore	Architecture - High Performance Computing	brian.haymore@utah.edu
David Heidorn	Systems Administration; Backups, Windows	david.heidorn@utah.edu
Sam Liston	Architecture - Storage	sam.liston@utah.edu
David Richardson	Virtualization	david.richardson@utah.edu
Rajeev Sahay	Intern - Security	rajeev.sahay@utah.edu
Brad Shelton	Intern - Storage	brad.shelton@utah.edu
Alan Wisniewski	Data Center Management	alan.wisniewski@utah.edu
Networking		
Joe Breen	Network Architecture	joe.breen@utah.edu
Jake Evans	Network Engineering	jake.evans@utah.edu
Alan Navarro	Intern - Networking	alan.navarro@utah.edu
Aaron Pabst	Intern - Networking	u0893501@utah.edu
Raja Vazrala	Intern - Networking	rajasekhar.vazrala@utah.edu

THE UNIVERSITY OF UTAH

Research Computing & CHPC
155 South 1452 East, RM #405
SALT LAKE CITY, UT 84112-0190

Welcome to CHPC News!

If you would like to be added to our mailing list, please fill out this form and return it to:

Colette Durrant
THE UNIVERSITY OF UTAH
Center For High Performance Computing
155 S 1452 E ROOM 405
SALT LAKE CITY, UT 84112-0190
FAX: (801)585-5366

Name:
Phone:

Department or Affiliation:
Email:

Address:
(UofU campus or U.S. Mail)

Thank you for using our Systems!

Please help us to continue to provide you with access to cutting edge equipment.

ACKNOWLEDGEMENTS

If you use CHPC computer time or staff resources, we request that you acknowledge this in technical reports, publications, and dissertations. Here is an example of what we ask you to include in your acknowledgements:

"A grant of computer time from the Center for High Performance Computing is gratefully acknowledged."

Please submit copies or citations of dissertations, reports, pre-prints, and reprints in which the CHPC is acknowledged to: Center for High Performance Computing, 155 South 1452 East, Rm #405, University of Utah, Salt Lake City, Utah 84112-0190