Center for High-Performance Computing NEWS

THE UNIVERSITY OF UTAH FALL 2014

Research

CHPC Support of Biomedical Research Enterprise

By Ram Gouripeddi, Randy Madsen, Bernie LaSalle, and Julio Facelli

The University of Utah's Biomedical Informatics Core (BMIC) is a core within the Center for Clinical and Translational Science (CCTS), University of Utah that provides biomedical informatics and information technology support to clinical and translational researchers through a variety of means, including education, innovation, informatics research and service. The CCTS serves as an academic home for clinical and translational research, developing innovative health services for the community and researchers, and training a new generation of clinical and translational investigators. The CCTS is supported by an award (1ULTR001067) from the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health (NIH). The Education component of the BMIC uses a mix of formal courses and informal approaches to education in biomedical informatics and clinical research informatics for translational research. The Innovation component of the BMIC aims to provide universal access to data from our partner institutions (Intermountain Healthcare, Veteran's Administration Medical Center and the Utah Department of Health) through the use of open source tools and applications and customized development when required. The informatics research component develops generalized research questions with regard to data quality, integration and harmonization into practices and tools for translational research. The Service component of the BMIC works across all CCTS cores to provide resources and services for the planning and implementation of translational research.

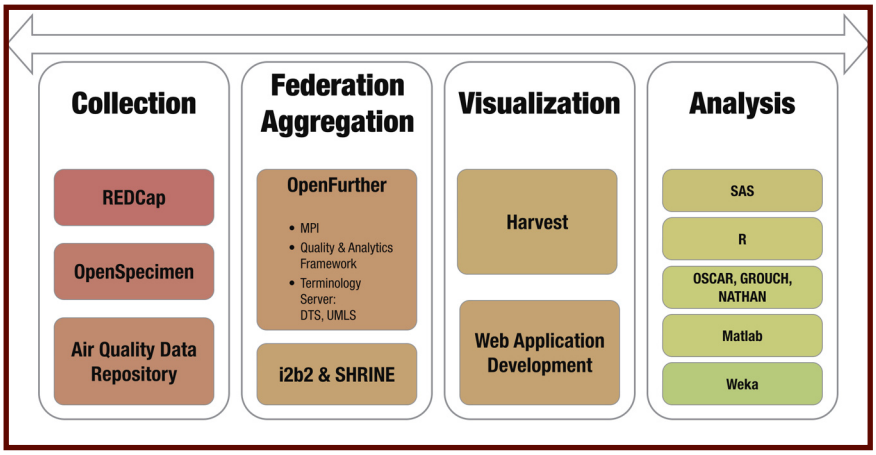


Figure 1: Overall Pipeline of BMIC Data Services.

The mission of BMIC is to improve the quality of research, patient care and population health by using biomedical informatics methodologies and best practices. In essence, BMIC broadly supports generation of new biomedical knowledge from data by providing platforms and services inclusive of data federation and integration, data transformation/translation, biomedical metadata and terminology management, data collection, data exploration and statistical analysis. Figure 1 is an overview of the overall pipeline of BMIC data services.

BMIC utilizes many of the computation services provided by the Center for High Performance Computing (CHPC), University of Utah in accomplishing its mission. These include hardware, software, software development frameworks, project management and issue tracking, databases, and processing power. CHPC's virtual machine (VM) farm provisions our software infrastructure and our entire development environment. This includes development, testing and production ready systems. We utilize the Atlassian infrastructure for continuous integration builds of our software, source code management, documentation and issue and feature tracking for software management. The VM farm provides performant disk storage for multiple instances of JAVA enterprise application servers, such as Apache ServiceMix, Jboss, and Tomcat, and database installations. Most of these applications are housed within CHPC's Protected Environment that complies with the Health Insurance Portability and Accountability Act (HIPAA) and provides encrypted storage to operate in a secure environment without impacting performance.

BMIC's primary tools for data collection are REDCap and OpenSpecimen. REDCap is a secure web application for building and managing online forms and surveys, and for storing and transferring protected health information. BMIC's REDCap currently has over 2,300 users and 1,500 projects – many of which involve external collaborators. Data within the REDCap application includes details on more than 25,000 individuals who have invited to participate in different survey instruments. OpenSpecimen, previously known

(Continued on Page 2)

(Continued from Page 1)

Figure 2: An overview of applications and services utilized by BMIC within CHPC's Protected Environment for software development and production ready support of biomedical research.

as caTissue, is being used by researchers to uniformly collect and store bio-specimen information. Both REDCap and OpenSpecimen data are stored within CHPC's protected environment.

OpenFurther (OF) is our enterprise service oriented application that supports federation and integration of disparate data sources. It can link heterogeneous data types, including clinical, biospecimen, environmental and patient-generated data, empowering researchers to assess the feasibility of particular clinical research studies, export biomedical datasets for analysis, and create aggregate databases for comparative effectiveness research. Data can be transformed to popular common data models such as Observational Medical Outcomes Partnership, Mini-Sentinel and Informatics for Integrating Biology and the Bedside (i2b2) to participate in national networks or as exports. OF supports both static and dynamic federation. OF is open-source and has been developed using the development environment within CHPC. It is currently deployed at the University of Utah as FURTHeR where it is being used for finding cohorts and to assess the feasibility of prospective biomedical research. A similar deployment has been undertaken at the University of North Carolina. A third deployment is being used in the PHIS+ project where it is being used to create a centralized database for performing comparative effectiveness research studies. This PHIS+ OF federation consists of clinical data from six pediatric hospitals from across the country and at this time the PHIS+ database consists of over 2 million pediatric patients.

In addition to these major projects, BMIC is deploying an instance of the i2b2 platform within the protected space for participation in the NCATS Accrual to Clinical Trials (ACT) Project. This will include a periodically updated copy of clinical data in the i2b2 model from the University of Utah Enterprise Data Warehouse. BMIC also provides a Knowledge Generation Service that supports researchers with their data requests and utilizes the protected space for storing, visualizing and analysis of clinical datasets.

CHPC provides access to a comprehensive application development and production environment that enables us to meet the requirements of an expanding biomedical research enterprise. The research community is expanding its interests in involving the patient in clinical care and research, as well as the role of the environment in health. These, along with the increased availability of biomedical data, will take our efforts into big data integration and wrangling, and the development of ultra large scale systems that rely on decentralized data and continuous evolution, utilizing the foundation blocks at CHPC.

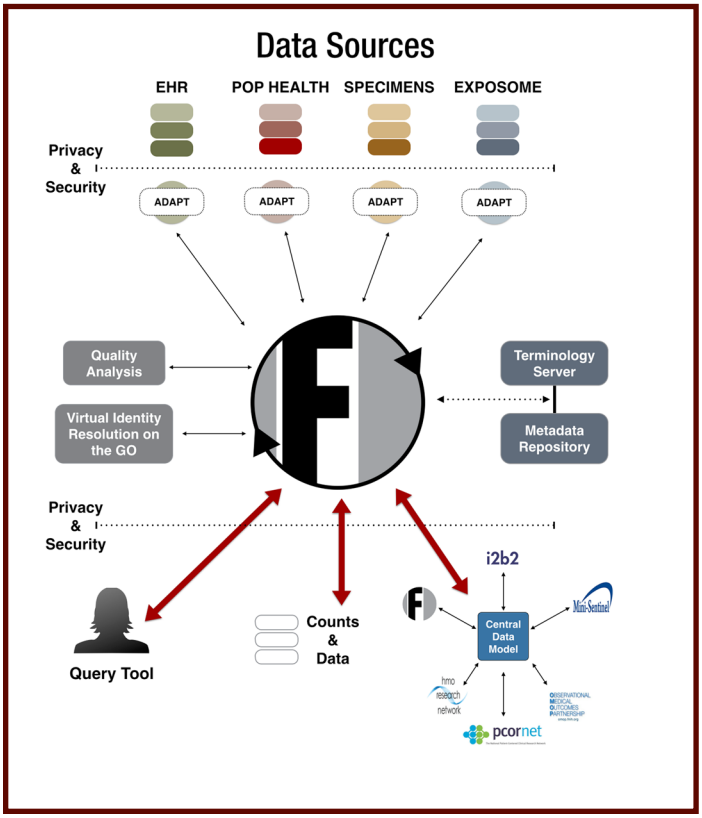


Figure 3: OpenFurther software ecosystem developed and deployed within CHPC's Protected Environment

Projects

Grant Awards at CHPC

CC*IIE - Identity Access Management (IAM)

University of Utah and Clemson researchers have been awarded a two-year $299K NSF grant designed to enable sets of investigators to easily provision and manage their own collaborations. There are three main objectives: 1) provide an InCommon-based login for accessing campus clusters using a shell environment; 2) prototype access to a campus compute resource in the campus Innovation Platform from the Global Environment for Network Innovations (GENI) portal; and 3) continue conceptual development of the new framework through regular meetings.

The program will develop a federation design, known as "FeduShare," from the individual's perspective rather than from the organizational/administrative perspective. The FeduShare framework builds on existing Identity Management solutions such as the InCommon federation, Shibboleth servers and the GENI solutions for people/resource management/authorization that build upon them.

Utah team members include Steve Corbato (co-PI); Wayne Bradford, Joe Breen, Steve Harper Julia Harrison, Cassandra Van Buren, and Bryan Wooten.

by Cassandra Van Buren

CC*IIE - Integration

Colorado State University (CSU), the Idaho Regional Optical Network (IRON), the University Corporation for Atmospheric Research (UCAR), the University of Colorado Boulder (UCB), and the University of Utah (UU) were successful in their application for a NSF CC*IIE regional proposal. This two year funded project will allow the project team to conduct four regional workshops targeted at smaller institutions. The focus will be on High Performance Networking (HPN) as an enabler of scientific discovery through computational modeling and simulation, data-driven analysis, collaboration, and community building. The funded project will also allow for site visits and some engineering support for the institutions. HPN and the other components of advanced CI are key enabling technologies vital to each university's and college's ability to function and prosper in a rapidly evolving scientific and technical environment. The success of this project provides a blueprint for future outreach to small and minority institutions efforts in the Intermountain region to enable them to advance and succeed in the changing CI environment in which they must compete

by Joe Breen

ACI-REF Update

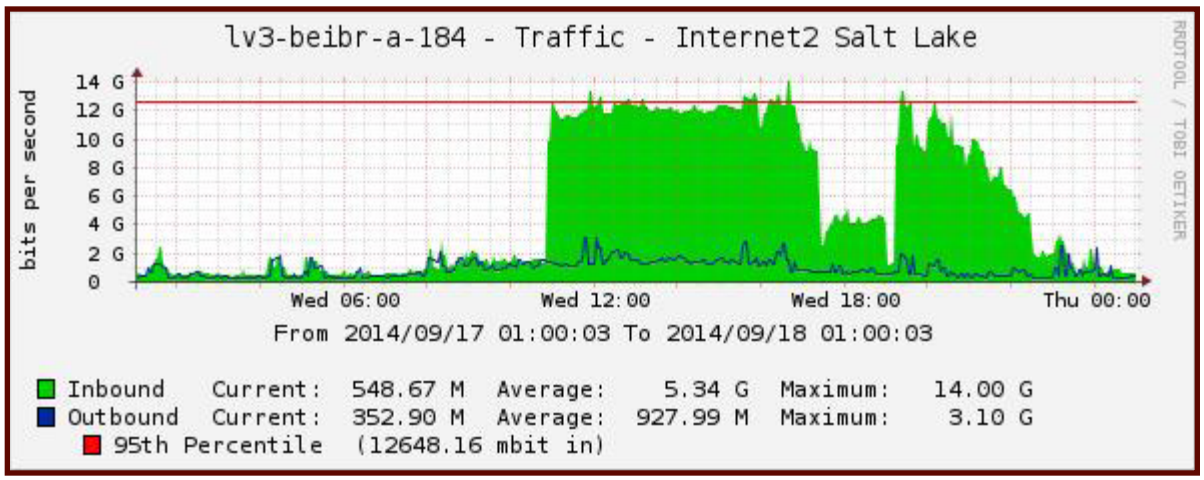
As part of the ACI-REF program, CHPC systems and networking staff have been working with their counterparts at Clemson, including Clemson researcher Alex Feltus, on the configuration and optimization of data transfer from NIH's National Center for Biotechnology Information (NCBI) to Utah as well as between Utah and Clemson. The other ACIREF schools (Harvard University, University of Hawaii, University of Southern California, and University of Wisconsin-Madison) are also doing similar experimentation. So that the results can be compared, all schools are working with the same 12TB data set, and using the same workflow (cURL and Aspera transfer clients).

Preliminary results for a transfer from NCBI to CHPC file systems have been very promising. By optimizing and parallelizing the flow, we have been able to download this test data set in 2.5 hrs (the transfer starts at 19:00 on graph below and rates decrease as individual flows finish), achieving average transfer rates between 10 gb/s and 14 gb/s.

These initial tests were all performed over our standard Internet2 paths. Plans to run these same tests over a software defined network (SDN), which should increase transfer rates even more, are in the works.

If you want more information, please contact us.

by Sam Liston



Test data results on ACI-REF data transfers at 19:00

CHPC Collaborates with School of Computing's Flux Group on Adaptable Testbed Project

by Joe Breen, Brian Haymore and Anita Orendt

Last Fall NSF funded a collaboration between the School of Computing's Flux group, headed by Prof. Rob Ricci, and CHPC to develop an "adaptable profile-driven testbed." This testbed will allow computer research teams to use the same resources for different missions, e.g., networking, high performance computing and security experiments. The Flux group's emphasis is to create a low barrier manner of creating very reproducible experiments. CHPC's emphasis is to create an environment to roll out HPC images on demand, to scale the images dynamically, and to support multiple images with different HPC and security contexts.

The three-year Apt project consists of a hardware foundation, the Apt cluster, and a testbed control system built upon systems developed previously by the Flux group for the Emulab and GENI projects. Apt starts with technology and lessons learned from the previous Flux testbeds and expands the scope. By leveraging specific hardware and expanding the software, Apt provides an environment for researchers in the traditional network research community, the HPC community, and other communities through Apt's "on demand" profiles. Researchers can create, save, and re-use profiles to repeat experiments, as well as share profiles with other researchers.

Apt Deployment

The Flux group, with support from CHPC, deployed the Apt hardware at the University of Utah Downtown Data Center. The cluster contains two classes of nodes:

- 128 Dell PowerEdge r320 nodes, with a single Intel Xeon E5-2450 processor (8 cores, 2.1Ghz), 16GB Memory, and four 500GB Hard Drives
- 64 Dell PowerEdge c6220 nodes, with dual Intel Xeon E5-2650v2 processors (8 cores, 2.6Ghz), for a total of 16 cores, 64GB Memory, and two 1TB Hard Drives

These nodes connect via a 1Gbps Ethernet control network and have multiple high bandwidth data plane network options, including 10Gpbs Ethernet, 40Gbps Ethernet and 56Gbps Infiniband. Additional details on the hardware are available at **http://docs.aptlab.net/hardware.html.**

Flux team members are developing software to dynamically provision this hardware to meet the needs of the researchers. The users define a profile which includes all the information needed to run an experiment. This profile includes the description of the hardware and software resources needed for the experiment, and provides the mechanism to enable repeatable research.

The hardware specification of the profile includes information on the properties of the nodes, the storage, and the network. The software environment of the profile consists of the operation system, and can include additional software packages, data files, etc. needed for the experiment. The Apt project includes a number of standard profiles. Users will define other profiles for their own use and to share.

Researchers use either an application programming interface (API) or a web interface to configure the profile, then create an image of this profile, and define the experiment. The experiment belongs to the experimenter for the specified duration. When the experiment is complete, the Apt software de-provisions the hardware, and makes it available for future requests. For more information on the project see http://docs.aptlab.net/. (Continued on Page 5)

The Apt Cluster at the DDC- photo by Sam Liston

(continued from page 4)

Apt HPC Environment

CHPC is developing HPC profiles, initially based upon its existing cluster images. CHPC manages multiple HPC images that fall into two main security contexts: traditional computation, e.g., Kingspeak, and compliance regulated computation, e.g., Apexarch.

CHPC is establishing the traditional HPC profile as a cluster called Tangent to launch jobs on the C6220 nodes, which are the same hardware as the 16 core nodes of CHPC's Kingspeak cluster. This profile will have CHPC applications available and mount current CHPC file systems. From the user's perspective, access to this resource is obtained via a login to an interactive node for the Tangent cluster. The Tangent interactive nodes are local to CHPC and allow users to submit batch jobs that will spin up dynamic HPC images on the Apt hardware.

For managing the workflow of each job, CHPC is developing a set of scripts to create an Apt experiment and to provision the resources. A script is run to watch for jobs being submitted to the batch queue on Tangent. When a user submits a job, the script checks if the requested resources are available on the Apt cluster. If not available, the job will sit idle in the batch queue until the next cycle of the script. When the resources are available, the script will create an experiment using CHPC's HPC profile and provision the resources for the requested wall time. At this stage, the user will see that the system is allocating the node(s) for the job. Once provisioned, the script launches the job. The job will run until complete or until the job reaches the requested wall time. Once complete, the script marks the node(s) as unavailable, de-provisions them from the Tangent cluster environment, and returns them to the Apt resources for use by another experiment.

CHPC is using this experimental HPC image to explore other resource management tools such as Simple Linux Utility for Resource Management (SLURM) for the scheduling of jobs. CHPC is also testing several scientific applications using this system, in order to evaluate the systems in terms of the user experience and learn more about the overhead for the provisioning and de-provisioning of nodes in a dynamic manner. For information on the HPC profile, see **https://wiki.chpc.utah.edu/display/DOCS/Tangent+User+Guide.**

The Apt Story Continues...

The Flux group, in collaboration with Clemson University, University of Wisconsin Madison, and University of Massachusetts Amherst, was awarded funding for the CloudLab project (**http://www.cloudlab.us/**), a special case of the Apt story, with a focus on cloud technologies. Cloudlab will create a facility that will federate three large clusters together, and will also federate these clusters with a wealth of smaller, more distriubted GENI resources that have been established through pre-existing Flux GENI work.

Selection of Recent Research Using CHPC Resources

Armentrout, P. B., Yang, B., Rodgers, M.T. (2014). "Metal Cation Dependence of Interactions with Amino Acids: Bond Dissociation Energies of Rb+ and Cs+ to the Acidic Amino Acids and Their Amide Derivatives." J. Phys. Chem. B 118(16): 4300 - 4314.

Bess, E. N., DeLuca, R.J., Tindall, D.J., Oderinde, M.S., Roizen, J.L., DuBois, J., Sigman, M.S. (2014). "Analyzing Site Selectivity in Rh2(esp)2-Catalyzed Intermolecular C–H Amination Reactions." J. Am. Chem. Soc. 136: 5783 - 5789.

Borodin, O., Bedrov, D. (2014). "Interfacial Structure and Dynamics of the Lithium Alkyl Dicarbonate SEI Components in Contact with the Lithium Battery Electrolyte." J. Phys. Chem. C 118(32): 18362-18371.

Good, S. P., Mallia, D.V., Lin, J.C., Bowen, G.J. (2014). "Stable Isotope Analysis of Precipitation Samples Obtained via Crowdsourcing Reveals the Spatioitemporal Evolution of Superstorm Sandy." PLoS ONE 9(3): e91117.

Jin, J., Miller, J.D., Dang, L.X. (2014). "Molecular Dynamics Simulation and Analysis of Interfacial Water at Selected Sulfide Mineral Surfaces under Anaerobic Conditions." International Journal of Mineral Processing 128: 55-67.

Jones D.E., Igo., S., Hurdle, J., Facelli, J.C. (2014). "Automatic Extraction of Nanoparticle Properties Using Natural Language Processing: NanoSifter an Application to Acquire PAMAM Dendrimer Properties." PLoS ONE 9(1): e83932.

CHPC's New Director - Prof. Tom Cheatham

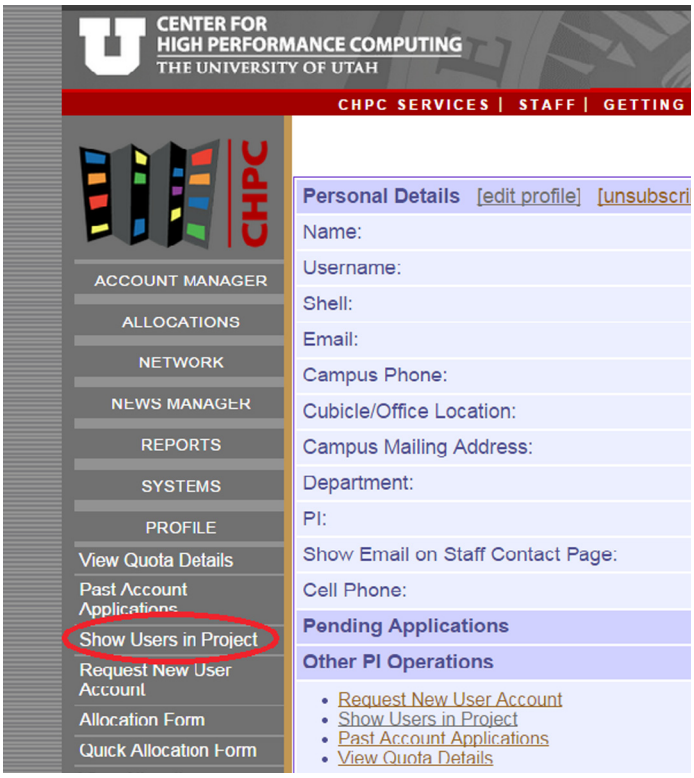
CHPC is delighted to announce that Professor Thomas E. Cheatham III has been named director of CHPC. His HPC expertise will be a great asset to the center and university researchers. You will hear from him in our next newsletter.

Info for CHPC's PIs

CHPC has an on-going project to keep our accounts up-to-date to address both security concerns and the need for accurate statistics on usage. We request your assistance in this endeavor.

CHPC Principal Investigators (PIs) or delegates can log in to the CHPC website via the PROFILE link and then choose "Show Users in Project," as shown in the screen shot to the right, to get a list of all user accounts in their group. PIs can request the removal of any users who no longer need access by sending a list of names and uNIDs to issues@chpc.utah.edu.

The CHPC account deletion process is detailed at the following link: **https://wiki.chpc.utah.edu/display/policy/1.2.2+Account+Removal+and+Locking+Policy**



What's New at CHPC?

by Anita Orendt

► In June CHPC provisioned a cluster named lonepeak.chpc.utah.edu that is run unallocated. This node is comprised of 16 nodes. Eight of these nodes have 12 cores with 96GB memory and the other eight have 20 cores and 256GB memory. All users can run jobs on these nodes. This cluster has its own 33 TB scratch space, /scratch/lonepeak/serial. The Lonepeak User Guide at **https://wiki.chpc.utah.edu/display/DOCS/Lonepeak+User+Guide** .

► CHPC added twelve general nodes with 20 cores to kingspeak in August. This is an addition to the original 32 nodes with 16 cores. All of these nodes have 64GB memory. In addition, there are also four new special purpose nodes that have been added to kingspeak. These nodes have 20 core, 384GB memory and 20TB disk space each, for a total of 80TB in storage. These nodes will be in the general node pool, but also can be requested or targeted for Hadoop and large memory needs, by adding the property "hadoop" to the node specification, e.g., #PBS –l walltime=33:00:00, nodes=4:ppn=20:hadoop

CHPC is in the process of establishing a usage policy on these nodes. Send a request to **issues@chpc.utah.edu** if you are interested in running Hadoop on this resource.

► Information about running on clusters with mixed core counts is available in the "Resource Specification Options" section of the Kingspeak User Guide found at **https://wiki.chpc.utah.edu/display/DOCS/Kingspeak+User+Guide**. This information is also applicable to running on ash either as an owner or a guest.

► The ash cluster has been expanded. This cluster, owned by Prof. Phil Smith, is now composed of the original 253 nodes having 12 cores, 24 GB memory along with 164 new nodes having 20 cores and 64GB memory, for a total of 417 nodes and 6316 cores. When not being used by the Smith group, these nodes are accessible using the smithp-guest account; there are two guest interactive nodes, accessible at ash-guest.chpc.utah.edu.

► For users running owner-guest on ember or kingspeak, there is a script to show you the current status of the owner nodes per group and the syntax to target specific nodes. The script is: /uufs/chpc.utah.edu/sys/pkg/chpcscripts/std/bin/owner_guest_guide.sh.

► CHPC has added additonal nodes to the FastX frisco cluster, bringing the total to six. This includes two nodes (frisco05 and frisco06) with graphics cards so you can use virtualgl. See the updated information on the FastX usage page, **https://wiki.chpc.utah.edu/display/DOCS/FastX .**

CHPC Staff Directory

Administrative Staff	Title	Phone*	Email	Location
Tom Cheatham	Director			
Julia D. Harrison	Associate Director			
Guy Adams	Assistant Director, Systems & Network			
Anita Orendt	Assistant Director, Research Consulting & Faculty Engagem			
Janet Ellingson	Administrative Manager & Newsletter Editor			

Scientific Staff	Expertise	Phone*	Email	Location
Wim Cardoen	Scientific Applications			
Martin Cuma	Scientific Applications			
Sean Igo	Natural Language Processing			
Albert Lund	Graduate Assistant			

Project Leads		Phone*	Email	Location
Wayne Bradford	Security			
Joe Breen	Advanced Network Initiatives			
Steve Harper	Virtualization			
Brian Haymore	HPC & Storage			
Sam Liston	Special Events/SC			

Technical Support Staff		Phone*	Email	Location
Amanda Allen	Help Desk Assistant			
Irvin Allen	Systems Admin			
Nathaniel Ellingson	Tech Assistant			
Jake Evans	Network Engineer			
David Heidorn	Systems Admin			
David Richardson	Systems Admin			
Walter Scott	Software Developer			
Alan Wisniewski	Network Support			

*All phone numbers are preceded by area code 801 unless otherwise noted.

What is CHPC?

The University of Utah's Center for High Performance Computing (CHPC) purview is to support University faculty and research groups whose main focus requires computing and advanced networking as core instrument(s) central to their research. The Center provides large-scale computer systems, storage, networking, and the expertise to optimize the use of these high-end technologies. CHPC facilitates advancement in academic disciplines whose computational requirements exceed the resources available in individual colleges or departments. CHPC also provides a protected environment for health science researchers. Since 1996 these resources have resulted in more than 900 scientific publications.

THE UNIVERSITY OF UTAH

Center for High Performance Computing
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:

[redacted]
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