Center for High Performance Computing NEWS

THE UNIVERSITY OF UTAH FALL 2009

Article

The HIPAA Cluster: Ensuring Data

by Sean Igo, CHPC

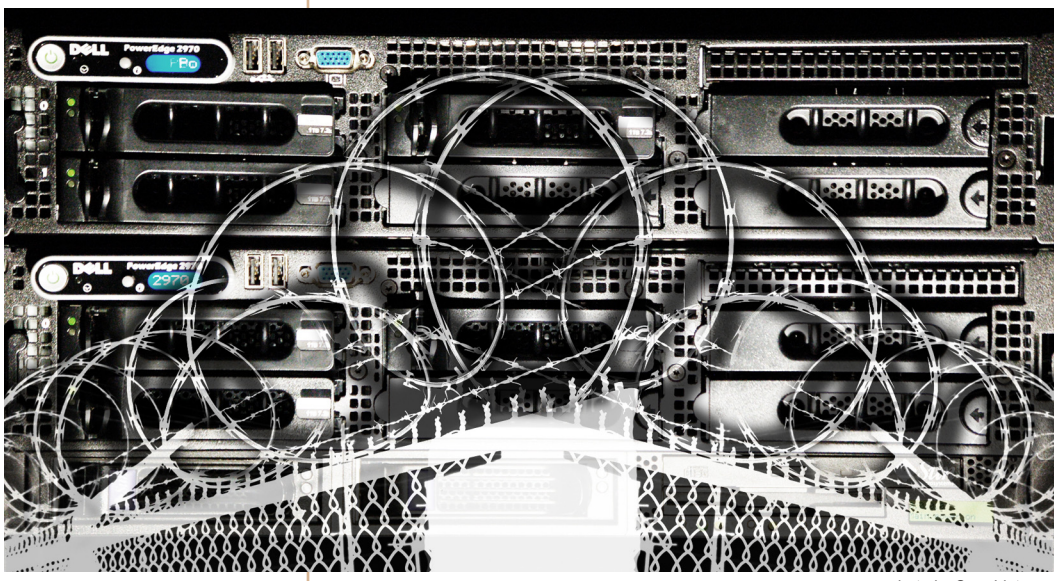
Research in health-care related fields, including clinical, biomedical, nursing, and public health research, is an increasingly important and well-funded endeavor. For example, the Obama Administration's stimulus effort has added $18B for biomedical research to the NIH's existing $30B budget. The University of Utah is a strong participant in health care research and has long been a leader in the integration of health care research with computational methods. The U has one of the world's earliest and largest Biomedical Informatics departments.

Research in the biomedical sciences has certain complications. Chief among these is the fact that experimentation often involves human subjects or, in the case of computer-based research, data about human subjects. Any experiments involving humans and the resulting data are required to conform to stringent legal and ethical standards, as set out in federal regulations from the Department of Health and Human Services. These regulations are monitored locally by the University's Institutional Review Board (IRB).

For example, the Health Insurance Portability and Accountability Act of 1996, known by its acronym "HIPAA," is a set of regulations that govern - among other things - privacy of personal healthcare data. Such data, called Protected Health Information (PHI), consists of any healthcare-related data that can be linked to a specific individual. The HIPAA Privacy Rule regulates how PHI may be used and the conditions under which it may be distributed. There are 18 classes of identifiers detailed by HIPAA, including patient name and address as well as less obvious information such as dates of medical procedures.

Research involving PHI must conform to HIPAA privacy standards. Therefore, it is important that researchers using PHI have access to a secure computing environment in which they may store and manipulate PHI. At the same time, clinical research is increasingly making use of computationally intensive techniques such as data mining, machine learning, statistics, and operations on large data sets. These requirements - large capacity, secure storage and high performance computing - make CHPC the ideal organization to maintain a HIPAA-compliant computational research environment.

Stimulated by NIH grant funding, CHPC now maintains such an environment, named "homer" after University of Utah Professor Emeritus Homer Warner - a pioneer in the field of Biomedical Informatics. Informally, the environment is known as the "HIPAA sandbox." Created through a collaboration of CHPC and the University's Department of Biomedical Informatics, it is isolated from the main CHPC clusters and access to it is highly controlled.



Homer, the HIPAA sandbox, has secured access. photo by Sam Liston

What is in the HIPAA Sandbox?

The HIPAA sandbox consists of three kinds of computers:
- Windows interactive nodes, which run the Windows XP operating system,
- Linux interactive nodes, running Red Hat Enterprise Linux 5, and
- Compute nodes, which constitute a parallel-processing supercomputing environment organized around the MPI parallel application library.

Interactive nodes are machines used to run software direct-

ly, much as you would run it on a personal desktop or laptop computer, with two differences. First, they sit behind the sandbox's security perimeter, accessed through a secure Virtual Private Network (VPN). Second, they're multicore, server-grade machines much more powerful than the typical personal computer. Interactive nodes are also used to submit batch processes to the compute nodes.

The HIPAA sandbox compute nodes are powerful machines just like the Linux-based interactive nodes, except that they are networked to operate in concert on parallel processing tasks. Instead of interacting directly with the computer nodes, users submit jobs to be run in batches, as with CHPC's other computing clusters. The compute nodes will soon be expanded to 32 cores, communicating over a high-speed Infiniband network.

The sandbox also includes a multi-terabyte storage system. Regular backups are performed, and the backup media are securely stored, handled only by HIPAA-trained CHPC staff.

As of this writing, the HIPAA sandbox is in its early stages. In addition to the three new compute nodes, new interactive nodes are planned by researchers starting new projects requiring HIPAA compliance. As with CHPC's other resources, researchers are able to buy additional hardware for their particular needs.

What kind of research will be done in the Sandbox?

One type of research is the application of Natural Language Processing (NLP) techniques (see CHPC's spring 2009 newsletter) to understand clinical text. There is a large volume of text created during the course of treatment of patients, including admission notes, nursing notes, surgical narratives, discharge summaries. There is much that could be learned through analysis of this text. For example, by tracking events that occur during the course of a patient's treatment, a computer system might be able to support or warn against subsequent courses of action. Another possibility is tracking medications the patient has taken and raising a red flag against prescribing drugs that interact harmfully with the earlier medicines.

There are several biomedically-oriented text processing packages currently available. One such, MetaMap, is an application available online from the National Library of Medicine. It segments input text into phrases and attempts to categorize them according to standard medical terminology. For example, it can perform such actions as recognizing the phrase "lung cancer" as a neoplastic process.

A common shortcoming among these currently available medical NLP packages is that they expect the text given to them to be well-formed: grammatically correct and with words spelled properly. In practice, this is not always the case. One project currently hosted on the HIPAA sandbox is a system which attempts to address this problem. The project is called "POET" for Parseable Output Extracted from Text. The goal of the project is to develop tools that convert poorly formed text into a more suitable form for input to these medical NLP applications. It consists of several subprojects, including a module to recognize medical acronyms and abbreviations and modules to extract grammatical structure from tabular or list data.

The Sandbox is rapidly acquiring new users from various organizations on campus including the Departments of Biomedical Informatics (DBI) and Radiology, the College of Nursing, and the School of Medicine's FURTHeR project, an infrastructure being built under the large NIH translational research grant called the Clinical and Translational Science Award (CTSA). When CHPC's director, Dr. Julio Facelli, and his colleague at the DBI, Dr. John Hurdle, canvassed PIs of NIH-funded projects at the Health Science Campus last spring, they were able to recruit no fewer than eight PIs who were willing to shift their biomedical computing environment to a homer-like setting. For these researchers, the real appeal of using CHPC as a computing resource is the fact that the CHPC handles systems management issues (e.g., rapid response to electrical power issues, provision of reliable cooling and heating, VPN support for a work-anywhere computing experience, ensuring a highly secure HIPAA environment compared to their office computers or departmental servers, and automatic upgrades of key software) in addition to potential access to high performance computing power.

How to get started using the sandbox?

If you are conducting research that uses data governed by HIPAA privacy rules and believe that CHPC's HIPAA-compliant environment would be useful to you, please contact us to set up a CHPC account.

Permission to use a given dataset is governed by the approval of the University's Institutional Review Board (IRB). Researchers must submit a proposal to the IRB listing the data to be used and the people who will have access to it. If the IRB approves the use of the data in question, the researcher is given an IRB number. In order to store the data in CHPC's HIPAA Sandbox, the researcher must provide CHPC with this number and a list of the users who will be permitted to see the data. Thereafter, the data may be transferred to CHPC and only the IRB-approved users will be able to work with it.

The initial response to the HIPAA-compliant homer environment has exceeded expectations. Homer was designed to grow gracefully, but the CHPC and faculty at the DBI are pursuing ways to secure more hardware to support the growing interest. Policies governing use of homer are expected to evolve as interest grows.

User Services at CHPC

CHPC's primary mission is to provide University of Utah faculty the computing and networking resources they need to achieve their research goals. With nearly a thousand people who depend on these clusters, the User Services group at CHPC is structured to provide immediate help when users have problems either accessing or using the resources. Below is a brief summary of the responsibilities of the members of the group.

Julia Harrison, in addition to her duties as associate director of CHPC, manages the User Services Group. She also serves as an ex officio member of the CHPC Allocation committee, and the Updraft council providing reports and information to inform the committees' decisions. Julia also defines and implements policies and procedures, and oversees account maintenance and access issues.

Walter Scott is CHPC's web programmer. He maintains www.chpc.utah.edu, the site where users will find information about current cluster status, usage information, software documentation, CHPC policies, and all necessary forms. He also maintains Jira, the issue tracking system. When users experience a problem, they can report it to CHPC via email to issues@chpc.utah.edu. Walter also develops in-house tools for internal CHPC processes as needed. Thanks to Walter's efforts, users can now submit allocation requests online.

Martin Čuma consults users on parallel code development. He maintains program development software on CHPC machines including compilers, libraries, and development tools. Several times a year Martin delivers presentations and leads classes that deal with basic Linux OS operation, parallel computing and code development. He also holds a researcher position at the Consortium for Electromagnetic Modeling and Inversion (CEMI) at the Department of Geology and Geophysics.

Wim Cardoen, CHPC's newest scientific staffer, came on board in October to give additional support to the scientific staff. He received an MA in analytical chemistry from the University of Ghent (Belgium), his home-town school and received his Ph.D. from the University of Utah. His graduate research, supervised by Prof. Jack Simons, was the study of highly-correlated systems using density-matrix renormalization group theory (DMRG). After completing two years of post-doc research, first at Cornell University and then at the University of Rhode Island, he returned to the U when his wife accepted a faculty position in the School of Nursing.

Anita Orendt is the expert on the chemistry software packages available on the CHPC clusters. These include Gaussian03/09, NWChem, Gamess, Molpro, Dalton, and Amber, as well as support packages that include GaussView, ECCE, Cambridge Structural Database, and Dock. She advises users on the use of these packages and also gives an overview lecture on their use, as well as a lecture focusing on Gaussian and Gaussview. Anita also collaborates with a number of chemistry groups, including those of Dave Grant, Jon Rainier, Chuck Grissom and Peter Stang. Anita is also an adjunct assistant professor in the Chemistry Department and active in the leadership of the Salt Lake local section of the American Chemical Society.

Sean Igo, who received his MS in Computer Sciences here at the U, is CHPC's expert in Natural Language Processing and is the CHPC liaison with the Biomedical Informatics and other health science users, especially those who need access to the highly secured HIPAA sandbox. Sean's own research is in artificial intelligence and data mining.

Byron Davis is the staff scientist/consultant who helps users with the statistical and methodological aspects of their research. He oversees CHPC's statistical server Turrettarch and the statistical software installed there. He is also an adjunct associate professor in the Department of Family and Consumer Studies where he regularly teaches statistics.

Janet Ellingson staffs the help desk, creates new accounts, helps with grant applications, and moderates issues as users email their problems to CHPC. She also edits the CHPC newsletter. Janet is an adjunct assistant professor in the Department of History and occasionally teaches American history.

User Services Group -- Front row: Sean Igo, Janet Ellingson, Julia Harrison. Back row: Wim Cardoen, Walter Scott, Martin Cuma, Anita Orendt.

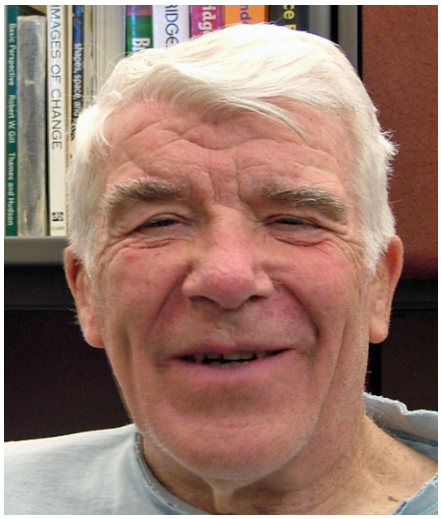


photo by DeeAnn Raynor

Robert McDermott

A Farewell to Robert McDermott

This summer Robert McDermott retired from CHPC after 19 years as the visualization staff scientist. During his time at CHPC he helped researchers create animated visualizations of their scientific data. Robert's own research has focused on the visualization and animation of transforming geometric shapes. Through the years Robert has experienced extraordinary changes in hardware and graphics software. Someday he can tell his grandkids about the time when the only work tools he had were paper and colored pencils. Although retired from CHPC Robert plans to continue developing physical models and running workshops to help teachers present geometric concepts to their students. His collegues at CHPC wish him a long and healthy life and the freedom to continue doing whatever he darn well pleases.

User Services

Video Conferencing

by Jimmy Miklavcic

Video and audio conferencing technology has expanded into a wide spectrum of hardware and software products for laptops and large meeting facility with a cost ranging from free to more than fifty thousand dollars. Here at CHPC we can offer services with systems that span this wide range.

The Access Grid TM (AG) is ideal for individual to large network wide meetings and seminars. It is an extremely scalable system that can be configured on a laptop or on multiple computer nodes for large meetings. In the INSCC Auditorium (RM 110) we have a full AG system for large-scale meetings, presentations and remote seminars. The Access Grid node can now be used with real time High Definition Video.

There are two other active Access Grid systems on campus. The Eccles Library has a full node in the History of Medicine room. Primarily used for regional medical library meetings and educational activities, it is available to others in Health Sciences. The second system is located in the Dibona Center for Educational Technology. This system is part of the College of Humanities distributed education program and is currently being used to teach Hindi to students at the University of Utah and Brigham Young University.

Researchers of Utah Science Technology and Research (USTAR) have received an Integrative Graduate Education and Research Traineeship Program (IGERT) grant requiring the installation of an Access Grid system. CHPC is assisting in designing and implementing a full Access Grid node in the current USTAR building at University Research Park.

The Mobile Access Grid (MAG) Box is available for small informal videoconference sessions.

CHPC also provides a portable video conference system for smaller, more immediate meetings. It is equipped with Access Grid, Polycom PVX (H.323), Skype and other common videoconference software.

CHPC staff is available to assist any department that would like to investigate installing an Access Grid node. For more information you can visit http://www.accessgrid.org or contact Jimmy Miklavcic at jimmy.miklavcic@utah.edu.

photo by Sam Liston

CHPC at SC09

by Sam Liston

CHPC will be hosting a booth again at the upcoming ACM/IEEE SC09 conference that will be held in Portland, OR. The booth will highlight CHPC's collaboration with the High Performance Computing Center at Utah State University and other higher education institutions across the state, as well as the scientific work being done on HPC machines at those institutions.

One of our staff members, Tom Ammon, will be participating as a volunteer for SCiNet. SciNet is responsible for setting up, in a week, the most capable network in the world to facilitate all the extreme network needs of the conference and then, a week later, tearing it down again.

Stemming from Tom's efforts with SCiNet, CHPC will be demonstrating several new technologies in our booth. The first is NFSoRDMA (Network File System over Remote Direct Memory Access), which uses RDMA as the remote procedure call (RPC) transport layer. Using NFS with this transport allows improved throughput compared to NFS over standard transport control protocol (TCP).

For the second tech demo, CHPC will partner with Bay Microsystems and the Utah Education Network to demonstrate the performance of Infiniband-attached storage over wide-area networks (IBoIP). IBoIP is an appliance that encapsulates IB traffic in IP packets for transport across a long-haul IP network. On the other side of the IP network, the appliance decapsulates the IB traffic and drops it off on the destination Infiniband network. This IBoIP bridging technology shows great promise for enabling collaboration that demands high-performance storage over WAN distances.

The third demo will explore another approach to extending IB networks over WAN distances. High-performance research networks, such as Internet2 and National Lambda Rail operate optical networks that provide WAN transport facilities that can accommodate non-IP networks. Infiniband range extenders allow IB networks to operate over longer distances than IB standard specifies. These range extenders, connected to long-haul optical networks, make it possible to extend IB fabrics over WAN distances. CHPC, in partnership with researchers from NASA-Ames Laboratory, and other institutions, will demonstrate this kind of long-haul native IB fabric.

Research Assistants

CHPC received support for two undergraduate research students from the Undergraduate Research Opportunities Program. Colin McDermott, a student in Film Studies, and Josh Bross, studying Computer Science and Film are working with researchers Elizabeth Miklavcic, Project Principal Investigator, and Jimmy Miklavcic, faculty sponsor, to develop the Animated Cinematic Display Interface for the Access Grid technology.

This project will enhance the cinematic component of the InterPlay research in telematic, distributed performance. The Animated Cinematic Display Interface will allow control of placement, motion and other attributes of live HD video streams from participating performance sites around the globe, creating a more immersive experience for audiences at all sites.

Colin McDermott and Josh Bross working on the Animated Cinematic Display Interface.

FYI

Published Research Using CHPC Resources

CHPC maintains on its web site a list of publications and presentations that acknowledge the use of CHPC's resources. You can find the current listing at the following address:

http://www.chpc.utah.edu/docs/research/CHPCBibliography.pdf

If you utilize CHPC resources in your research, please include an acknowledgement in your publications and presentations. Also, please send us a copy for our records.

Dedicated Access Time (DAT) Available on Updraft Cluster

On the updraft cluster there are "Dedicated Access Times" or DATs available where you may utilize the full system. The times are set aside three times each month. Two have restricted access, available only to the researchers (labeled "uintah") who were instrumental in funding the cluster but the third is available to the general user community. They usually run from Monday at Noon to Wednesday at Noon (48 hours) on the weeks they occur. The first and third weeks of the month are set aside for the uintah users and the last week of the month is available for general users.

To request access to a DAT, send an email to issues@chpc.utah.edu with the details of what you plan to do with the time. Please keep in mind that you must have sufficient allocation to cover the use of the full machine (over 2000 cores) for 48 hours.

The upcoming DAT schedule is available on the message of the day or "motd" on the updraft cluster. You see this message either when you first login to updraft, or you can issue the command: "cat /etc/motd" at anytime to view it. If you have any questions about the DAT, please contact CHPC.

What is CHPC?

The Center for High Performance Computing provides large-scale computer resources to University faculty and research staff to facilitate their research. CHPC is located in the INSCC building (just north of the Park administration building) and is responsible for the operation, maintenance and upgrade of the facilities housed in INSCC. CHPC also has data centers at the Komas buidling in Research Park and in the Student Service building.

The projects currently supported by CHPC come from a wide array of University disciplines, including Chemistry, Physics, Atmostpheric Sciences, Geology and Geophysics, Biomedical Informatics, and Pharmaceutics, that requiring large capacity computing resources, both for calculating the solutions of large-scale, two and three dimensional problems and for graphical visualization of the results.

If CHPC resources would be of use in your research, please go to our website: www.chpc.utah.edu for more information about our computing clusters and the software we support. CHPC resources are available free of charge to University faculty.

CHPC Staff Directory

Administrative Staff	Title	Phone	Email	Location
Julio Facelli	Director	585-3791	julio.facelli@utah.edu	410 INSCC
Julia D. Harrison	Associate Director	585-3791	julia.harrison@utah.edu	430 INSCC
Guy Adams	Assistant Director, Systems	554-0125	guy.adams@utah.edu	424 INSCC
Joe Breen	Assistant Director, Networking	550-9172	joe.breen@utah.edu	426 INSCC
DeeAnn Raynor	Administrative Officer	581-5253	dee.raynor@utah.edu	412 INSCC
Janet Ellingson	Admin. Program Coordinator & Newsletter Editor	585-3791	janet.ellingson@utah.edu	405 INSCC

Scientific Staff	Expertise	Phone	Email	Location
Wim Cardoen	Scientific Applications	N/A	wim.cardoen@utah.com	420 INSCC
Martin Cuma	Scientific Applications	587-7770	martin.cuma@utah.edu	418 INSCC
Byron L. Davis	Statistics	585-5604	byron.davis@utah.edu	416 INSCC
Julio Facelli	Molecular Sciences	585-3791	julio.facelli@utah.edu	410 INSCC
Sean Igo	Natural Language Processing	N/A	sean.igo@utah.edu	405-16 INCSS
Anita Orendt	Molecular Sciences	231-2762	anita.orendt@utah.edu	422 INSCC
Ron Price	Software Engineer & Grid Architect	560-2305	ron.price@utah.com	405-4 INSCC

Technical Support Staff	Group	Phone	Email	Location
Ty Adams	User Services	N/A	ty.adams@utah.edu	405-18 ISNCC
Irvin Allen	Systems	231-3194	irvin.allen@utah.edu	405-40 INSCC
Tom Ammon	Network	674-9273	tom.ammon@utah.edu	405-22 INSCC
Robert Bolton	Systems	528-8233	robert.bolton@utah.edu	405 -24 INSCC
Wayne Bradford	Systems	243-8655	wayne.bradford@utah.edu	405-41 INSCC
Erik Brown	Systems	824-4996	erik.brown@utah.edu	405-29 INSCC
Steve Harper	Systems	541-3514	s.harper@utah.edu	405-31 INSCC
Brian Haymore	Systems.	558-1150	brian.haymore@utah.edu	428 INSCC
Derick Huth	User Services	N/A	derick.huth@utah.edu	405-19 INSCC
Samuel T. Liston	Systems, Multimedia	232-6932	sam.liston@utah.edu	405-39 INSCC
Jimmy Miklavcic	Multimedia	585-9335	jimmy.miklavcic@utah.edu	296 INSCC
Beth Miklavcic	Multimedia	585-1067	beth.miklavcic@utah.edu	111 INSCC
Victor Morales	User Services	N/A	N/A	405-14 INSCC
David Richardson	Network	550-3788	david.richardson@utah.edu	405-38 INSCC
Walter Scott	User Services	309-0763	walter.scott@utah.edu	405-13 INSCC
Steve Smith	Systems	581-7552	steve.smith@utah.edu	405-25 INSCC
Neal Todd	Systems	201-1761	neal.todd@utah.edu	405-30 INSCC
Alan Wisniewski	Network	580-5835	alan.wisniewski@utah.edu	405-21 INSCC
Paul Vandersteen	User Services	N/A	N/A	405-19 INSCC

The University of Utah seeks to provide equal access to its programs, services, and activities to people with disabilities. Reasonable prior notice is needed to arrange accommodations.

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:

Janet Ellingson
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

(room 405 of the INSCC Building)

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.*"

If you use the NIH portion of Arches (delicatearch, marchingmen or tunnelarch), please add: "partially supported by NIH-NCRR grant # 1S10RR17214."

Please submit copies of dissertations, reports, preprints, 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