Center for High Performance Computing NEWS

THE UNIVERSITY OF UTAH SPRING 2009

Article

CHPC Supports Natural Language Processing (NLP)

by Sean Igo, CHPC

CHPC is pleased to announce the introduction of Natural Language Processing (NLP) support as a new service.

NLP, a branch of Artificial Intelligence (AI), is the practice of using computers to analyze or generate human (natural) language. Our current emphasis is on the analysis of electronic text, rather than on generation or speech.

The earliest work in NLP, in the 1950s, was concerned with automatic translation of foreign language text - mainly Russian - into English. It was, like other efforts in early AI, part of an enthusiastic vision that computers would soon be able to think as humans do. Disappointing results from those first efforts led researchers to realize that natural language understanding is a far more difficult problem than it initially seemed.

Today, most NLP efforts are concerned with much smaller tasks than a wholesale understanding of human language, although they can be considered subtasks which are either directly or indirectly applicable to it at some point in the future.

Typically, NLP applications are organized as a "pipeline" of software tools. The input is raw text, which undergoes some sequence of processes that perform more and more sophisticated linguistic analysis. Generally, NLP pipelines begin with these steps:

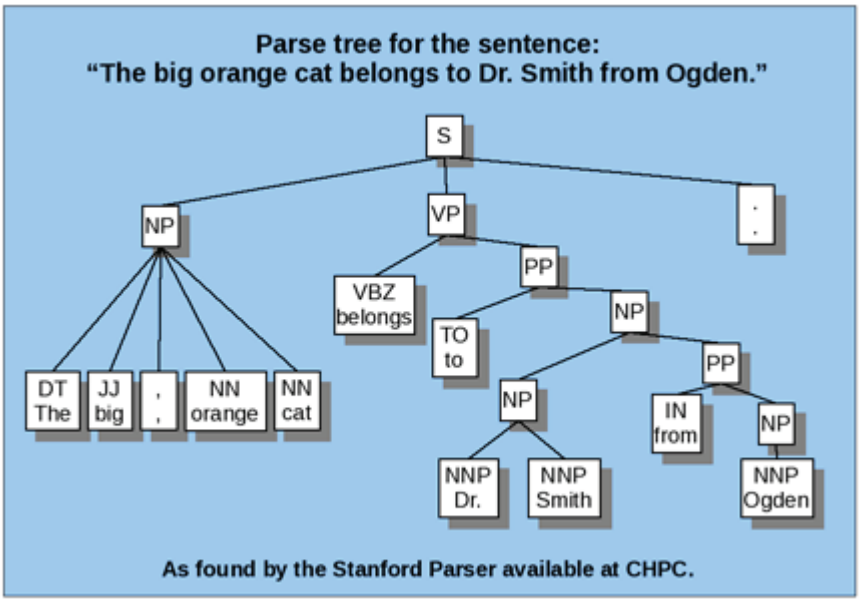
Sentence Splitting - recognizing the boundaries between sentences in a token or character stream.

Tokenization - the division of a stream of characters into "tokens", which may be words, numbers, punctuation symbols, etc. This is harder than it sounds, because some tokens, like "Dr.", include a period, when usually a word followed by a period should be two tokens.

Part-of-Speech (POS) Tagging - assigning to each token in a stream a part of speech such as noun, verb, etc.

Once the input text has been converted into parallel streams of tokens and POS tags, higher-level analysis may be performed.

One such is parsing, or discovery of sentence structure. Parsing can be considered either high or low level NLP, as there are different theories about how sentence structure should be represented. When sophisticated structure is sought -- the "full parsing" approach -- parsing is often a research end in itself. However, many NLP systems perform "shallow parsing", using a simpler view of grammar, because it tends to be faster and more robust in the face of badly-written text than full parsing is. After a sufficiently detailed understanding of a text's structure and meaning are achieved, there are many things that can then be done with it. Here I will describe three popular areas of research and application:



1. Question Answering (QA) - QA systems allow a user to submit a question in natural language form and attempt to find the answer. Their pool of knowledge may be limited to a certain domain, such as geography, or unlimited (e.g. systems that use the Web as a database.)

2. Information Extraction (IE) - similar to QA, IE systems are more specialized and attempt to find predefined kinds of information in a given corpus. For instance, an IE system may examine a set of business magazine articles to find mentions of corporate mergers, then extract for each

such mention the names of the companies involved, sums of money or stock transferred and to whom, and so on.

3. Machine Translation (MT) - In many ways the "holy grail" of NLP, MT is the automatic translation of text from one human language to another. Great strides have been made in recent years, resulting in freely-available, reasonably passable translators such as Google's language tools as well as high-quality professional tools for translating materials for international use.

Machine Learning:
Many modern NLP techniques involve Machine Learning (ML). ML is a collection of techniques designed for computers to learn by example, as opposed to earlier AI methods involving knowledge engineering conducted entirely "by hand". Often, examples are provided by human experts; this is called "supervised learning". Supervision can be "weak" or "strong" - depending on the degree of expertise necessary to prepare the training examples and / or the volume of examples needed for a learned model to work well. For the last fifteen years or so, ML approaches have been dominant in NLP research, now that sufficient volumes of training data and the higher computer power necessary are available.

Who should be interested in NLP:
If your research could benefit from discovering any kind of information from human-language input, NLP could be useful to you. Researchers at the U have used NLP techniques to extract descriptions of terrorist activity and disease outbreaks from newspaper articles and other documents, attempt predictions of adverse drug interactions from clinical notes, and analyze linguistic typology.

Advantages of CHPC Resources for NLP:
CHPC's computing clusters and large scratch storage capacity are ideal for dividing large corpus-based NLP jobs into manageable segments. This can either be done through simple allocation of a large number of independent tasks among several compute nodes or through development of custom software to parallelize the algorithms.

NLP / ML Software at CHPC:
Currently, the installed base of NLP software at CHPC is not extensive, since we are just getting started with it. The packages currently available are:

Berkeley Parser
Stanford Tools: Parser, Tagger, NER
Weka: General multi-algorithm machine-learning package
SVM-light: Highly optimized Support Vector Machine software
Metamap: Parsing / semantic tools for biomedical NLP.

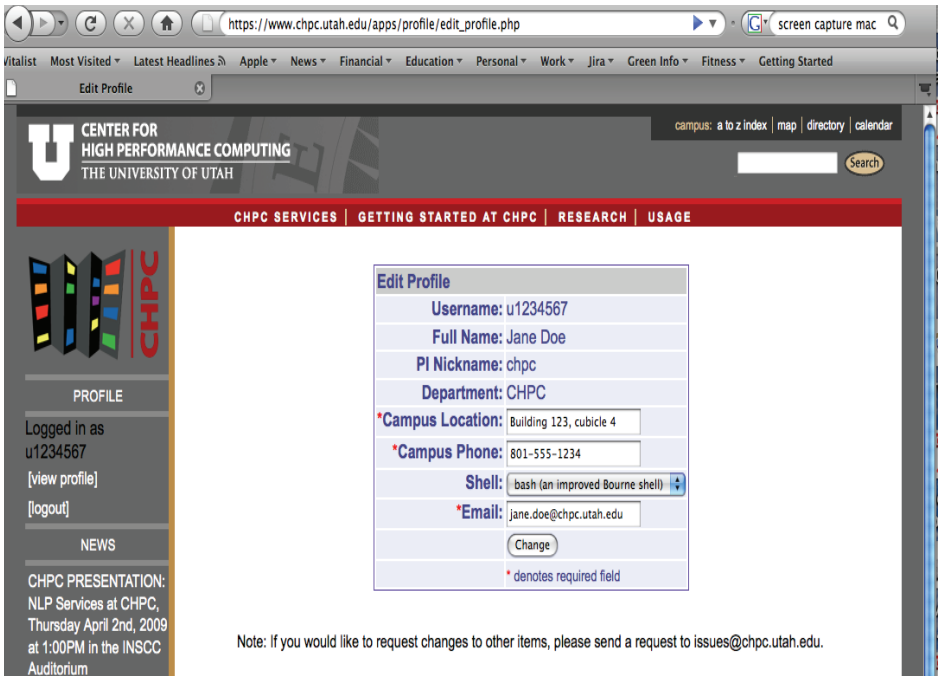
Support for NLP Research:
If you are interested in using CHPC resources for NLP, please contact Sean Igo at Sean.Igo@utah.edu.

FYI Editing Your CHPC Profile

Did you know that now you can edit your chpc profile, including changing your default unix shell on our systems? You start by going to our home page at http://www.chpc.utah.edu. Here are the simple steps:

1. Click on the left side of the screen where it shows "Profile," just under the CHPC logo. This will expand the area and allow you to login.
2. Enter your User Name (uNID) and password (CIS) and click on the "login" button. This will display your profile information in a Personal Details display box in the middle of the screen.
3. To make changes, click on "edit profile" at the top of your Personal Details display box. This will allow you to update your campus location, campus phone number, shell and email address.
4. Make your desired changes and click on the "Change" button.
5. You can click on "view profile" to see that the changes are saved.
6. Click on "logout".

For changes to any other information, please send a request as usual to issues@chpc.utah.edu.



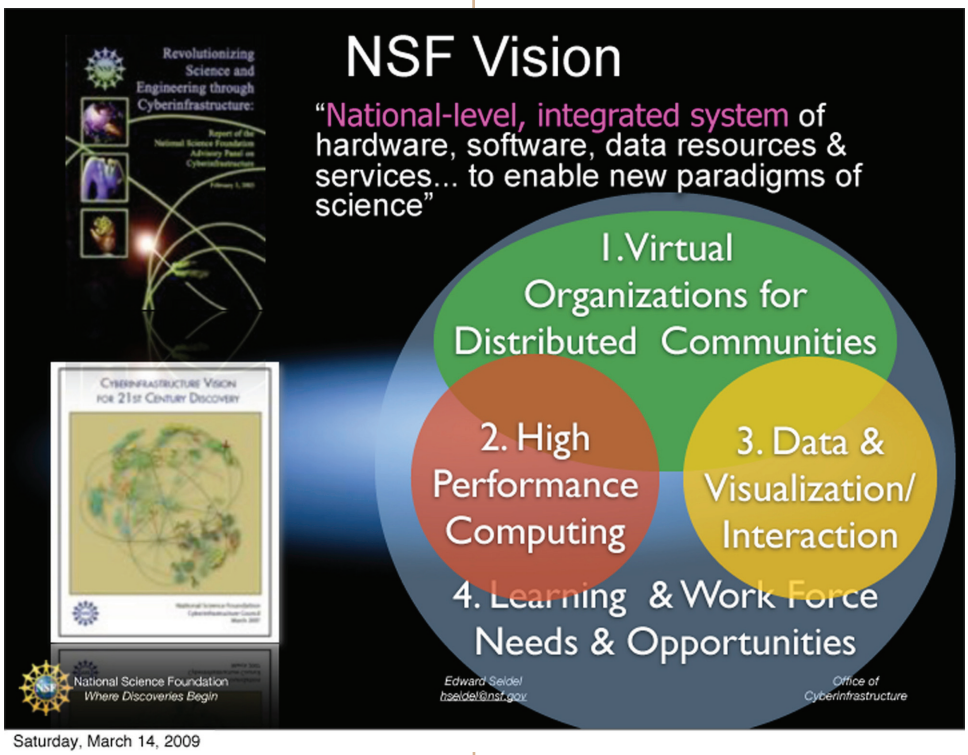
Article

Campus CI Day - Enabling Research in a Data Driven World

By Janet Ellingson, CHPC

The University of Utah held its first Campus Cyberinfrastructure Day on March 13th. Sponsored by the Office of Information Technology, including support from CHPC, the series of discussions began with the keynote speaker, Prof. Edward Seidel, the director of the National Science Foundation's Office of Cyberinfrastructure. He is well qualified to address the theme of this year's gathering: enabling research in a data driven world. To highlight the enormous change in scientific research over the past two decades, Prof. Seidel contrasted Stephen Hawking's work with black holes and that done by the scientists working with the Large Hadron Collider (LHC). Hawking's research group consisted primarily of himself and a few graduate students. Seidel estimates that they generated 50 GB of data. The research community working with the LHC is composed of thousands of scientists in over 33 countries who will collaboratively generate 25 petabytes of data a year (a petabyte is 10^{15} bytes). The overarching issue that must be addressed today is how to handle the enormous amount of data produced by scientific groups in a way that allows continuing collaborative research. Data must be shared. This requires that data be generated, stored, transported, accessed and visualized with common tools and software. The NSF is committed to facilitating a national and global cyberinfrastructure that will make possible the collaborative research that will answer the great questions of our century and solve our complex problems.

The National Science Foundation envisions "a National-level, integrated system of hardware, software, data resources and services…to enable new paradigms of science." [See the NSF publication "Cyberinfrastructre Vision for 21st Century Discovery" found at http://www.nsf.gov/pubs/2007/nsf0728/] Prof. Seidel identified four areas of focus that will move us toward the end goal:



Saturday, March 14, 2009

1) Virtual Organizations. The NSF defines a virtual organization as a coherent community of researchers from disparate fields who may not even know each other but who are working on a common problem. Prof. Seidel used as an example the various agencies and scientists concerned with hurricane prediction and disaster response during hurricane Katrina. The community that needed to share data included scientific research groups, engineers, and social service providers. This experience forcefully demonstrated the need for more coherency and improved communication. NSF is sponsoring studies that address the elements of effective virtual organizations and how such groups enhance research and innovation.

2) Computer environments. Prof. Seidel noted that cores are not getting significantly faster, we are simply making more of them. Parallelism, therefore, must be addressed at the desktop level and this provides an incentive for the creation of scalable applications that will also work at the national and supercomputing level. "There is no free lunch for traditional software. Without highly concurrent software, it won't get any faster," he said. In response to a question from the audience, Prof. Seidel stated that the recent federal stimulus money included an additional $300 million to fund grants in the Major Research Instrumentation (MRI) program. Although this funding is a one-time event, it will finance considerable improvements in the entire spectrum of computer environments.

3) Data handling. Perhaps the biggest challenge in this new research environment is the handling of the enormous quantities of data. Moving, managing, analyzing, mining, and storing data on the petabyte scale will require new tools and services. The NSF program DataNET has begun to address these problems. DataNet, a $100 million 5-year program based at the University of New Mexico and Johns Hopkins University, has the goal of creating a "sustainable [over 50+ years] digital data preservation and access network partnership" that includes the creation of a

"new generation of tools and services facilitating data acquisition, mining, integration, analysis, [and] visualization." The program will provide resources for replicating the infrastructure nationwide.

4) A new workforce. The training of people to work in the new data-driven environment is essential. Beginning as early as high school our educational system must prepare a new workforce. In response to a question from the audience, Prof. Seidel also acknowledged that changes may need to occur in the tenure process that will encourage the development of faculty who share data. Perhaps credit can be given for the publication of data sets just as credit is given for journal articles. For a detail description of the NSF goals in this area see "Fostering Learning in the Networked World," at http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf08204.

Prof. Seidel praised the University of Utah's efforts to meet the challenges of the new data-driven research environment. CHPC supports many collaborative research projects and we are using our experience and expertise to help the University of Utah's Campus Cyberinfrastructure Council meet the challenge to "rethink our campus infrastructure and how they can support new modalities of research, collaboration, education."

In addition to Prof. Seidel's address, the day included presentations by University faculty Valerio Pascucci, Juliana Freire, and Lewis Frey, Professor William Michener from the Univ. of New Mexico, and Dr. Sayeed Choudhury from Johns Hopkins Libraries. For the archived streaming video of the presentations, please go to http://www.it.utah.edu/leadership/research/ci_event.html.

User Services

Tips and Tricks for Updraft

by Martin Cuma, Ph.D., CHPC

Last fall, CHPC took delivery of a new cluster called Updraft. It has 256 nodes connected with Qlogic InfiniBand network. Each node has two quad-core Intel Xeon 2.8 GHz processors and 16 GB of RAM. There is also a dedicated NFS file server with a total of 16 TB of disk space. While similar to our other clusters setups, there are some important specifics that we detail in this article.

Updraft has been bought partly with funds from the Institute for Clean and Secure Energy (ICSE), and as such, the cluster CPU allocation and disk space are divided between the ICSE group (uintah allocation) and general use (general allocation). The Uintah group gets 2/3 of the nodes and 12 TB of disk space at /scratch/uintah, the general allocation gets 1/3 of the nodes and 4 TB in /scratch/general. The Uintah group members get assigned to Quality of Service (QOS) qos=uintah, other users get qos=general. Priorities in these two qos's are the same, except that one qos can't run on the nodes allocated to the other qos. To facilitate possibility of running larger jobs we implemented special qos called "bigrun", which allows to use up to 256 nodes and is specified in the PBS job script as:

```
#PBS -l nodes=256:ppn=8,walltime=12:00:00,qos=bigrun
```

Users need to contact CHPC to be added to qos=bigrun.

Users can also request Dedicated Access Time (DAT) that grants exclusive use of the cluster. DATs are set for 48 hours (Monday 12pm to Wednesday 12pm), 2 times per month (the 1st and the 3rd weeks of each month) for Uintah users and the last week of the month for General users. Those interested in DAT need to contact CHPC in advance at issues@chpc.utah.edu.

A new feature that we implemented on the Updraft cluster is preemption of jobs. That is, a running job can be terminated if a job with higher priority is queued up. All freecycle jobs are preemptable, that is jobs that don't have any allocation (recall that CHPC allows users without CPU allocation to run at a very low priority in the freecycle mode). Note that a preempted job is not requeued – that is, user must manually resubmit a job that was preempted. Users with allocation can also specify their job to be preemptable. This will keep the job's priority high (so it'll run sooner), but, it can be preempted by another job. To specify job as preemptable, add qos=preemptable as:

```
#PBS -l nodes=16:ppn=8,walltime=12:00:00,qos=preemptable
```

The advantage of this setup is that preemptable job will get charged only ¼ of the allocation that it uses. This option is beneficial for users whose jobs are restartable and want to use as little allocation as possible. Some of our users are successfully using this option, having scripts or cron jobs automatically monitoring progress of their jobs and resubmitting them as necessary.

From the hardware standpoint, Updraft uses Intel Xeon CPUs, as opposed to AMD Opterons on our older clusters. From our extensive testing, Intel compilers produce the fastest code on the Intel CPUs, thus, we recommend using Intel compilers to build your codes. For details, see Updraft user's guide at http://www.chpc.utah.edu/docs/manuals/user_guides/updraft/.

While most of the application stack on Updraft is the same as on all other CHPC clusters, one difference is the MPI.

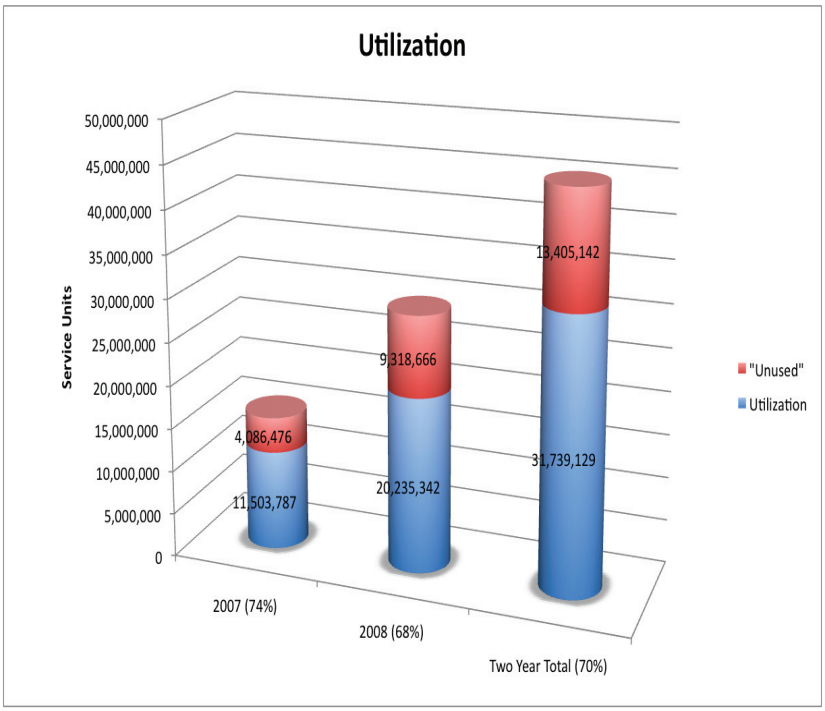
(continued on Page 6)

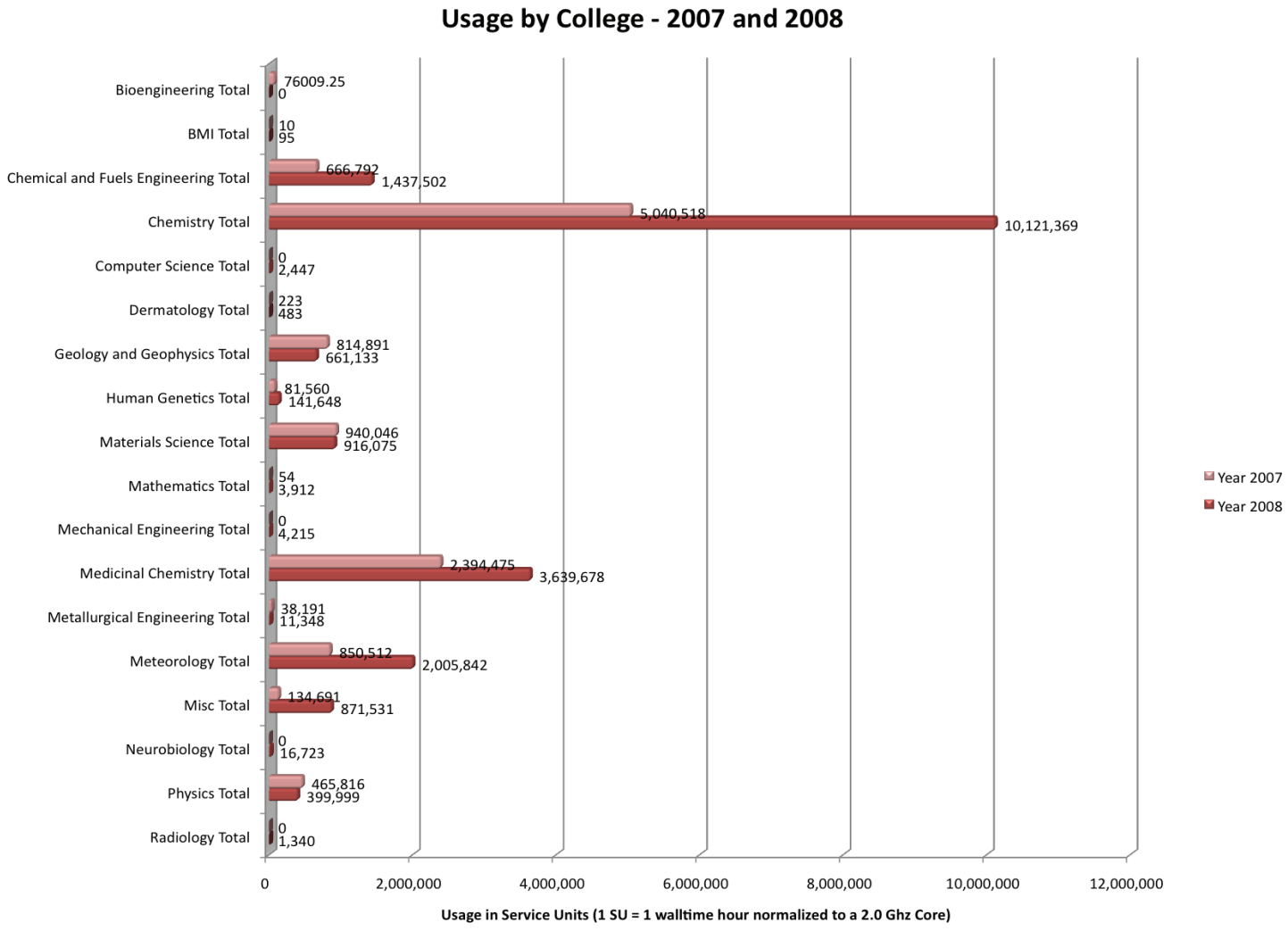
Staff Research

Utilization of CHPC's HPC Systems

by Julia Harrison, Associate Director, CHPC

CHPC HPC resources have nearly doubled over the past two years. Here are two different views of the HPC Usage Data from 2007 and 2008. This is only usage for the large, computational clusters, and does not represent utilization on smaller servers or desktop systems. The table below is by department and table to the right shows the usage in relation to the available cycles. All usage numbers are in Service Units, where 1 service unit (or SU) is equal to one wallclock hour per core, weighted relative to a 2.0 Ghz core.





"Unstill Life" by Beth Miklavcic

Beth A. Miklavcic, a member of CHPC's multimedia and vizualization team, has created "E-scape" canvases that are designed so viewers can escape into a multipoint view of different worlds, as well as pull back and view the complete composition of the surreal landscape. Using the Maya canvas and paintbrushes, she began to experiment with creating 2D visual images within a 3D program. Using multiple layers of imagery she has created numerous visual stimuli throughout the canvas grid. Although many of the compositions she has created are rich with content, some of her works contain a single item, which can appear to be as complex as the multi-layered canvases. For further information about Beth's work email beth.miklavcic@utah.edu

FYI

Published Research Using CHPC Resources

CHPC maintains on its web site a listing of publications and talks 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 give us a copy for our records.

UPDRAFT tips and tricks (cont.)

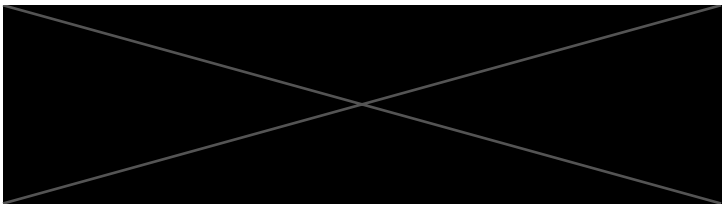
We use the Qlogic InfiniPath MPI binaries, that are installed as a part of the system, so, they reside in the standard paths (/usr/bin, /usr/lib64, etc.). InfiniPath was historically associated with Pathscale, which also makes compilers, so, by default, InfiniPath MPI tries to use Pathscale compilers.

To use Intel compilers instead, there are flags for each MPI compiler wrapper that let one specify preferred compiler. For example, mpicc -cc=icc. See Updraft user's guide for more details on this.

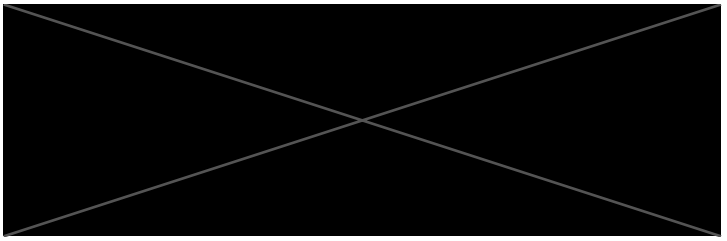
We are excited to have this new computational resource available for users and hope that this article both helps newcomers to come up to speed faster and experienced users to become more familiar with the specifics of the Updraft. If you have any questions on Updraft or anything else related to CHPC, please, contact us at issues@chpc.utah.edu.

CHPC Staff Directory

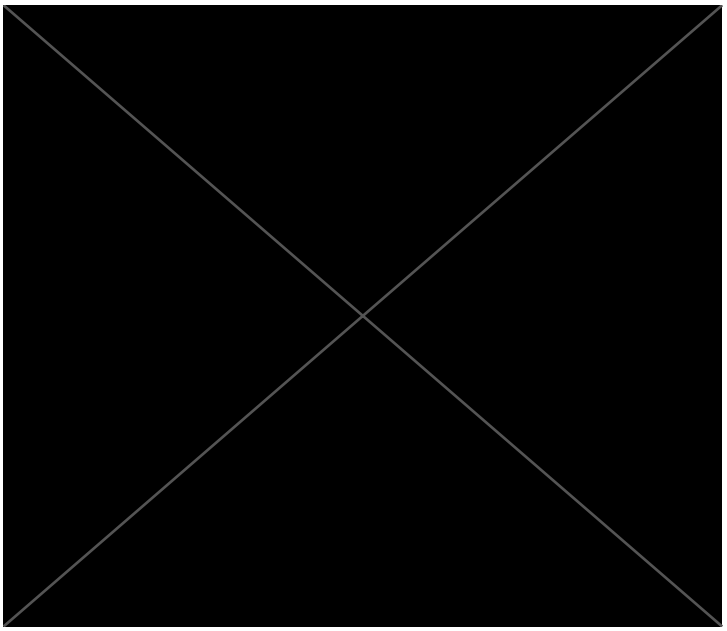
Administrative Staff	Title	Phone	Email	Location
Julio Facelli	Director			
Julia D. Harrison	Associate Director			
Guy Adams	Assistant Director, Systems			
Joe Breen	Assistant Director, Networking			
DeeAnn Raynor	Administrative Officer			
Janet Ellingson	Admin. Program Coordinator & Newsletter Editor			



Scientific Staff	Expertise	Phone	Email	Location
Martin Cuma	Scientific Applications			
Byron L. Davis	Statistics			
Julio Facelli	Molecular Sciences			
Sean Igo	Natural Language Processing			
Robert McDermott	Visualization			
Anita Orendt	Molecular Sciences			
Ron Price	Software Engineer & Grid Architect			



Technical Support Staff	Group	Phone	Email	Location
Ty Adams	User Services			
Irvin Allen	Systems			
Thomas Ammon	Network			
Robert Bolton	Systems			
Wayne Bradford	Systems			
Erik Brown	Systems			
Steve Harper	Systems			
Brian Haymore	Systems.			
Derick Huth	User Services			
Samuel T. Liston	Systems, Multimedia			
Jimmy Miklavcic	Multimedia			
Beth Miklavcic	Multimedia			
Victor Morales	User Services			
David Richardson	Network			
Walter Scott	User Services			
Steve Smith	Systems			
Neal Todd	Systems			
Alan Wisniewski	Network			
Paul Vandersteen	User Services			



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:

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