CENTER FOR HIGH THROUGHPUT COMPUTING

Engagement Report

July 1, 2011 – June 30, 2012

Brooklin Gore

Introduction

The Center for High Throughput Computing (CHTC), established in August 2006, provides computing infrastructure, operations, middleware and consulting to advance the research of UW-Madison faculty and external collaborators. Condor, the distributed high throughput computing software developed by Prof. Miron Livny and his team at the UW-Madison, matches the resources in the CHTC to the differing computing needs of the campus. For the past six years, the CHTC team has been actively engaging with campus researchers to help accelerate their science through the use of the CHTC's capabilities.

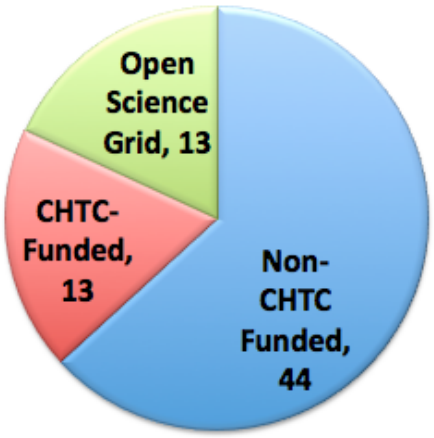
This report includes over 100 groups that we have worked with during the reporting period to better understand the impact of the CHTC on their research and to learn where we can apply future resources to best meet the evolving needs of the campus. Results are drawn from internal accounting data and survey responses from 32% of those groups. Internal knowledge from our research computing facilitators was used for non-responding groups. The report methodology is detailed in the appendices. The report includes the following sections:

- **Resources.** Provides an overview of resources offered by the CHTC.
- **Usage.** Highlights how the campus is using the CHTC year over year.
- **Impact.** Shows how we impact scientific research.
- **Findings.** Highlights key take-aways from the report, and proposed areas of future work.
- **Appendices.** Provides the methodology for the report and comments from survey responders on what is working and what we can improve.

Resources

Condor's power lies in its ability to effectively manage large collections of computing resources across many administrative domains, or pools, and present them to the end user as a single collection of managed resources. The CHTC is comprised of a number of federated resource collections, unified by Condor, consisting of multi-core Linux-based computers with a small percentage of Windows machines. The largest of these collections are included in Figure 1 and are comprised of resources purchased by the CHTC as well as those purchased by other university groups and managed by the CHTC. Prof. Livny is also the Technical Director for the Open Science Grid (OSG). In this reporting period we significantly increased the number of OSG cycles delivered to campus researchers. **The CHTC provided 70 million hours of computing to campus researchers and their collaborators in the 12-month period.**

Fig.1: Major CHTC Resources
(Million Hours)



Usage

During the reporting period from July 1, 2011 through June 30, 2012, **106 research projects in 52 departments on the UW-Madison campus used CHTC resources.** We increased the number of hours, research projects and departments served over the previous reporting period as shown in the CHTC Quick Facts to the right. We sustained resource-sharing arrangements with campus departments and collaborated with 10 U.S. institutions and 3 international ones. Table 1 includes the 25 most active groups (in terms of computing hours) in order of decreasing consumption that is presented in million core hours[1]. The top 25 users accounted for 65.2 million hours, 93% of hours consumed.

Jun'10-Jul'11	Jun'11-Jul'12	CHTC Quick Facts
45	**70**	Million Hours Served
54	**106**	Research Projects
35	**52**	Departments
10	**13**	Off-Campus

Top 25 Groups by Usage	
22.5	**LHC Compact Muon Solenoid (CMS)**
9.0	**Chemical Engineering - Nanomaterials & Molecular Models**
6.5	**Laboratory for Molecular and Computational Genomics**
4.6	**Biological Magnetic Resonance Data Bank**
4.4	**Ice Cube Neutrino Observatory**
3.9	**LHC Atlas**
2.7	**Physics - Magnetic properties of Heisenberg-Kitaev model**
1.5	**Computer Science Department**
1.0	**Computer Science - Deep Linguistic Processing**
1.0	**Economics - Understanding markets through auctions**
0.9	**Berkely Open Infrastructure for Network Computing (BOINC)***
0.7	**Electrical & Computer Engineering - Brain Connectivity**
0.7	**Technion - Israel Institute of Technology**
0.6	**Chemistry - Applications using CHARM and GROMACS**
0.6	**Center for High Throughput Computing - Testing**
0.6	**Chemistry - Polymers**
0.5	**Engine Research Center - Engine Efficiency/Performance**
0.5	**Statistics - Penalized likelihood research**
0.5	**Physics - Phenomenology**
0.5	**Physics - Quantum Algorithms**
0.4	**Biochemistry - Diabetes**
0.4	**Chemistry - Computational Chemistry**
0.4	**Botany - Root Growth Genetics**
0.4	**Chemistry Department**
0.4	**Morgridge Institute for Research - Regenerative Biology**

* Primarily jobs for the Einstein@Home project (http://en.wikipedia.org/wiki/Einstein@Home)

Table 1: CHTC Usage in Million Core Hours

[1] A core hour is defined as a single core of a multi-core processor running for 1 hour.

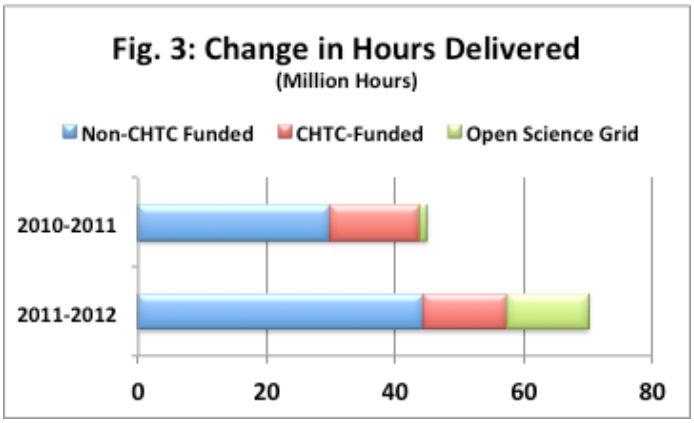
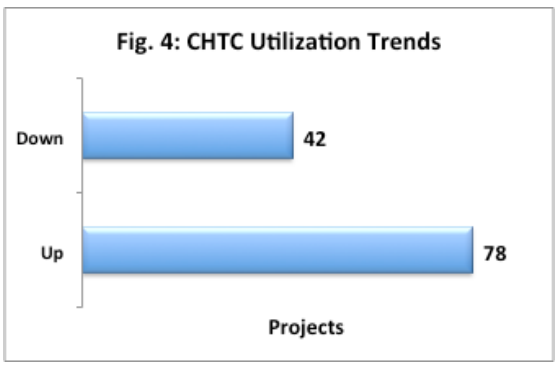


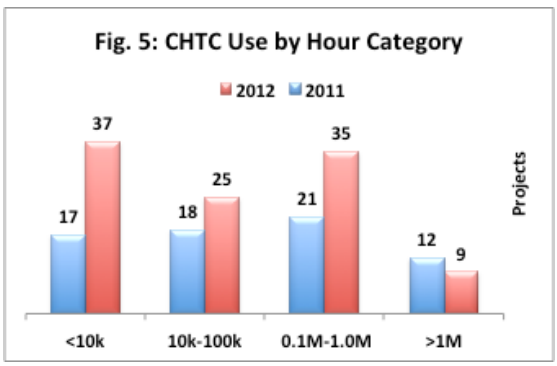
Figure 3 shows that **the CHTC delivered 55% more hours** to researchers in the current reporting period. It also shows that we more effectively delivered cycles from the Open Science Grid. A comparison of top users in the current and previous reporting periods, presented in Table 2, indicates a few interesting trends. Both Large Hadron Collider (LHC) experiments (Atlas and CMS) approximately tripled their use as the search for the Higgs Boson neared discovery in early 2011. Every group in the top ten for the previous period increased their utilization. Also encouraging is the fact that three researchers in Physics, Computer Science and Economics were not even users of the CHTC in the previous period!

	Rank		Hours	
Project	**2011**	**2012**	**2011**	**2012**
LHC - Compact Muon Solenoid (CMS)	2	1	8,291,696	22,520,074
Chemical Engineering	1	2	8,569,501	9,031,699
Lab for Molecular and Computational Genomics	3	3	4,425,542	6,478,077
Biological Magnetic Resonance Databank	25	4	244,236	4,633,386
IceCube Neutrino Observatory	5	5	2,686,398	4,414,772
LHC - Atlas	7	6	1,771,244	3,873,554
Physics – Magnetic Property Simulations	N/A	7	0	2,650,640
Computer Science Department	13	8	822,831	1,514,343
Computer Science - Deep Linguistic Processing	N/A	9	0	1,047,898
Economics - Analysis of Auctions	N/A	10	0	953,192

Table 2: CHTC Top Users by Reporting Period

Figure 4 shows that **utilization is trending up for 65%** of 120 users active in the past two years**.** We see from Figure 5 that for the current period, fewer groups are using over 1 million hours, while **a much larger group drives overall usage, comprising over 55% of users** between 10,000 and one million hours. Finally, we are seeing a much larger group in the under 10,000-hour category possibly indicating a larger number of new users over the previous period.

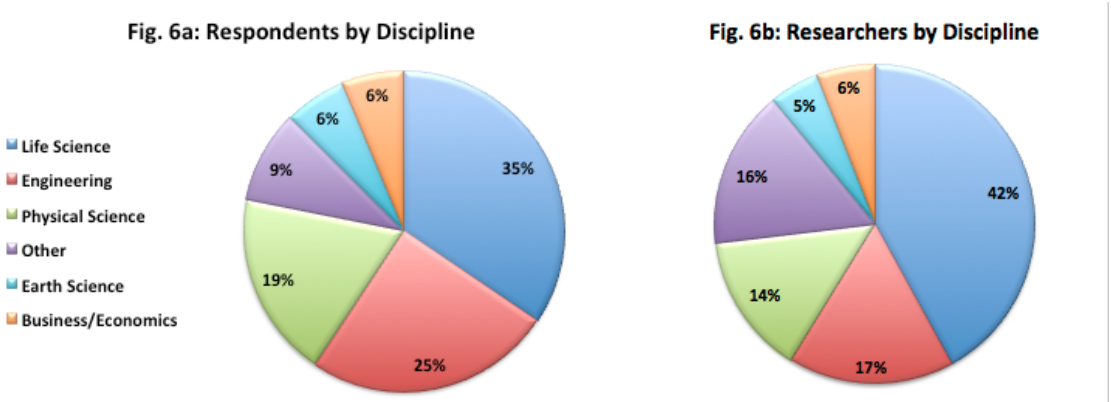




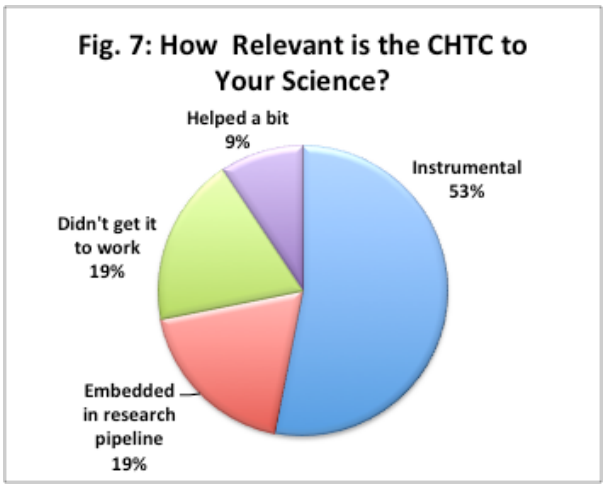
Impact on Science

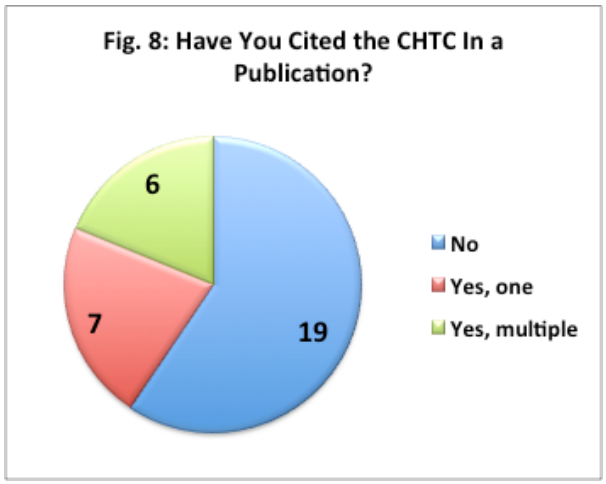
As mentioned in the previous section, the CHTC helped analyze data from two Large Hadron Collider experiments: Atlas and the Compact Muon Solenoid (CMS) in their search for the Higgs Boson. Collectively, this work consumed over 26 million hours delivered by the CHTC – about 37%! In life science, the Biological Magnetic Resonance Databank expanded their use of the CHTC for protein-folding simulations using the CS-Rosetta program. Another life science user, the Laboratory for Molecular and Computational Genomics used the CHTC for optical mapping of whole genomes. A clinical health example of our work is Tyler Churchill's use of not only campus but also OSG resources to improve the performance and quality of cochlear ear implants.

To gain additional insights into the CHTC's impact on science during the reporting period, we conducted a web-based survey. Of 99 projects surveyed, 32 responded. Figure 6a shows how these respondents represent the various science communities. Figure 6b shows the distribution across science disciplines for 119 researchers who have used the CHTC in the past two years.

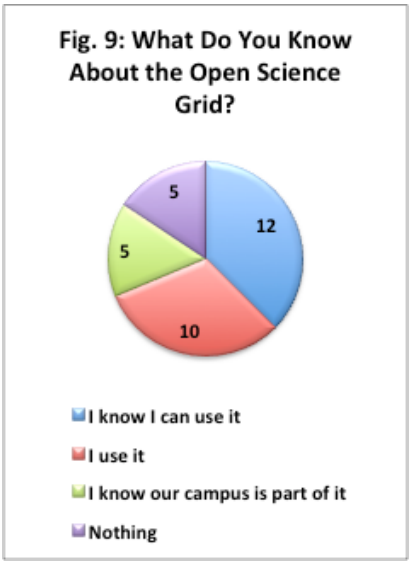


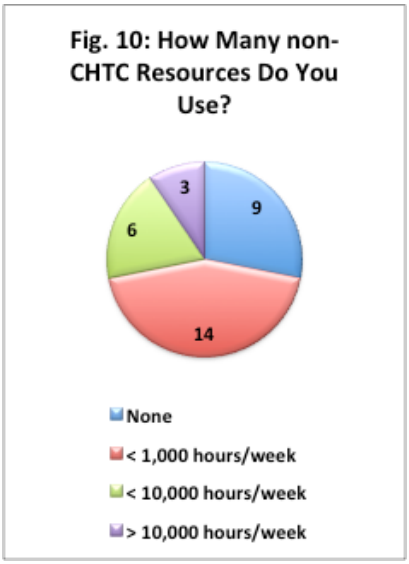
We see from Figure 7 that for almost two thirds of respondents that the CHTC is instrumental to their science or is embedded into a research pipeline. We also note there is still work to do to better help 19% of respondents. Appendix A provides additional insights into what we can do to better help these researchers. Figure 8 indicates that **access to high throughput computing had a significant enough impact on science to be cited in a publication for over 40% of respondents**.





We also wanted to gauge how well campus researchers where aware of our relationship with the Open Science Grid, which is providing a growing number of compute resources, and to understand if researchers who used the CHTC's high throughput computing resources used other non-CHTC provided resources. Finally, we wanted to understand if the researchers using our facility were aware of our research computing facilitators who can help get applications, algorithms and data analysis running in a high throughput computing environment.





Findings

In our 2nd annual report, we are just beginning to be able to compare results year over year. We are collecting better data and learning what questions to ask to more effectively judge our impact on science. It is rewarding to see the growth in hours delivered to the research community and our use by a growing number of researchers, departments and external collaborators. Clearly, the CHTC is helping researchers on the UW-Madison campus and their external collaborators do better science. This is evident from the researcher comments included in Appendix A, our internal data and survey results. For example:

- Our reach increased: We delivered 55% more compute hours, we almost doubled the number of researchers using our facilities and serviced 48% more departments on campus and 30% more external collaborators. While not scientifically correlated to our growth, we did hold two outreach meetings for the period in Oct, 2011 and Feb. 2012.
- While 14 researchers stopped using our facilities, we added 38 new researchers. We emailed the 14 who stopped using us to find out why, but did not hear back from any of them.
- We are the primary computing resource for 72% of the researchers using our facility.
- More researchers are aware of (84%) and using (31%) resources provided from the Open Science Grid with overall usage up from 1 to 13 million hours in this reporting period.
- A higher number of survey respondents (19%) than last period (14%) were NOT able to use the CHTC's resources. Appendix B provides some insights into why. The primary reasons cited where lack of large memory and tightly coupled systems for parallel processing. As of this writing, we are adding larger memory systems to the CHTC and working with campus to provide shared high performance (parallel) computing (HPC) resources to campus researchers.

Methodology

In the previous reporting period, we surveyed our community via email, while this year we used a web survey. The questioning was a bit different, as we wanted to get more insights into the true impact on science, for example if the CHTC was cited in published papers. Not necessarily by design, we got less specific information on why people were not able to use our resources – there was only an open-ended question on what we could do to improve. The survey included the following 10 questions:

1. **How would you categorize your science?** (Business/Econ; Earth Science; Engineering; Life Science; Physical Science; Other-specify)
2. **How would you classify your use of CHTC resources?** (Not a fit for my work; Tried but never got it to work; It helped advance my science a bit; It is instrumental to my science; It is embedded in my research pipeline)
3. **If you use computing resources in addition to the CHTC (and OSG), what is your estimated weekly usage?** (I only use CHTC resources; Less than 1,000 hours a week; Less than 10,000 hours a week; More than 10,000 hours a week)
4. **What do you know about the Open Science Grid (OSG)?** (Nothing; I've heard of it; I know our campus is part of it; I know I can get compute cycles from it; I use it to augment my local computing needs)
5. **How would you classify your experience with members of the CHTC team?** (I didn't know we had such a team; I didn't need help from them; I tried to use them but had some trouble; They were helpful; They were indispensible)
6. **Did you cite the use of the CHTC in a publication?** (No; Yes, one; Yes, multiple)
7. **What are we doing right?** (free text entry)
8. **How can we better support the computing needs of your research?** (free text entry)
9. **Do you have any other comments or questions you'd like to share with us?** (free text entry)
10. **If you wish to provide your email address, we can follow-up with any questions or concerns you may have.** (free text entry)

We used a free survey service from surveymonkey.com that was limited to 10 questions and survey complexity. We have an on-campus survey tool that we plan to use next year to get more information on why people have trouble using our resources. We will also collect more demographic information and not make contact information optional which makes it hard for us to follow-up with respondents regarding issues they reported.

In addition to the survey, we perform daily usage reporting by user and group which enables us to provide usage growth statistics and classify users by type, for example, department, on campus, off campus, etc. As we make reporting an annual activity, we also are improving our ability to compare results year to year

Appendix B – Full Text Responses

Responses to the survey question: “How can we better support the computing needs of your research?”

1. Include shared-memory machines with RAM on the order of several hundred Gigabytes and several dozen cores. Serve licenses for fee-based software (e.g. Comsol, Abaqus) along the lines of what has been done for Matlab.
2. Need to do Monte Carlo simulation better
3. Provide documentation material to help other institutions/groups to emulate your operations.
4. The output of my simulations require more than a few GB of space, so I can only run a few of my simulations at the same time. Being able to use Gnuplot would be of great help for me. Thanks.
5. We desperately need infrastructure for constructing web interfaces for biologists to have access to the high-volume data from our projects. We currently spend lots of money outsourcing this costly activity.
6. Not sure
7. A lot of my runs are in the ~1000 hours category. I wish I could more directly interface my local development and computing environment with the larger OSG and CHTC compute resources. I'm currently using a local Condor pool for this, but if I could double or triple its size (from 26 cores to ~100) that could help increase my iteration time.
8. 1. storage space 2. backups 3. CPUs with higher memory
9. Doing an great job!
10. Much easier install and configuration (currently needs sysadmin + interactions with CHTC to get things up and running)
11. Better documentation for how to operate automated system for running jobs would reduce staff inquiries and free up time for other tasks.
12. I would like to run a model made in C through a MATLAB script (tens of thousands of times). I think this will require the ability to compile mex files with chtc_mcc, a capability not yet available.
13. Making Condor more friendly to distributed data processing. I hope data on working nodes can be persistent and don't need to be transferred every time.
14. This is just a thought off the top of my head but.... perhaps (??) some support personnel that come from domain science backgrounds, but with skill sets that are closer to computer science. Such individuals would have some familiarity with both the culture and science of their domain, and can help integrate your capabilities from that viewpoint.
15. Better documentation of basic condor usage.
16. Limit to one day per job seems a little strict. Perhaps, for users that do not overtax the system a little longer time can be allowed.
17. Everything worked out as needed.
18. Keep doing what you are doing.
19. It is very good already!
20. It'd be nice to have a easy-to-access link on the CAE or COE website for information regarding: current status, FAQ, Q/A board, and so on.
21. Our jobs require parallelization. It seems as though CHTC is better for many serial jobs. An Intel compiler might have made our program compile more easily, but even so we probably would not have used CHTC that much. Condor and the wrapper scripts for submission, and specifying the input files to send, seemed clumsy compared to PBS.
22. In order of priority: 1) Increasing "reliability" of jobs. Many jobs (even < 24 hours) get interrupted / restarted without warning, causing some frustration. 2) Increased flexibility in job time limits, e.g. allowing jobs > 24 hours. 3) Increased flexibility / availability for multi-core jobs. Obviously concerns 2-3 highlight on the need for campus "high performance" computing resources to supplement true "high throughput" computing.

Responses to survey question: What are doing right?"

1. Supporting users. Serving licenses.
2. Have lots of computing resources
3. Your availability and willingness to engage is right on target
4. Having CHTC in the university and available to grad students is great and very helpful. Thanks.
5. You've been very responsive to my questions and helping me get our systems off the ground at our facility. Note that I am not currently performing work at CHTC or OSG but using Condor locally with the help and guidance of CHTC.
6. Customer focus, rapid response
7. Condor itself is a great product. Simple, free, useful. CHTC is a very available and well-managed pool of compute resources that I know I can use when I need it.
8. LOTS of CPUs with low memory
9. Pretty much everything.
10. Providing tools for easy use
11. Very friendly, active staff. People were willing to help me acclimate to the system for running jobs and meet deadlines.
12. Everything
13. I appreciate Bill enormously for his help on everything about condor. He is indispensable.
14. Helpful and offering resources.
15. Great support from Bill Taylor setting up our pipeline.
16. This a very well-run service that we only recently started to use. The support has been excellent, especially as no one in my lab had prior experience with this form of computing. I foresee thiks being increasingly integral to the work my group does.
17. You guys keeping updating the condor to make it works better. It is very helpful! And Bill Taylor responses our emails and addresses our problem quickly. He is good and helpful!
18. Bill Taylor was very helpful in getting python to work on chtc resources.
19. Excellent team. Bill Taylor has provided awesome help.
20. The system is running pretty smooth without problems. It's really nice to have a high computing power at hand.
21. You are on campus.
22. Your willingness to engage with us when we knew nothing about your resources and skills was fantastic. I hope our work together will result in publications where CHTC will be cited or team members will be included as authors, but we only began working together earlier this year.
23. Providing extra overflow computing.