Engagement Report

July 1, 2010 – June 30, 2011

Brooklin Gore

Introduction

The Center for High Throughput Computing, 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 four years, the CHTC team has been actively engaging with campus researchers to help accelerate their science through the use of the CHTC's capabilities.

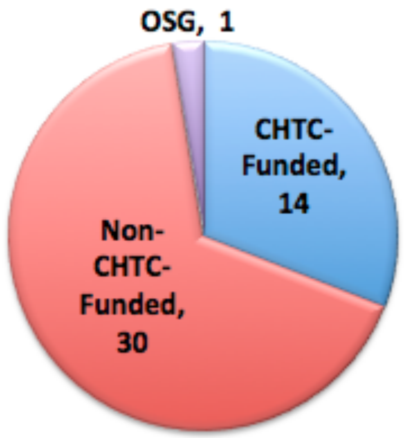
This report covers the engagement experience of 74 groups during the reporting period to better understand the impact of the CHTC on research and to learn where we can apply future resources to best meet the evolving needs of the campus. Results are drawn from internal accounting data and survey responses from 40% of all engaged groups. Internal knowledge of engagement status 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.
- **Impact.** Shows how we impact research.
- **Findings.** Highlights key take-aways from the report and areas of future work.
- **Appendices.** Provide the methodology for the report, a full listing of all users by utilization and comments from unsuccessful and successful engagements.

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, which are unified by Condor. Most resources are multi-core Linux-based computers, but there are a small percentage of Windows machines as well. 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 February 2011, we simplified the way UW researchers can access those off campus cycles via the CHTC. **Together, the CHTC provided 45 million hours of computing to campus researchers and their collaborators in the 12-month period.**

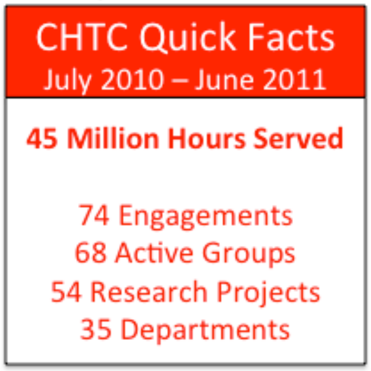
Fig. 1: Major CHTC Resources
(Million Core Hours)



Usage

During the reporting period from July 1, 2010 through June 30, 2011 **we actively engaged with 54 research projects in 35 departments on the UW-Madison campus.** We sustained resource-sharing arrangements with campus departments and participated in several national and international science programs. 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]. Appendix B provides a listing of consumption for all engagements.

Fig. 2: CHTC Quick Facts



CHTC Quick Facts
July 2010 – June 2011

45 Million Hours Served

74 Engagements
68 Active Groups
54 Research Projects
35 Departments

Top 25 Groups by Usage	
8.6	Chemical Engineering, Nanomaterials & Molecular Models
8.3	Large Hadron Collider – Compact Muon Solenoid (CMS)
4.4	Laboratory for Molecular and Computational Genomics
2.9	Technion - Israel Institute of Technology
2.7	IceCube Neutrino Observatory
1.8	Unspecified campus affiliations*
1.8	Large Hadron Collider - Atlas
1.4	Physics, Theoretical Physics
1.3	Botany, Genotype-to-Phenotype mapping
1.2	Chemistry Department
1.2	Backfill**
1.0	Chemistry, Polymers
.8	Computer Science Department
.7	Physics, Phenomenology
.7	Open Science Grid
.7	Physics, Quantum Algorithms
.7	Medical Physics, Brain MRI
.6	Electrical & Computer Engineering, Carrier Field Dynamics
.5	Electrical & Computer Engineering, Brain Connectivity
.4	Computer Aided Engineering Department
.3	Genetics, Developmental Genetics
.3	Chemistry, Condensed Phase Systems
.3	Chemistry, Computational Chemistry
.3	Engine Research Center, Engine Efficiency/Performance
.2	Biological Magnetic Resonance Data Bank

* A significant number of users of Computer Science resources pre-date our more formal engagement approach and are not yet classified by research group or specific department.
** Backfill is a technique used to provide idle cycles to low priority applications, in this case, we run 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.

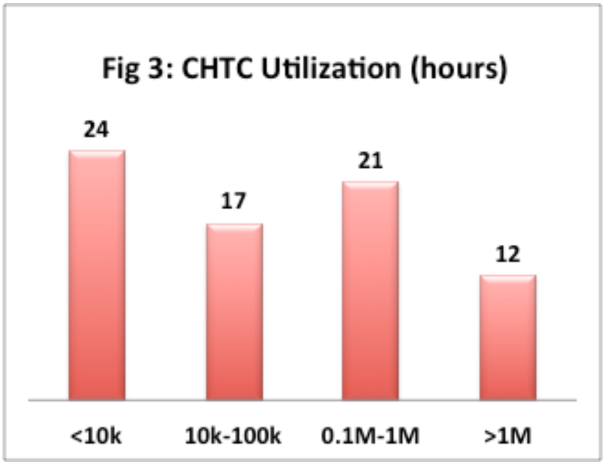


Fig 3: CHTC Utilization (hours)

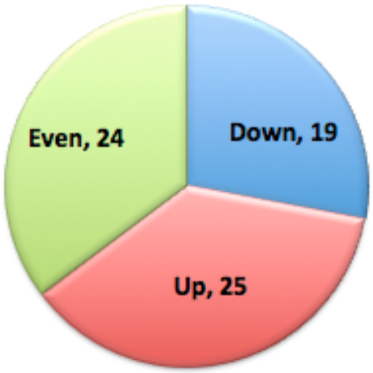
Of the 74 engagements performed during the period, 68 actively used CHTC computing resources. Figure 3 shows the number of groups in four utilization bands. **Almost half used over 100,000 hours and 12 groups appear 'hooked' on high throughput computing with over 1 million hours of usage each.** About a third of all groups used less than 10,000 hours over the 12-month period.

As is typical in academic settings, many research projects have one of two over-arching goals: publication of a PhD thesis or publication of a research paper. This situation often creates utilization waves for a given group. To better understand if CHTC use is increasing or decreasing, it is helpful to look at the usage trends of all projects over the period, since just one of the heavy users can affect the overall trend. Figure 4 shows that **utilization is either up or even for more than 70% of the active groups.**

Impact

31 groups responded to our survey, a 42% response rate. Of those, 25 indicated that the CHTC had helped them advanced their science. In most cases, we help by enabling researchers to analyze more data, analyze data more thoroughly, analyze data more quickly, or some combination of all three. **Figure 5 shows that over 80% of respondents indicated that their CHTC engagement led to better science.**

Fig. 4: CHTC Utilization Trends



Of the 25 groups who responded that we helped them do better science, 16% were able to do some good science to support a grant or publish a paper, generally a single event. More importantly, 84% of respondents indicated that the CHTC had become a key resource for their research. Figure 6 shows that 21 respondents have actually begun to rely on the CHTC as a foundational resource for their research. In fact, some of the comments in Appendix D indicate that **without the CHTC, some research would simply not be possible.**

Fig. 5: Advancing Science?

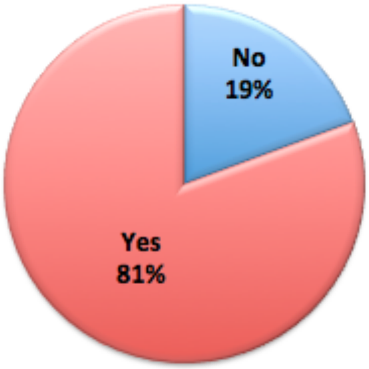
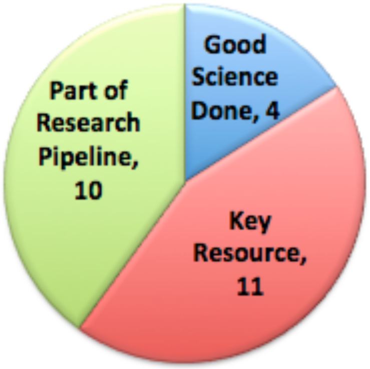
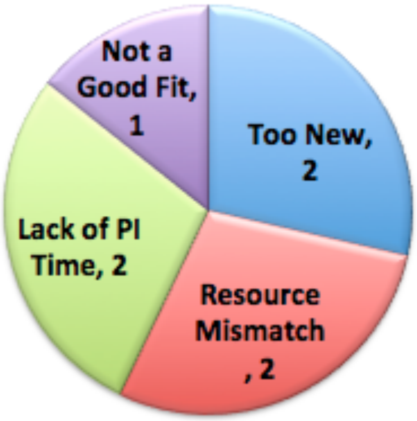


Fig. 6: Impact on Science



For the 6 respondents who indicated that their CHTC engagement did not advance their science, Figure 7 breaks out the various reasons. (Note: the number of causes is greater than the number of responses because a researcher could cite multiple reasons). **High Throughput Computing does not appear to be a good fit was selected by 14% of the responders that also reported that the engagement did not lead to better science.** For less than a third of the responses that did not advance, the CHTC lacked appropriate hardware (resource mismatch), such as large memory systems, Graphic Processing Units (GPUs) or efficient storage for very large temporary files.

Fig. 7: Why no Impact



Findings

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 D and from our internal and survey data. For example:

- Over 84% of researchers who reported that they are advancing their science indicate that the CHTC has become a key resource for their work, or is actually embedded in their data analysis pipeline. In some of these cases, research could not be done without the CHTC.
- CHTC engagements are requiring more cycles (25 trending up, vs. 19 trending down).
- High throughput computing was not a good fit for only 14% of engagements that reported that science was not advanced. We try to pre-screen engagements to ensure a good fit so as not to waste researcher time. This prescreening helps keep this number low.
- Our simplified access to OSG provided a significant number of cycles to campus researchers. This approach should be further developed and expanded.

While producing this report helped us better understand where and why we are succeeding and where and how we can improve, it also raises new questions as about how to measure and evaluate the impact of CHTC on the campus:

- What is a reasonable percentage of all research that could benefit from high throughput computing?
- We don’t have enough data to make statements like: “Researchers who use HTC publish more papers than those who don’t”, or “Researchers who have access to HTC resources get more grants, or more grant dollars than those who don’t”.
- We don’t have enough data to say: “If we doubled the number of engagement personnel, we could double our number of engagements”, or “If we doubled the number of CHTC resources, we could do twice as much science”.

Methodology

We were inspired to produce an annual report of our engagement activities by an engagement report created by the Open Science Grid (OSG), of which we are active members. We chose to generally follow their approach with some minor modifications. The categories used by the OSG to classify their engagements were:

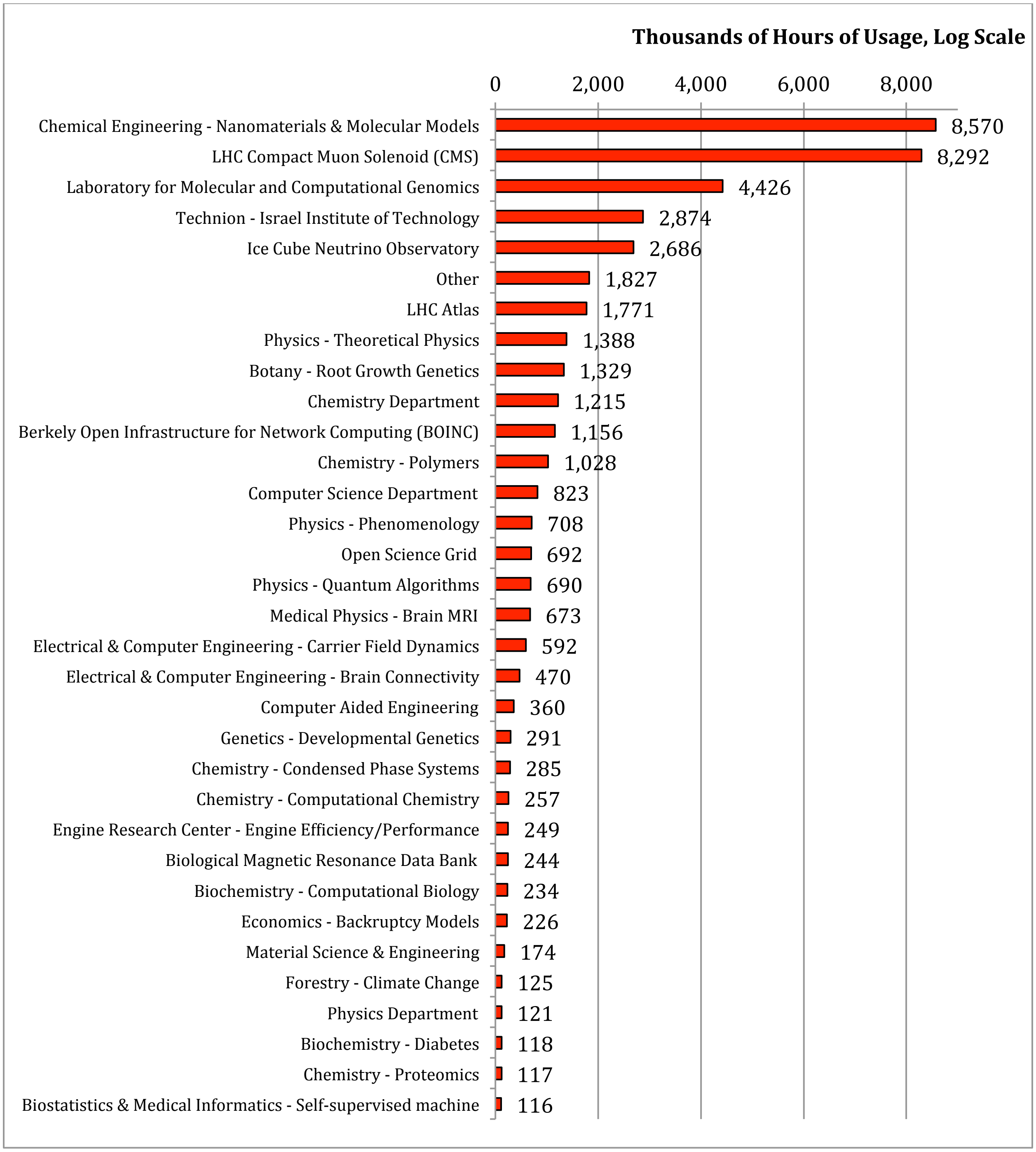
0. Unknown or no real interaction – we have no information
1. Interaction, but no real work (e.g. tried to engage but lack of interest from scientist or it was clear OSG was not the answer)
2. Engaged with scientist but didn't move forward (e.g. application not a good fit)
3. Engaged and got application to run on OSG, but scientist did not move real work to OSG
4. Real science work was done on OSG
5. Major science work continues to be done on OSG

We wanted to capture more information for the cases in which our engagement did not lead to active use of the CHTC. Further, we added a 6th category based on our experience.

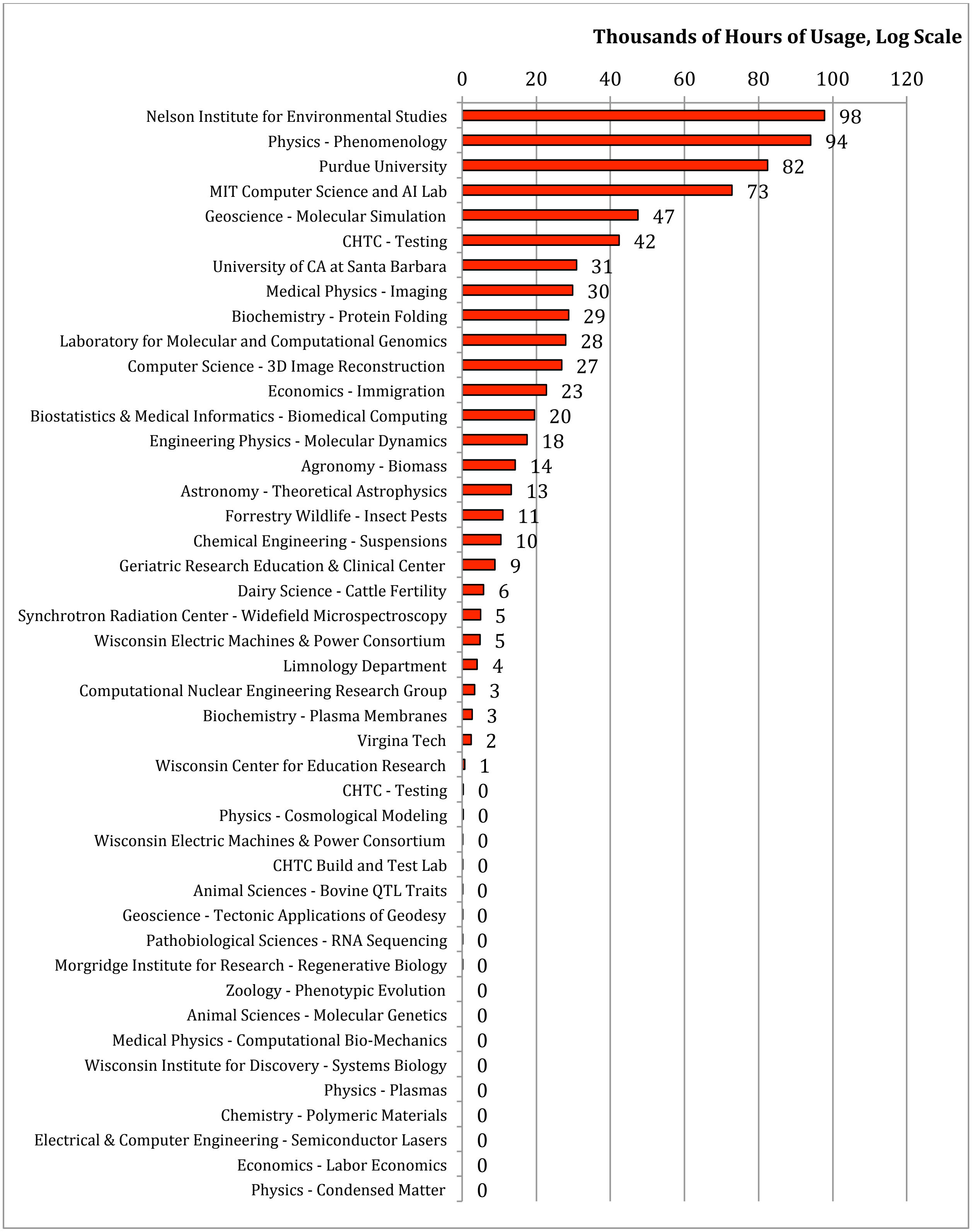
0. Not sure of state, no information from PI and lack of documentation on our end
1. We engaged, but weren't able to do useful work, because:
    a. The engagement is too new to tell for this reporting period,
    b. We (the PI's group) just didn't have the time or interest to continue,
    c. The problem wasn't a good fit for the CHTC
2. We engaged, tried to port our application, but it didn't work out, because:
    a. The CHTC didn't have sufficient time to move forward,
    b. We (the PI's group) didn't have sufficient time to move forward,
    c. There was some other issue preventing the work to move forward
3. We engaged and got our application to run, but didn't move forward, because:
    a. The CHTC didn't have the right computing resources (memory, etc.),
    b. The CHTC didn't have sufficient time to move forward,
    c. We (the PI's group) didn't have sufficient time to move forward,
    d. There was some other issue preventing the work to move forward
4. The CHTC helped us get some good science done
5. The CHTC is a key, ongoing resource for advancing our science
6. The CHTC has become part of our production pipeline

As a campus-focused organization, we have close, face-to-face interactions for most engagements, which let us internally classify each one. As a sanity check for how well we knew our customers, and as a way to 'hear it from the horses mouth', we also surveyed each group via email. We sent a single follow-up reminder to those groups who did not initially reply. We had a 42% response rate from groups for whom we did not help advance science and a 40% response rate from those we did. We incorrectly categorized only 2 out of the 30 responses we received.

Appendix B – Listing of All Engagements by Usage



Appendix B – Listing of All Engagements by Usage



My situation is best described by category 3(a), where a lack of efficient scratch storage was the limitation. We can efficiently split up a large calculation into many small ones, but each one generates a few (<10) GB of data. This data is subject to a global reduction at the end, but while waiting for it to finish, we need to store all the data somewhere. If we scale to 100s of jobs, then we need TB's of storage and would also prefer not to bring all that back through our network to a single location (even if we had the space available there). I am interested in re-engaging and can devote some staff and student time to the effort.

-- Computational Nuclear Engineering Research Group

Thanks for the email. Sorry to say our work with your center has not been able to move forward. The problem was mainly on my side. I guess my overall impression was that "our problem was not a good fit for the CHTC". Our work in computational bio-mechanics typically is High performance computing. I felt your scheduler and probably, the system are more or less not designed for this kind of work. This is my naive thought and I may be mistaken. We also need machines have larger amount of memory (>48G; ideally 128G) to run large models. Running commercial software on Condor might have IT related problems. I do hope that future development of cluster would allow us to take advantage of this wonderful resource. Please let me know if there is anything I can assist you further. Thanks for the email again.

-- Medical Physics, Computational Bio-Mechanics

I think we are still interested in moving forward with this, but the main holdup on our side is getting our software and simulation strategy set up for use. The meetings will be helpful in the set up on our side. 1) a and b, but we hope to continue at some point in the near future, 2) c, we are not yet prepared to use the high throughput computing environment. Thanks for keeping in contact with us.

-- Wisconsin Institute for Discovery, Systems Biology

I'm confident that Condor will be useful for my work, though I still have a few weeks or months of work to do before I'm to the point of organizing some runs. I guess I'd have to say 1a.

-- Physics, Plasmas

For us it was a combination of: 1.c: Simulated Annealing is intrinsically sequential, thus not a Directed Acyclic Graph suitable for Condor at CHTC. The attempt to use the Master Worker Framework was awkward. 2.c: Our application could use proprietary software called ABAQUS for finite-element modeling. Although we pay for several licenses at CAE, the staff there was unable (or unwilling?) to configure the license manager (flexlm) to serve license tokens to Condor Nodes at CHTC. Since then, however, we have been able to implement a Simulated Annealing algorithm that runs on a single condor node with 8 cores. In this way, we have developed an approach that can be written as a DAG. So that we can continue to use this approach, please keep the our logins active at CHTC.

-- Geoscience, Tectonic Applications of Geodesy

The CHTC is an integral part of the BMRB production services provided to the world wide biological NMR community through the BMRB web site. The CHTC provides the unique computation capabilities required to make the compute intensive protein structure determination software tool CS-Rosetta available to a large audience. The CHTC is a key to the success of the BMRB campus project and is being used to stress test the capabilities of the CS-Rosetta algorithm for computing protein three-dimensional structures. In the past year, approximately 400 successful CS-Rosetta jobs have been executed on the CHTC by scientist. This has contributed greatly to the understanding of the applicability of the CS-Rosetta tool for determining three-dimensional structures of proteins using NMR chemical shift data either alone or in combination with other kinds of data. Educators have used the CS-Rosetta CHTC implementation to train upper level undergraduates in the use of protein chemical shift data to determine protein three-dimensional structures in a classroom environment.

-- Biological Magnetic Resonance Data Bank (BMRB)

I'd like to say that the CHTC system helped us get some good science done. I can get results faster with the system because it can simulate multiple models in parallel.

-- Wisconsin Electric Machines & Power Consortium

Right now CHTC is a key resource for at least two of my research projects. So I would say that I am in category 5). (It changes with time, but over the past five years, the overall trend of my use of CHTC is definitely upward.)

-- Physics, Quantum Algorithms

Simply put, the CHTC resources enabled us to conduct an important study that we otherwise would not have been able to do. We hope to continue to be able to make use of these invaluable resources. Thank you!

-- Electrical & Computer Engineering - Carrier Field Dynamics

We simply could not do the things we are doing now without CHTC. It literally rescued a grant - we had proposed this work and it ended up being much more computationally involved than we had anticipated. Fortunately CHTC came to the rescue and allowed us to accomplish what we had promised and more. All of our brain connectivity data is now processed through CHTC, and future research will certainly rely on it. The support of CHTC staff has also been excellent and has recently positioned us to take advantage of the OSG. CHTC is a critical piece of infrastructure for my research program.

-- Electrical & Computer Engineering – Brain Connectivity

4) Doesn't yet apply. I'd like to think we are doing good science, but we are not done--CHTC is helping us get done faster. 5 and 6) Yes definitely. We boast about it in grant applications and use it extensively for some intensive jobs, allowing us to create pipelines for image analyses that do several hundred jobs in the time it would take us to do one job. Thanks much for providing this service!!

-- Geriatric Research Education & Clinical Center

4) The CHTC helped us get some good science done - This has been the case for several years. Problems that would have been beyond our infrastructure are now put in the trivial category.
5) The CHTC is a key, ongoing resource for advancing our science - We do use these resources on a regular basis. It has changed they way we look at solving infrastructure problems. 6) The CHTC has become part of our production pipeline - While the CHTC has become a part of our research production pipeline, it has also provided the proof of concept that is helping us to make a bid for commercialization. Some of or work will be moving to a for-profit research firm that uses practices developed with the CHTC to create a commercially viable solution.

-- Wisconsin Center for Education Research

Sorry for the delay - I've been away for too long:-) I think we are somewhere between 5 and 6.... We have been doing some GPU computing and the results are very promising. Could certainly use many more....even without parallel GPU computing....

-- Chemistry Department

I would choose option #4, CHTC has helped us get some good science done. I specifically work on doing Genetic Algorithm optimizations of internal combustion engines using Computational Fluid Dynamics (CFD). Engines today are required to meet strict pollutant emissions regulations and there is a growing demand to improve fuel efficiency. Unfortunately, there is typically a trade-off between pollutant emissions and fuel efficiency, so uncovering a operating strategy that yields low emissions and good fuel efficiency is challenging, especially experimentally. Thus, we use detailed CFD models coupled to a genetic algorithm to optimize the engine system to run with simultaneously low emissions and good fuel efficiency. An optimization typically takes 30 days and over 1000 CFD simulations. At any given time during the optimization, 32 designs are being simultaneously evaluated by the CFD model, thus this work would not be possible without CHTC.

-- Engine Research Center, Engine Efficiency/Performance

I'd say based on the TF-Cluster work, we can at least claim 4) below, and we are working toward 5) with the implementation of bowtie (which will eventually then become part of our production pipeline). Good stuff!

-- Morgridge Institute for Research, Regenerative Biology

To sum up the effect of the CHTC on our work, it's fair to say that it helped us transition from traditional biology in which we made detailed measurements on one or two genetic/y distinct forms of seedlings, to large scale biology in which increased throughput of measurement enables population-based investigation. The results can be studied with statistical genetic methodologies to map genotype to phenotype. It is quite clear that our grant success increased as we convincingly made that transition. And it's going to be increasingly clear in the future.

-- Botany, Root Growth Genetics

I'd say my experience with CHTC falls most closely into (6), since I couldn't have gotten the project done without Bill Taylor's help. He was very generous with his time and assistance, even getting extra computing resources together when I was up against deadlines on more than one occasion. Hopefully I'll be able to finish the paper up in the next month or so and sent him a copy, so he'll be able to see what all his labor was for. Working with CHTC has been great. I couldn't have asked for a better experience.

-- Economics, Bankruptcy Models

Gee, can I say all of the above? #5 is definitely true, in that CHTC has become more than just "overflow" compute capacity, but (at least for certain types of jobs) really enables new science for us. The only thing that would be helpful is perhaps additional nodes that are allocable as a whole. In addition to our molecular dynamics / Monte Carlo calculations, for which Condor is perfect, we often run electronic structure jobs where it would be nice to get all the procs / memory / disk for a node. While there are some CHTC nodes available this way, there number is much smaller. As such we often reserve our personal cluster for these types of jobs, and do "everything else" (true high throughput stuff) on CHTC.

-- Chemistry, Computational Chemistry

#6 best describes the CHTC for me. When I need to analyze my results for my science, I use the CHTC as part of my production pipeline. The CHTC has allowed me to perform more analysis for my science than I would have been able to do otherwise. Bill Taylor and SOAR have been crucial to my work, permitting me to access many more machines than I otherwise have access to. These additional machines have provided me with additional results that would have taken otherwise a prohibitively long time to calculate.

-- Astronomy, Theoretical Astrophysics

A resounding "yes" on all counts, except (4); no, we did not get some "good" science done, we got GREAT science done.

-- Laboratory for Molecular & Computational Genomics