The UC Santa Cruz Hummingbird Project: Research, Outreach and Instructional Computing (ROIC)

(Return On Investment Computing)

Submitter

Andrea Hesse, Academic Divisional Computing Director, ITS, UCSC, ahesse@ucsc.edu

Project leaders

Andrea Hesse, Academic Divisional Computing Director, ahesse@ucsc.edu
Stephen Hauskins, Computing Director, Division of Physical & Biological Sciences, hauskins@ucsc.edu

Team members

Alison Lindberg, Research Consultant, Physical & Biological Sciences, alindbe@ucsc.edu
Brad Smith, Director of Research & Faculty Partnerships, brad@ucsc.edu
Ted Buchwald, System Administrator, Baskin School of Engineering, buchwald@ucsc.edu
Doug Niven, Academic Computing Expert, Social Sciences, dniven@ucsc.edu
Eric Shell, Software Support, Baskin School of Engineering, eshell@ucsc.edu
Joshua Sonstroem, Unix Systems Administrator, Core Technologies, jsonstro@ucsc.edu
Kirk Loftis, Data Center Operations, kirkl@ucsc.edu
Michael P Edmonds, Computing Director, Social Sciences, medmonds@ucsc.edu
Michael Usher, Storage System Administrator, Baskin School of Engineering, musher@ucsc.edu

The Hummingbird Research/Instructional Support Project

UC Santa Cruz Hummingbird cluster provides an open access high performance computing environment that offers a range of resources for use by all UC Santa Cruz students, researchers and faculty members on demand. (https://www.hb.ucsc.edu)

The Need Looking for a Solution

High end computational capabilities and support are rapidly becoming a necessity for UC Santa Cruz faculty, students, researchers and collaborators.

This type of capability is disproportionately distributed. We have those who obtain research grants or start up funds that allow them to purchase what they need. On the other hand there are faculty, students, postdocs and research staff that do not have access simply based on their financial situation. Hummingbird is the response to this inequality, providing a computational resource to the many that don't have local access to anything else. The concept is centered around the idea of sharable resources supported by all of the Academic Divisions and Information Technology Services (ITS). Shareable computing resources translate into overall

lower cost and at the same time giving access to more researchers, students and instructors while also helping to manage the support load associated with growing demand.

Hummingbird represents a convergence of resources and needs. UC Santa Cruz has a very well established Science DMZ (NSF), ceph file storage system, data center and network services. Hummingbird offers faculty, researchers, graduate and undergraduate students in multiple divisions and departments access to research and instructional computing resources to leverage the other services and is managed by the local information technology staff who have the skills to provide the needed support.

The Hummingbird support team recognized this overarching need and responded. The team membership covers all of the academic divisions for the campus. Each member represents the particular needs of their constituency: science, engineering, social sciences, humanities and arts. The team also recognizes the many overlapping components of doing high level computing. The result is the leveraging of resources that allows a shared model that addresses the majority of need. Several academic divisions along with ITS supply funding for hardware components.

Hummingbird enables the campus to support and promote STEM projects and give resources to a student population that would otherwise not be available. From linguistics to economics, from genomics to computational media, the team supports a diversity of clients, by providing an accessible computational resource to all campus academic members. This spring, the cluster supported both graduate and undergraduate courses in applied math and statistics. This summer we will have the BD2K Summer UP workshop via the UC Santa Cruz Genomics Institute using Hummingbird (https://bd2ksummer.ucsc.edu/). There is also new faculty connected with the UCSC Treehouse Childhood Cancer Initiative that have expressed strong interest in utilizing Hummingbird (https://treehousegenomics.soe.ucsc.edu/). MCDB labs are also interested in utilizing the cluster for DNA analysis.

The Hummingbird team supplies clear documentation on how to use the service by way of our website and offers a “best effort” platform for training and education on the usage of Hummingbird cluster, from basic terminal command line interaction, software installation assistance, to submitting computational jobs. We participate directly in courses for cluster usage instruction. Having discovered that many students, including graduate students, arrive on campus without the necessary basic skills, we are developing open workshops for each academic quarter to cover basic cluster usage.

How a Solution Drove Innovation

Hummingbird represents the bridge between local clusters and those located at various institutions: San Diego Supercomputer Center, TACC, NERCS, and national research laboratories. We have modeled the architecture after larger open access computational clusters, such as Comet and Stampede to enable interoperability, providing a local environment

where researchers can develop proof of concepts to support their application to these larger environments. It can also provide a path to using cloud based services like Amazon and Google. The team is pursuing the best avenues to utilize cloud computing to augment our computing environment. The future is certainly a hybrid model of local resources, large scale HPC facilities and cloud service providers.

Hummingbird allows for full support of individuals that don't need high end supercomputing environment, as well as, those that are doing prototyping for research projects on local and nationally available clusters and students that want to learn and hone their skills in cluster usage while completing research projects that support their senior and graduate theses. We support numerous courses that need computational capabilities beyond the desktop for pure research instruction that covers Chemistry, Astronomy, Physics, Genomics and Machine Learning. Course instruction needs cover quantum mechanics to economics: Hummingbird fills the bill.

Hummingbird is responsive and nimble to client needs. We can rapidly install and upgrade compute nodes, install or update software, analyze client needs and create a solution. When the cluster as configured cannot support a particular need we can create containers that allows them to use the cluster via a specialized compute environment (https://singularity.lbl.gov/). This solution arose from the need of a linguistics faculty member that needed a particular computing environment for their software. We recently introduced GPU computing to the environment and the response has been excellent: course and individual requests wanting to do machine learning and digital media projects.

We are embracing the research facilitator role, an NSF funded model at several universities that holds annual workshop/conferences at UC San Diego to give participates the tools and knowledge to improve and assist our local research community (http://research-it.ucsd.edu/2018workshop.html). We send team members to participate in the conference annually. There is a new mantra on the horizon: let researchers do research without the burden of having to know an overwhelming amount of computer linguistics.

We think UC Santa Cruz has a unique advantage in being on the frontline of supporting emerging cross-discipline collaborations across the campuses and across the country, and even to some degree across the world (CERN). We seek to recognize what opportunities exist and how we can leverage them to support as fully as possible the computational needs of our campus clients using local and external resources.

Deployment Timeline

The prototype for Hummingbird (known as CampusRocks) was built on donated hardware to validate the need. With the Physical and Biological Sciences, School of Engineering, Social Sciences and Information and Technology services all committing resources in 2016, we began to develop a service team, public facing web site and a service roadmap. In 2017 we began implementing against that roadmap in earnest, migrating to OpenHPC, deploying the SLURM scheduler and establishing separate research and instructional queues. In 2018 we added our

first GPU nodes. The system has seen over 750 unique users since our upgrade over the summer of 2017. The service team meets regularly to plan for upgrades and maintenance. Skill development to provide support is now written into staff professional development plans. Web based documentation for both research and instructional use is updated routinely and example SLURM scripts are provided for “quick starts” to commonly used implementations. Users also receive quarterly updates on the service by way of newsletter.

What Clients are Saying...

“Hummingbird has been invaluable for my bioinformatics research with marine metagenomics data. The cluster has enabled me to investigate new ways of assembling and annotating 40-50 of these large datasets, with great speed (both due to fast cores, lots of memory, and parallelization) and reliable backup of scripts and results. I could not have done the same experiments in a reasonable time on my laptop, which would have been unusable for other research tasks had I tried. Finally, the cluster computing skills I have developed by working on Hummingbird — my first such experience — will be essential for my bioinformatics work after graduate school. Thanks for maintaining such an important resource!”

“I use the cluster for my research in medical imaging. I run simulations in a program called Geant4 which simulates particle interactions in an imaging system. Each simulation requires modeling the behavior and interactions of approximately 200 million proton events, and requires at least 90 cores, therefore, there is no other resource on campus that allows me to do these simulations in a timely manner. Likewise, it is essential to my work that the cluster function efficiently since time is of the essence. Some of the machines, namely (02 and 04) function at 1/3-1/5 the speed of some of the other machines which is extremely frustrating.”

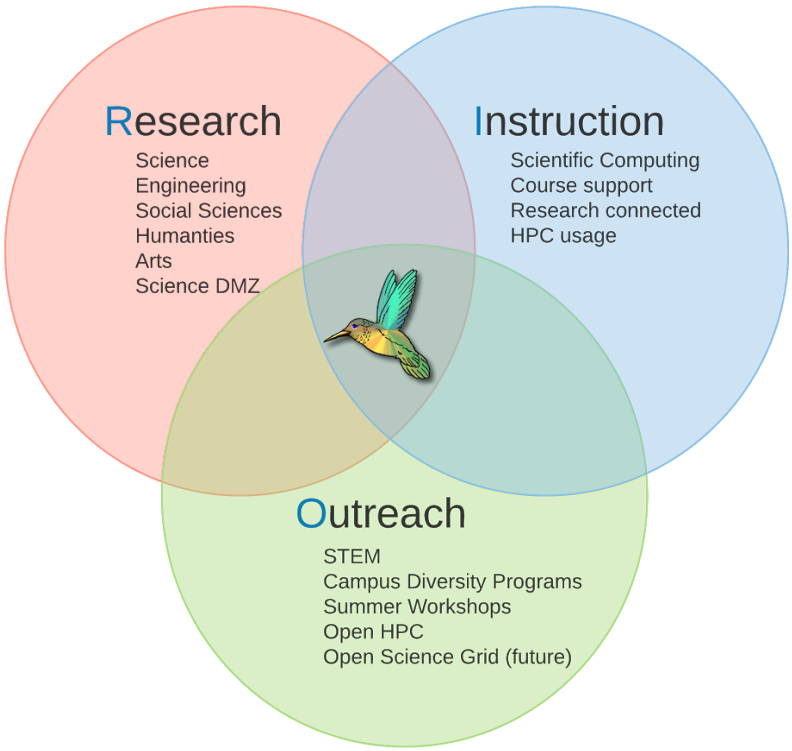
“I use this cluster for parallelized monte carlo simulations of high energy particle physics processes, especially related to dark matter. I also utilize the cluster for simulating and processing large sets of gamma-ray data in order to search for astrophysical signatures of dark matter. A significant expansion of these resources would be of great value to UCSC's research programs.”

“I am an undergraduate working with Dr. Camps in METX. My cluster utilization involved analyzing co-variation in cancer databases, namely cbioportal, to provide functional context clues for an orphan gene. Proper analysis requires using the entire genome as a query set, which can be computationally intensive. Hummingbird is a great resource, thanks for your work.”

“I am a graduate student in Scott Lokey's lab. We use the cluster for running molecular dynamics simulations on virtual libraries consisting of thousands of members. While each individual simulation is fairly brief and computationally inexpensive, the numbers mandate parallelism. The campus rocks cluster provides a wonderful and free resource for running these simulations. We greatly value its functionality and will continue to use it in whatever capacity we can.”

Hummingbird creates a partnership between faculty, researchers, students, educators and information technology services staff to deliver a research computing environment that supports as many as possible. We encourage and invite the addition of hardware to the cluster from any faculty member or campus group.

HUMMINGBIRD'S COLLABORATION SPHERES



The Anatomy of Hummingbird

- 15 Intel nodes (360 cpus)
- 4 AMD nodes (224 cpus)
- 1 node with 4 GPUs - machine learning, faster scientific software execution
- 1 storage chassis for home directories, software, and scratch (temp) usage
- Head node for submitting computational jobs
- 10 gigabit per second cluster network
- Queue batch system for handling job submissions
- Ability to easily setup a client environment
- Installed software includes. matlab, stata, python and scientific packages
- Self help and information website: https://www.hb.ucsc.edu/