

Fall 2017

Research Highlight

3D Reconstruction of Biological Macromolecules

by Peter Shen, Department of Biochemistry

All biological functions depend on atomic-scale processes. These processes are orchestrated by biological macromolecules, including proteins, nucleic acids, lipids, and carbohydrates, and their larger complexes. The field of structural biology focuses on studying how three-dimensional structures of macromolecules carry out their various functions. Understanding molecular structure can provide important insights into the causes of disease and inform ways to develop treatments.

Structural biology is now in the midst of a golden age ushered in by recent and ongoing developments in electron cryo-microscopy (cryo-EM). Developed in the 1980s, cryo-EM has long been an invaluable tool to directly visualize large complexes, including viruses and ribosomes. However, cryo-EM was traditionally limited by its relatively poor achievable resolutions compared to traditional high-resolution structure determination approaches such as X-ray crystallography or NMR. In 2012, the resolution barriers in cryo-EM began breaking down with the development of direct electron cameras—highly sensitive and fast detectors that preserve atomic-resolution information during the imaging process. In addition to improved detectors, the cryo-EM field has also enjoyed improvements in sample preparation methods, microscope technology, and image processing algorithms to generate the 3D structures. Due to these improvements, the cryo-EM field is now exploding as demonstrated by the sudden increase of high-resolution structures resolved within the past five years, the selection of cryo-EM as the Method of the Year by Nature Methods in 2015, and the award of the 2017 Nobel Prize in Chemistry for the development of the technique.

Cryo-EM at the University of Utah

In recent years, the University of Utah has established an outstanding core base of cryo-EM expertise and compiled a strong track record of performing impactful cryo-EM research (see Box 1 on page 2). These efforts have resulted in the University of Utah being awarded one of five \$2.5 million grants from the Arnold and Mabel Beckman Foundation to establish a world-class cryo-EM facility. The facility will be housed in the Crocker Science Center and will contain state-of-the-art instrumentation, including a FEI Titan Krios Transmission Electron Microscope equipped with a Gatan K3 Direct Electron Detector. Similar setups at other leading institutions have already produced atomic-resolution structures. Installation of our instrumentation will begin in November 2017 and is expected to take two to three months.

From Biochemical Sample to 3D Reconstruction

A cryo-EM project begins with a purified biochemical sample. The sample is applied to the surface of a metal mesh grid with a supporting layer of a thin film, typically amorphous carbon. The film consists of tiny holes that support a thin layer of the sample, which is plunge-frozen into a cryogen such as liquid ethane, thus trapping the particles in vitreous, amorphous ice (see Figure 1). The cryo-EM grid is loaded into the transmission electron microscope and hundreds to thousands of movies are recorded. After the imaging session, the individual frames of movies are realigned to correct for movement that occurs during electron beam exposure, thus recovering the signal that would otherwise be blurred out. Each motion-corrected movie is then summed into a single image, or micrograph. The remaining workload of a cryo-EM project is data processing and analysis.

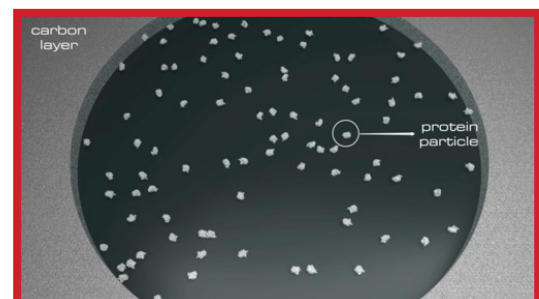


Figure 1. Illustration of freestanding, randomly oriented particles trapped within a cryo-EM grid. Illustration created by Janet Iwasa.

Each cryo-EM micrograph typically contains hundreds of freestanding particles randomly oriented throughout the thin ice layer. The particles can be identified using manual or automated selection methods and are then isolated into an image stack. Cryo-EM dataset stacks often consist of tens of thousands to a few million particle images and thus their processing represents a daunting computational task. The overall goal of data processing is to assign orientations to each particle image so they can be back-projected into the correct 3D structure. The task becomes even more challenging when considering that cryo-EM images are inherently noisy because specimens are unstained, have low contrast, and are imaged using low electron doses to minimize radiation damage from the electron beam.

Box 1. Cryo-EM highlights at the University of Utah

Clockwise, from top left:

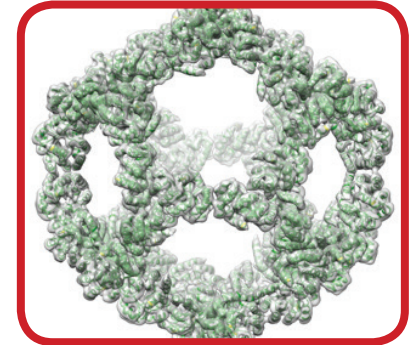
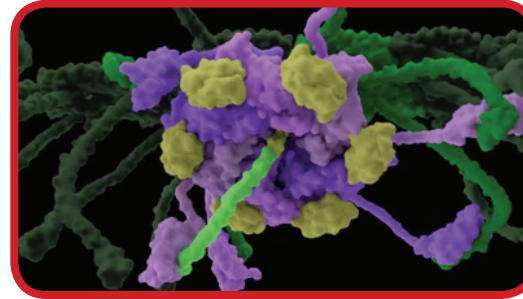
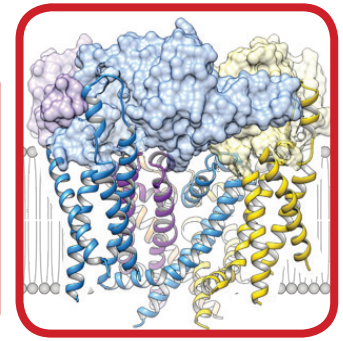
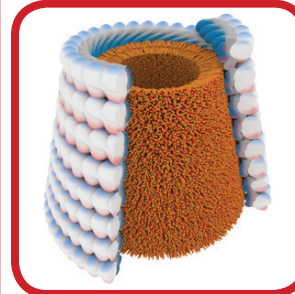
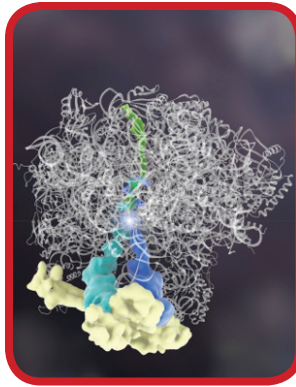
Discovery of mRNA-independent peptide synthesis by the Ribosome Quality Control Complex. This structure shows that the Rqc2 protein (yellow) directs protein elongation without an mRNA template. (*Shen et al., Science 2015, doi: 10.1126/science.1259724*)

Membrane remodeling by ESCRT-III filaments. Model illustrating how ESCRT-III spirals (dark and light blue) can assemble structures of decreasing diameter, as might be required to constrict membranes (orange) together to a fission point. (*McCullough et al., Science 2015, doi: 10.1126/science.aad8305*)

Structure of the Polycystic Kidney Disease channel PKD2. Mutations in PKD2 cause autosomal dominant polycystic kidney disease, one of the most common, life-threatening genetic diseases. (*Shen et al., Cell 2016, doi: 10.1016/j.cell.2016.09.048*)

Cryo-EM reconstruction of a designed, self-assembling nanocage (grey) fitted with the computational design model (green). The nanocages can be used as delivery vehicles to transfer molecules between cells. (*Votteler et al., Nature 2016, doi: 10.1038/nature20607*)

Model of the Vps4 molecular machine (purple), which powers membrane remodeling events including HIV budding and cell division. (*Monroe et al., eLife 2017, doi: 10.7554/eLife.24487*)



After the particles have been selected, the data processing begins with enhancing the signal-to-noise ratio of particle images via 2D image analysis. Most 2D analysis procedures are derived from K-means clustering algorithms that iteratively assign all particle images to a defined number of classes. Ideally, images within the same class are similar to each other and each class is as distinct as possible. During this process, images are averaged in each class, leading to so-called 2D class averages that begin to reveal high-resolution features of the particle. These class averages are often useful to identify bad particles, artifacts, contaminants, or pure noise that can then be excluded from the dataset prior to 3D reconstruction.

From a computational standpoint, 2D class averaging procedures are relatively straightforward. The algorithms assign particle images three parameters: x and y translational shifts and an in-plane angular rotation. However, reconstruction in 3D represents a significantly more demanding challenge because each particle must be assigned five parameters: x and y translational shifts and three angular rotations to orient the particle in 3D space. The principle behind 3D reconstruction is rooted in projection matching algorithms and can be considered as a version of K-means clustering: each particle image is matched against 2D projections of a 3D model and the best match gives its orientation assignment. After the orientations have been assigned to all particles, the 3D reconstruction is computed by

back-projection of all images, which creates a new model for the next iteration of projection matching and reconstruction. The process iterates until convergence criteria are met, for example when the angular assignments stop changing.

Processing Cryo-EM Data at CHPC

Most of the cryo-EM data analysis procedures at the University of Utah are performed using CHPC resources. CHPC supports many software packages used in cryo-EM data processing, including RELION, EMAN2, SPIDER, FREALIGN, BSoft, and cryoSPARC. These packages offer varying degrees of functions, flexibility, and use different approaches, but the general principles underlying data processing are more or less the same.

One of the most heavily used 3D reconstruction programs at CHPC is RELION, which implements an empirical Bayesian approach for structure determination. In RELION, the 3D reconstruction problem is treated as finding the most likely 3D density map given the 2D particle images of unknown orientations (i.e., the observed data) and the expectation that biological structures have smooth characteristics (i.e., the prior information). Optimization of the posterior probability is achieved by an expectation-maximization algorithm. The expectation step is essentially the particle alignment step where each particle image is compared to

all projections from the reference map. Not surprisingly, this step can be computationally expensive depending on total number of particle images, the dimensions of the images, and the search space sampling rates. The maximization step is the reconstruction of a new density map, which then feeds into the next iteration of the expectation step. One of the major advantages of this Bayesian approach is that most of the statistical parameters are inferred from the data, thereby minimizing user intervention and improving ease of use.

To reduce computation times, RELION uses a hybrid parallelization scheme at two separate levels during the expectation step. First, RELION uses distributed-memory parallelization (via MPI) to divide a dataset into subsets of images that are processed in parallel across different nodes. Second, RELION uses shared-memory parallelization via threading, which parallelizes tasks among multiple cores across a single node. Threading has the advantage of sharing the memory of one computer.

Despite these approaches to speed up calculations, the underlying regularized likelihood optimization algorithm often requires a massive amount of processing power. In 2015, our efforts to reconstruct a ribosome complex (using ~80,000 420²-pixel images) required over a week and more than 40,000 SUs. More recently, a GPU-optimized version of RELION was released that implements its parallelism in CUDA. By using this version on the CHPC GPU resources our data processing throughput has been dramatically boosted, to the extent where jobs that normally take a week or longer on CPU nodes can now be finished within a couple of days. Earlier this year, a new software package, cryoSPARC, was introduced that enables even faster orientation assignments to

particle images. The program applies a branch-and-bound algorithm to quickly screen through large areas of search space and exclude regions that cannot contain the correct orientation assignment for each particle. Within a couple of computationally-inexpensive iterations, the search space is narrowed down to a small area and further iterations are spent to fine-tune the orientation assignment. As of September, we now have a dedicated cryoSPARC machine hosted at CHPC. In preliminary benchmark tests, we went through the entire process of 2D class averaging to generating an atomic-resolution 3D reconstruction within a few hours. Another remarkable feature within cryoSPARC is its ease of use through its intuitive browser-based GUI. Inputs and outputs are managed under the hood and the user can perform about a dozen of different cryo-EM data processing functions within a few mouse clicks.

Altogether, these exciting developments are clear indicators that cryo-EM is still on the rise. As the field continues to grow, it will be increasingly important to make hardware and software tools more accessible to newcomers. One major focus in the field is to fully automate the entire pipeline of recording cryo-EM movies, de-blurring the images, identifying the particles, and reconstructing them in 3D. Perhaps the time is not far off when high quality 3D reconstructions are attainable within hours after the cryo-EM imaging session. Our ongoing collaborations with CHPC will certainly play an important role for this dream to become a reality here at the University of Utah.



User Services – CHPC’s New Cyberinfrastructure Facilitator Brett Milash

CHPC has added a new cyberinfrastructure facilitator to its staff in August. Brett Milash, previously the Co-Director of the Bioinformatics Shared Resource at the Huntsman Cancer Institute, has joined the User Services team at CHPC as a Scientific Consultant and brings with him over 20 years of experience in bioinformatics, scientific computing, statistics, database design, and programming skills in python, R, and C/C++. In his new role at the CHPC, Brett will be specializing in workflow management tools (such as Snakemake, makeflow, and systemPipeR) and processing sensitive data in the protected environment, in addition to fielding bioinformatics-related support issues.

Brett’s position was made possible by a recent NSF award, “CC* Cyber Team: Creating a Community of Regional Data and Workflow Cyberinfrastructure Facilitators”. This award funds the creation of a three person “Cyberteam” of data and workflow facilitators, one each at the University of Utah, University of Colorado Boulder, and Colorado State University, with each institution having a specific focus:

UNIVERSITY OF COLORADO – data and compute workflows

COLORADO STATE UNIVERSITY – data curation preservation, metadata

UNIVERSITY OF UTAH – protected data and compute workflows; use of virtualization methodologies

The institutions and this team will work together to provide the infrastructure and support for the data needs of researchers at all Rocky Mountain Advanced Computing Consortium (RMACC) institutions. This award will focus on the need for researchers to be able to access, process, analyze, understand, curate, and/or preserve their data in a reliable and secure manner such that it is readily available for sharing and future use.

by Luan Truong



Jupyterhub is a multi-user version of Jupyter notebook, a web application that provides an environment that enables users to edit and run code from a browser and to attach the media representation of results (HTML, LaTeX, PNG, PDF, etc.) with the code that generates them. The notebook is served such that it is readily available to anyone with an CHPC account and work is automatically synced to the user's home directory. Users can also spawn a terminal to their home directories and use additional tools needed from the command line.

This CHPC Jupyterhub instance, notebook.chpc.utah.edu, was deployed at CHPC in response to requests from two chemical engineering faculty members, James Sutherland and Tony Saad, who both wanted to use Python via Jupyter notebook for their classes. However, many in the department were concerned that Python had a steep barrier for entry as students would need to install Jupyter notebook - unlike Matlab, which has been used in these classes in the past. This position changed when James and Tony pitched the use of a hosted instance of Jupyterhub with the idea that students would have access to a Python programming environment through a web browser, effectively removing all entry barriers.

The initial feedback has been positive:

From Tony Saad: *"I couldn't be more thrilled about telling my students in CH EN 2450 (Numerical Methods) that they only needed to point their browser to notebook.chpc.utah.edu and have access to a free, beautiful programming environment that allowed them to write text, do math, and code in one spot!*

*Given that my students had a Matlab-only background, I gave them the option to choose either Matlab or Python for their homework. I am pleased to report that over 79% of my students are now using Python! Of these **79%**, the vast majority (**80%**) are using notebook.chpc.utah.edu. Student feedback has been very positive and many have reported their newly found affinity for programming thanks to Jupyter and Python.*

By hosting JupyterHub, CHPC has helped place the University of Utah on the leading edge of using modern technologies for teaching scientific computing and I am very proud to have been part of this effort. I don't know of any similar endeavor in Utah."

From James Sutherland: *"Beginning Fall 2017, I began requiring python and Jupyter notebooks for all homework in ChEn 6703 (Advanced Numerical Analysis for Chemical Engineers). notebook.chpc.utah.edu has been a wonderful resource for the class. It provides a common python environment and allows a web-based computation tool with minimal barrier to entry for students. Jupyter is changing the way that I teach classes, and notebook.chpc.utah.edu has facilitated this in large part. Students are able to combine equations (with cross-referencing), discussion, and results of in-depth calculations in one self-contained notebook. I plan to continue using notebook.chpc.utah.edu as a resource for all of my classes!"*

Due to the success with Jupyterhub this semester, CHPC has plans for a more permanent installation that can better support the anticipated growth in usage. Give notebook.chpc.utah.edu a try and let us know if you have any questions, as well as, watch for announcements on the new instance which will be called juno.chpc.utah.edu.

Let us know if you are teaching a class for which you feel Jupyterhub will be a good fit. We can create accounts for all of the members of your class.

Getting started is as easy as 1, 2, 3 if you have a CHPC account

Log in with your CHPC credentials at notebook.chpc.utah.edu

Will be dropped in your home directory.
The New >> Python 3

Code!



Utah will again have a team participating in the Student Cluster Competition held at the annual SuperComputing Conference. This year's team is being mentored by CS Professors Mary Hall and Hari Sundar along with CHPC system administrator Brian Haymore, as was last year's team. The team's industry partners are again NVidia and Dell. The team name is the "SupercompUtes".

Teams are comprised of six members. Before the conference, teams work with their advisors and industry partners to design and build a small cluster within the constraints specified by the competition rules and work to optimize the performance on a set of specified HPC workflows. At the conference, the team works in a "real-time, non-stop, 48 hour challenge" to assemble the cluster on the exhibit floor and then complete a real-world workload across a set of applications. Teams are judged by their performance on the LINPACK and HPCG benchmark, as well as on their overall performance on a set of four applications and on their knowledge of both the operation of the cluster and the understanding of the applications based on interviews with the judges. Awards are given for the highest LINPACK performance and for the overall winner, a composite based on the team's overall performance.

The competition, started in 2007, was created as a mechanism to expose undergraduates to HPC. It is designed to introduce the next generation of students to the HPC community and to the breadth of skills it takes to build, maintain, and run a HPC system.

Last year the SupercompUtes placed second (out of 14 teams) in the overall performance with a score of 86 points out of a possible 100 – only 2.5 points behind the winning team from the University of Science and Technology of China. This was a great showing for a first time entry from the U – and with four returning members from 2016, this year's team hopes to improve on this result!



University of Utah SC16 Cluster Competition Team



New CHPC Grant Announcement

CIF21 DIBBs: EI: SLATE and the Mobility of Capability (NSF award# 1724821, 7/1/2017-6/30/2021)

This grant, a joint effort between the Universities of Chicago, Michigan and Utah, lead by the University of Chicago, was awarded to implement 'cyberinfrastructure as code' by augmenting the Science DMZ pattern with a generic, programmable, secure and trusted 'underlayment' platform. This platform will use best-of-breed data center virtualization components and, where available, network techniques such as software defined networking, to enable distributed automation of deployment and service lifecycle management tasks by domain experts. The goal is to simplify the creation of scalable platforms that connect research teams, institutions, and resources in order to accelerate science, while reducing operational costs and development cycle times. The Computational Physics Program within the NSF Directorate for Mathematical and Physical Sciences has supported this award jointly with the Office of Advanced Cyberinfrastructure. The PI from Utah is Joe Breen.

https://www.nsf.gov/awardsearch/showAward?AWD_ID=1724821&HistoricalAwards=false

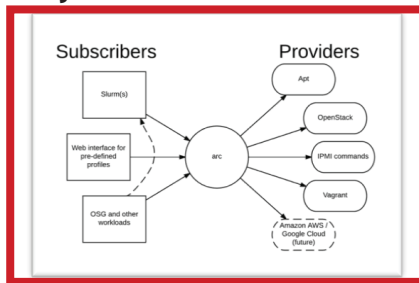
by The CHPC Student Staff

The University IT has a Student Employment Program, <https://it.utah.edu/student-program/>, which offers students a chance to supplement their education with real world IT experience. The program is mentored based, with each student receiving guidance from a full time staff member. Along with each student having a mentor, the program encourages student to work together, and holds a monthly social activity in order to promote student interaction beyond students working for the same IT department. Since the program started in 2015 with 15 students, it has rapidly grown to over 55 students. Each student-mentor pair meet monthly with a student lead, to review the progress of the student on their projects and tasks.

CHPC has been taking advantage of this program to hire student staff members since 2016; we currently have 13 students. These students work on a variety of projects – here are a few highlights provided by the students.

Kevin Parker – Arc tangent Project

The goal of this project is to provide a dynamic middleman between compute schedulers and resources. It is designed with a pluggable architecture, so both pre-existing software and software that does not yet exist can play nicely with each other, within the same pools of resources. Through this project, I have both been able to refine my skills in Python as well as learn about several new software packages and concepts that operate on different layers of the software stacks of compute clusters.



Paul Fischer – Exploring Storage Solutions

To meet the varied digital storage needs of HPC users, a diverse and robust suite of storage solutions is required. As a student at CHPC, I have spent most of my time investigating, deploying, and evaluating potential storage solutions. Recently, my focus has been on open-source and scalable Unix-compatible storage systems, primarily Ceph, ZFS, and Lustre. Demands for HPC resources are always increasing, and it is crucial that existing storage can be expanded smoothly. In working on these projects, I have greatly developed my knowledge of Unix systems and HPC-scale hardware and networking, as well as gained invaluable skills for resolving unfamiliar technical problems.

Rajeev Sahay – HPC Cluster Maintenance

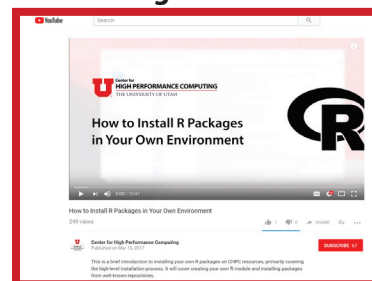
Managing the servers comprising the CHPC clusters is often difficult due to the hundreds of nodes in each cluster. In

an attempt to better manage the health of the servers, I was assigned the task of creating a bash script capable of delivering a daily report consisting of useful information such as the state of each node. In the process of creating this script, I learned how to script in bash, which I had no experience with beforehand. I also learned more about slurm, which is the back-end manager CHPC uses to retrieve node health information. Using the script has allowed CHPC to be more proactive in identifying node issues, sometimes before they are reported by users. Along with writing this script, I also work on the troubleshooting nodes when they are down.

Emerson Ford – Dynamic VM and Node Provisioning

For the past few months, I've been working on implementing a network testbed software called Emulab developed by the Flux Research Group at the University of Utah. Emulab allows for dynamic virtual machine and bare metal node provisioning to allow dynamic altering of the networks in which those virtual machine instances and bare metal nodes operate. Additionally, Emulab allows for the creation of "profiles" so you can save the images and configurations of your experiment. We hope to use this software to make HPC services more accessible as well as giving researchers another tool for HPC research. From this experience, I've learned a great deal about everything related to systems administration, such as how to manage networks, how to create images, and how to configure servers with tools such as IPMI.

Robben Migacz – CHPC Training Videos



One of my projects has been the production of a series of short training videos that cover topics such as connecting to computational resources, using software, and transferring data. The videos provide training for new users and are posted to CHPC's website and shared when relevant questions are asked on the help desk. The videos have also reached a global audience and have been viewed many thousands of times. They provide documentation and instruction that is more demonstrative and presentational than a manual or traditional documentation page and are well-received by users. Producing the training videos has taught me much about technical communication and has required improvement both in my knowledge of high-performance computing concepts and techniques relevant to the design and editing of multimedia content. You can see the videos at <https://www.chpc.utah.edu/documentation/videos/index.php>.

CHPC's Expanding Impact

CHPC Director Tom Cheatham is serving as the inaugural chair of the Campus Research Computing (CaRC) consortium. The consortium, organized through a National Science Foundation grant, has the goal of establishing a national forum for the exchange and dissemination of best practices, expertise, and technologies to enable the advancement of campus-based research computing activities. At the PEARC17 (Practice & Experience in Advancing Research Computing) in July, there was an initial CaRC Council meeting as well as an open "Birds of a Feather" session to gather input from the Research computing community on the role of the consortium.

Tom has also accepted the position of Vice Chair of the 2017-2018 RMACC Board of Directors. RMACC, the Rocky Mountain Advance Computing Consortium, is a collaboration among academic and research institutions in the intermountain states. The mission of the consortium is to facilitate the widespread, effective use of HPC throughout the region. As part of RMACC, researchers in the region have access to time on the Summit cluster at University of Colorado- Boulder.

Lonepeak Cluster Expansion

At the end of August an additional 87 general nodes along with 20 owner nodes were added to lonepeak. The general nodes are servers that were "repurposed" from other uses at the University outside of CHPC. As a reminder, the general nodes on lonepeak are run without allocation and without preemption, making them a good source of additional compute time if you have either no or limited amounts of general allocation and no owner node access.

For more information on these new nodes and lonepeak, see <https://www.chpc.utah.edu/documentation/guides/lonepeak.php#lpHardware>

2018 Open Science Grid All Hands Meeting (OSG-AHM)

CHPC will be the local host for the 2018 OSG-AHM which will take place at the University of Utah, March 19-22, 2018. As we approach the meeting, we will be sending more information to our users.

For information about OSG see <https://www.opensciencegrid.org/>. The AHM is an annual meeting to discuss various aspects of the open science grid, and includes sessions on research completed using this resource as well as training on the use of OSG resources.

New Cluster- Notchpeak

With the introduction of servers based on the Intel Skylake (XeonSP) processor, will be starting a new compute cluster this fall, "Notchpeak".

Martin Cuma obtained early access to a server with these processors and did benchmarking; his results can be found at: https://www.chpc.utah.edu/documentation/white_papers/skylake.pdf

As mentioned in this whitepaper, there are many choices of CPUs in this processor family. Based on price to performance findings, we are focusing on the Gold 6130, with 16 cores per processor and a base clock speed of 2.1GHz. Therefore, for a dual processor node, there will be a total of 32 physical cores.

Another change is the increase from four to six memory channels per CPU. Using all of the memory channels (recommended for performance and balance) this results in three options; **96GB** (12 x 8GB memory DIMMs), **192GB** (12 x 16GB), or **384GB** (12 x 32GB).

A final change is that we are moving from FDR to EDR for the Mellanox Infiniband fabric for Notchpeak. Current prices for servers with these new processors are listed below:

96GB memory	\$ 7,055.00
192GB memory	\$ 8,251.65
384GB memory	\$ 11,001.12

Price includes a 2TB local hard drive and comes with a 5-year warranty. To this price CHPC will add \$900/node for networking infrastructure and power connection costs.

If you are interested in learning more about the new servers, additional configurations, or our plans for Notchpeak, please contact CHPC.

Node-sharing Capabilities for Owner Nodes Partitions

CHPC now allows multiple batch jobs to share a single node. on owner nodes. We have been using the node-sharing feature of slurm since the addition of the GPU nodes to kingspeak, as it is typically most efficient to run 1 job per GPU on nodes with multiple GPUs. More recently, after working out the usage accounting details, we have offered node sharing to select owner groups for testing. At this time CHPC has no plans on adding node sharing on general nodes or top guest access to owner nodes.

Additional details on the use and performance impact of node sharing is given on the CHPC website at <https://www.chpc.utah.edu/documentation/software/node-sharing.php>. If your group is interested in learning more about implementing node sharing on their owner nodes, please contact us at issues@chpc.utah.edu.

**The University of Utah
University Information Technology
Center for High Performance Computing
155 South 1452 East, Room #405
SALT LAKE CITY, UT 84112-0190**

Thank You for Using Our Systems!

Welcome to CHPC News!

If you would like to be added to our mailing list, please provide the following information and send via the CHPC contact methods listed below:

Name:

Phone:

Email:

Department or Affiliation:

Address:

(UofU campus or U.S. Mail)

Please help us 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."

Please submit copies or citations of dissertations, reports, pre-prints, and reprints in which the CHPC is acknowledged one of the following ways:

Electronic responses

Email: colette.durrant@utah.edu

Fax: (801)-585-5366

Paper responses

U.S. Mail: 155 South 1452 East, Rm 405
Salt Lake City, UT 84112-0190

U Campus Mail: INSCC 405