Shared Research Computing Policy Advisory Committee (SRCPAC) – Fall 2015 Minutes
November 30, 2015, 10:00 a.m. – 11:30 a.m.
523 Butler Library

36+ Attendees: Kathryn Johnston, Halayn Hescock, Bruno Scap, Marley Bauce, Victoria Hamilton, Rob Lane, George Garrett, Rob Cartolano, Sander Antoniades, Dali Plavsic, Maneesha Aggarwal, Marc Spiegelman, Amy Nurnberg, Gaspare LoDuca, Raj Bose, Chris Marianetti, Greg Bryan, Alan Crosswell, Mahdad Parsi, Michelle Benson, Jennifer Brown, Ryan Abernathey, Michael Gaschler, Michael Weisner, John Villa, Serena Ng, Paul Blaer, Eric Block, Hugh Ediet, Jochen Weber, Mark Newton, Sharon Sputz, Roslyn Hui, Bob Mawhinney, Harmen Bussemaker, Eric Vlach. Julien Teitler, and some late arrivals.

Kathryn Johnston, Chair of SRCPAC and Professor & Chair of Astronomy, opens the meeting by welcoming all 36+ attendees, and asking everyone to introduce themselves. Each meeting attendee has been provided with a printed copy of the meeting agenda.

Yeti Operating & Executive Committees

Greg Bryan, Professor of Astronomy and Chair of the Yeti Operating & Executive Committees, presents an update on the November 5th meeting of Yeti Operating Committee, which provides guidance to Research Computing Services (RCS) on the operating policy of the Yeti high-performance computing cluster. Rob Lane, Manager of Research Computing, complements Greg's presentation by summarizing usage and recent changes. Nearly full capacity usage has held constant over the last eight months; over 300 individuals having submitted at least one job. Recent queue changes are working well, with two minor changes approved:

- To allow Free Tier users more CPUs at once, but for less time.
- To increase Infiniband queue time from 36 to 48 hours.

Greg notes that Yeti is working well, thanks to the work of RCS. **Appendix I: Fall 2015 Yeti Operating Committee Update**

HPC Expansion Schedule

Kathryn Johnston asked Rob to outline the approach to build the next generation HPC cluster after a false start six months ago (wherein SRCPAC did not receive sufficient interest, and so the expansion round was postponed until now). This will be the third cluster, after Hotfoot (#1) and Yeti (#2, and current). SRCPAC is seeking volunteers for a Faculty Advisory Committee (FAC).

The FAC Committee's first meeting will review a draft RFP (prepared by RCS), then meet two weeks thereafter to finalize the formal document for submission to select vendors, who will have one month to respond. The FAC Committee will choose a winner for the cluster and a winner for storage by the end of March, with configurations and prices posted by early April, and final orders in by the end of April. RCS can provide preliminary cost estimates earlier than April, but they will be "highly disclaimered," as was done in the previous round. Final pricing will be posted in early April so that PIs have most of April to procure and identify funds necessary to purchase nodes before the final day for orders on April 29th. This new system will once again include nodes for Free, Education, and Rental Tiers.

Following submission to Columbia's Purchasing Office, RCS expects one month for Purchase Orders to be generated, then another month for equipment to arrive, followed by two months of installation and testing. The Yeti Operating Committee will again allow the TIP program (Trial with Intent to Purchase),

which allows prospective users to gain access to Yeti if they have an intent to purchase into the new system. Those interested in TIP should contact RCS (rcs@columbia.edu).

Rob clarifies that this new machine will not be an expansion of the current cluster, but rather a completely new cluster. Current Yeti users will not have an account on the new cluster, and vice-versa.

Ryan Abernathey, Assistant Professor of Earth & Environmental Sciences, asks about the philosophical rationale behind this being a new cluster and not a Yeti expansion. Rob replies that we do not have storage capacity to expand Yeti, so the new cluster must be separate. Additionally, starting a new cluster will allow a "blank slate," whereby new decisions can be made. Rob gives the example that RCS may switch to a new scheduler, which can only be done with a new system (although Yeti will remain active for the four year lifespan, as promised.).

Rob asks whether a month is adequate time from announcing prices to having orders in; the committee replies that this timeline should be communicated via email at multiple times so faculty are well aware of impending deadlines, and that Chairs in particular should be well informed of the opportunity and schedule.

Kathryn notes that we will resurrect the previously constituted Faculty Advisory Committee, and anyone who wishes to join the committee should email RCS (rcs@columbia.edu). **Appendix II: HPC Expansion Schedule**

Reviewing SRCPAC Recommendations

Kathryn asks Halayn to outline responses to prior recommendations from other committee and subcommittee meetings.

- Intercampus Subcommittee has expanded to include A&S, SEAS, and CUMCIT.
- CUIT is nearly complete with establishing an enterprise agreement with Amazon
- CUIT and Research Initiatives are hiring an Intern to help track university options for external resources
- A survey for Data Science Institute faculty for their research computing power needs is not being conducted at this time, but DSI is included in their *ad hoc* Research Storage committee and other initiatives.
- RCS will redesign and rebrand rcs.columbia.edu, planned for early 2016.
- The Intercampus Subcommittee will not continue to meet.
- Maneesha Aggarwal now attends the monthly Basic Science and Engineering Chairs meetings.
- CUIT has completed a Cloud Pilot.
- SPA is developing guidelines for cloud and shared resources.
- RCS has established a Free Tier and has some new students already using it.
- CUIT is planning to establish Linux servers for both education and research.
- RCS will continue a number of training courses on introductory Linux, HPC and scripting.

CUIT has received additional central funding for staffing, which has allowed for eliminating the service charge for Yeti (previously $250 per node per year).

Raj Bose, Director of Research Computing for the Zuckerman Mind Brain Behavior Institute, notes that an additional Intercampus Subcommittee recommendation was to evaluate connectivity between CUMC and Morningside, but this is being worked on with Hugh Ediet (Systems Biology) and CUIT. Hopefully there will be good news forthcoming soon.

Amy Nurnberger, Research Data Manager within the Center for Digital Research & Scholarship, expresses enthusiasm about new training courses as an outcome of the Education Subcommittee, and that the Libraries will be interested in taking part in these courses in the future. **Appendix III: Prior SRCPAC Recommendations & Actions Taken**

Cloud Pilot Report

The Cloud Subcommittee met in 2014, and discussed how the Cloud can fit into the University's computing strategy; the outcome was to create a Cloud Pilot in Summer 2015. CUIT established accounts, billing and security protocols, provisioning, necessary software, and managing instances.

RCS reports the following key takeaways:

- Critical to immediately establish a workflow, but hard to get started;
- Amazon Web Services is "overwhelming" with features and services, not clear where to start;
- Complex set-up and configurations; time consuming to support.

RCS will create its new Cloud Service in early 2016, with three levels of support:

1. Advisory Services (giving recommendations);
2. Help with Start-Up (account, billing, cost explanation; image and software customization; running instances; data storage and transfer);
3. Provide Ongoing Support (similar to how HPC is currently supported, plus maintaining machine image).

Researchers are responsible for paying for all Amazon services used. CUIT's support will be provided at no cost to the researcher.

Kathryn asks about unlimited support, as there are certainly staffing limits. Rob replies that RCS reserves the right to cap the number of users supported.

Alan Crosswell notes that it is against University policy to use PCard for Amazon Web Services, so CUIT is configuring a process for paying with a ChartString. **Appendix IV: Cloud Pilot Report**

Secure Data Enclave

Julien Teitler, Associate Professor of Social Work and Sociology, gives a brief history of the Secure Data Enclave (SDE): Established a few years ago as a pilot to allow researchers to use their desktops as dumb terminals to remotely access secure data, as opposed to physically going to a secure room to access from a machine disconnected to the rest of the Columbia system, which caused time and space constraint problems. SDE is not a substitute for HPC, as it was created to respond to a different need; it is not equivalent in power. The system has passed CUMC-IT HIPAA certification, and met all data provider requirements except for Department of Education (as the DoE will always require that its data be accessed on secure and non-internet connected computers). An RCS-supervised person is embedded in the CPRC/PER/Economics/ISERP group.

Julien's recommendation is to roll out the enclave to make it available to all researchers. Costs have not been finalized, although his understanding is that CUIT will continue to cover the majority of operational/system administration costs, while faculty and schools will need to find funding for licensing

software and acquiring hardware. Julien prefers to not require small-scale individual users to pay additional costs as he feels it will dramatically curb use.

Julien asks committee for feedback and suggestions for how the SDE gets expanded and made available to new users. Raj asks what agencies provide data, and whether PIs from other disciplines could use this resource as well. Julien replies that most data contributors are government agencies or universities that produce data with individual or neighborhood identifiers, and/or with personal health information… data that must be accessed only under restricted conditions.

Ryan asks about technical conditions for securing data. Julien explains that data is physically housed in the same place as financial data, within the most secure room on campus. Data is brought over by a Data Security Officer; the SDE's software is the same as what is used at New York Presbyterian Hospital, which creates a secure tunnel from central server to any secure computer. Data cannot be downloaded to a local hard drive, but the data Security Officer moves data temporarily from secure server to a shared folder that requesting researcher can access. Turn-around time from request to access has been under 24 hours.

Julien notes that the pilot is still open, and Rob comments that approval to join the pilot is done on a case-by-case basis depending on each user's needs. Sharon asks about personal health identifier information; Julien reminds the committee that SDE is HIPAA certified. Julien hopes to meet the needs of any Morningside PI requiring sensitive information.

Rob hopes to roll the pilot into a regular service by halfway through 2016.

Education

Kathryn then asks Rob to outline available educational services for computational research. Everyone is in agreement that there is a need for more resources for educating in computational methods, ranging from Intro to Linux and understanding command line and learning basic scripting, up to discipline-specific methods. Thus, Rob identifies current activities across campus (although not exhaustive):

- Application Development Initiative: Student group for programming (weekly Cookies and Code)
- Columbia Data Science Society
- Libraries/RCS Workshops (each class has between 10-20 attendees)
- RCS Presentations and Information Sessions at special request (email rcs@columbia.edu)
- HPC Education & Free Tiers (only one dozen users at current)
    - In Spring 2016, RCS will work with a Genomics & Bioinformatics undergraduate course, which will utilize the Education Tier. If any faculty are teaching courses that would benefit from using the Education Tier, email rcs@columbia.edu.
- Summer Workshops not currently taught, but could be, with instructors from external organizations (funded through national programs, so no financial commitments from researchers). If this service would benefit anyone, email rcs@columbia.edu.
- Many courses across the larger University.

Rob Cartolano, Associate Vice President for Digital Programs and Technology Services within the Libraries, notes that we have an institutional subscription to Lynda.com, which offers an endless number of online training courses in computational methods, and researchers are strongly encouraged to utilize this subscription. For example, there are 27 Python courses on Lynda. Ryan agrees, but notes that face-to-face instruction is always preferable to online, in case questions arise.

RCS is currently working with the Libraries, EVP for Research, and SRCPAC to gather and share all information about training opportunities, then coordinate combined efforts at helping to build these skills.

Ryan Abernathey expresses the need that many faculty are facing to bring students up-to-speed in coding skills in order to teach disciplinary courses; in effect, the first few weeks of a disciplinary course are spent teaching coding. For graduate students he teaches in the Department of Earth & Environmental Sciences, he offers a full-day workshop or assembly of shorter sessions on Python. Demand has been huge (100 participants), although it is not a formal course, and thus does not satisfy Ryan's departmental teaching requirements. Ryan encourages filling the void in base-level programming and coding; this would be a huge value to students and enable Faculty to focus on the intended subject. **Appendix V: HPC Education Resources**

Electronic Lab Notebooks

Kathryn asks Maneesha Aggarwal, Executive Director of Academic IT Solutions within CUIT, to discuss the introduction of a new service, Electronic Lab Notebooks (ELN), to allow digital recording of lab activities. While this software is cost-prohibitive at an individual level, CUIT is making it a central service that it will lease out to researchers. Interested labs may contact Maneesha at maneesha@columbia.edu.

Harmen Bussemaker, Professor of Biological Sciences and Systems Biology, asks whether there are overlaps with other project management tools like Slack. Maneesha notes that ELNs fulfill a different and more robust service, such as HIPAA compliance, data entry tracking, and notetaking. Some will connect to instruments to download data automatically.

Research Computing Services

Kathryn then introduces Gaspare LoDuca, Vice President of CUIT and Chief Information Officer, and asks him to discuss the increased commitment to supporting research. In addition to the support of the HPC clusters, the workshops, the establishment of the Secure Data Enclave, exploration of Electronic Lab Notebooks, CUIT has recently hired two permanent administrative staff in order to further accommodate research needs, and has additionally set-aside money for hardware purchases and services. Additionally, CUIT's communications staff will devote more time to publicizing services in order to increase usage (including a recent poster displayed on College Walk).

Gaspare reports regularly to the Board of Trustee's Physical Assets Committee, and will further make the Committee aware of advances in the University's research computing infrastructure (as he did first in March 2015). Trustees are very excited about recent advancements, and believe this is an important endeavor for the future. Thus, as the research computing strategy evolves and expands (for example, to the Cloud), it is critical for CUIT to provide centrally-funded support at all stages.

Gaspare hopes to very soon sign a contract with Amazon Web Services and Procurement, which will allow the setting-up of Direct Connect.

Research Computing Services 2.0

Kathryn reviews how far the shared HPC research effort has come, and how researchers across the University are increasingly more frequently incorporating serious computing in their discovery research, as illustrated by the Data Sciences Institute. Kathryn met with faculty to understand how peer institutions are running research computing services, and whether Columbia can and should develop and evolve RCS in the future.

Both to make the Columbia community more aware of the resources and services to support research and to establish a reputation that facilitates recruitment, and to support grant proposals that want to reference research computing resources, Kathryn has been having a series of conversations to discuss how we can potentially elevate and publicize the shared computing resources. Kathryn would like to (re)brand the group, although any rebrand cannot include the words *Center* or *Institute*.

Kathryn asks that any brainstorms or wishlists be sent to her at kvj@astro.columbia.edu. The relaunch will likely coincide with the launch of new HPC cluster in 2016.

Raj notes it is "amazing" that so much has happened so recently, including two new Institutes – Zuckerman and Data Science – and how we are primed for future and continued successes. Multiple committee members echo this congratulatory remark, with special focus on the relationship between rebranding and fundraising initiatives.

Ryan suggests dropping 'services', and focusing on what the computing services enable: how research computing is vital to allow for accomplishments in X, Y, and Z. Identifying research highlights would be a critical thrust in a rebrand. Kathryn notes that the relaunch will include a symposium, which will highlight accomplishments, but that more can be done.

Sharon Sputz, Director of Strategic Initiatives within the Data Science Institute, suggests 'Facilities', which committee members enjoy. Gaspare does not like 'collective'.

Rob reminds everyone to report publications emerging out of Yeti use. Either Rob or Marley Bauce, Manager of Research Initiatives within EVP for Research, will contact researchers every semester requesting new citations.

Kathryn thanks everyone for their participation and adjourns the meeting.

Catalog of Appendices

- Appendix I: Fall 2015 Yeti Operating Committee Update
- Appendix II: HPC Expansion Schedule
- Appendix III: Prior SRCPAC Recommendations & Actions Taken
- Appendix IV: Cloud Pilot Report
- Appendix V: HPC Education Resources