COLUMBIA|RESEARCH
Shared Research Computing
Policy Advisory Committee

Shared Research Computing Policy Advisory Committee – Spring 2018 Meeting
April 16, 2018

Attendees: Chris Marianetti, Gaspare LoDuca, Marc Spiegelman, Tian Zheng, Kyle Mandli, Lorenzo Sironi, Bob Mawhinney, Alan Crosswell, Victoria Hamilton, Tom Chow, Maneesha Aggarwal, Rob Lane, George Garrett, Sander Antoniades, Michelle Benson, Peter Jorgenson, Rob Cartolano, Raj Bose, Michael Weisner, Mark Newton, Dali Plavsic, Jochen Weber, Marley Bauce, Michael Shelter, Halayn Hescock (Phone)

Chris Marianetti, Chair of SRCPAC, opens the floor by welcoming all attendees to the Spring 2018 SRCPAC meeting, and overviewing today's agenda. He then asks all attendees to introduce themselves.

Habanero Update

Kyle Mandli, Chair of the Habanero Operating Committee, thanks the Committee and introduces George Garrett, Manager of Research Computing Services. Habanero is comprised of four racks of computing nodes, plus power and cooling nodes. This is located in the basement of the Green Science Center, kindly offered by the Zuckerman Institute.

There are four ways to participate in HPC at Columbia: Purchasing their own nodes; renting nodes owned by CUIT; utilizing a one-year free tier (for $1,000); and hosting course-based research on an education tier.

Habanero launched in 2016 with 222 nodes; last year, Habanero expanded to add 80 nodes, totaling 302 nodes. Habanero also saw a storage expansion earlier in 2018 – 20 more disks, equating 100TB of usable storage. The order has already been placed with the vendor, DDN.

The Scheduler has been upgraded, and a new test queue has been added. Finally, CUIT is piloting Jupyterhub and Docker – any interested participants should email rcs@columbia.edu.

Shared HPC has continued to grow at Columbia, now including 44 groups (or, 1,080 individual users), who have altogether completed 2.1M jobs. In terms of core hours, since Habanero's December 2016 launch there has been a great increase in system utilization (partly due to Yeti Round 1's concurrent retirement). Monthly utilization is 80-94% of system capacity, making it highly utilized!

Business rules are established and governed by the Habanero Operating Committee, which meets every semester to review rules and vote on new proposed policies. Any Habanero user is welcome to attend these meetings. Any special requests should also be brought to this Operating Committee, and can be sent to kyle.mandli@columbia.edu.

CUIT offers multiple support services: The simplest is to email hpc-support@columbia.edu. Office hours are held on the first Monday of every month, and RCS can arrange information sessions for specific groups/departments.

RCS and the Libraries collaborate through running a series of 90-minute workshops on Linux, Scripting, and HPC. The next workshop series will be in October 2018, but the schedule has yet to be finalized.

The prior cluster – Yeti – has had its first portion retired in November 2017. 66 nodes are still in production, with modest utilization, set to retire in March 2019.

New Cluster Update

An RFP and Design Committee has been established to identify a new HPC system's specifications and guide RCS in the purchasing process; a full list of Committee members can be found on the corresponding presentation slide.

In Early-April, the Committee selected four finalist vendors, who will soon be visiting campus to provide in-person presentations. By the end of April, the Committee will select one compute and one storage vendor, then will begin accepting researcher orders from mid-May to mid-June; RCS will distribute a preview of this announcement shortly. The new cluster should go live in November.

The new cooling expansion project will allow Morningside's Data Center to accommodate new machines. This project has initiated with substantial support from CUIT and Morningside Deans, thus allowing us security for at least the next five years. Alan notes this builds upon the initial NIH grant that got us the initial power upgrade, which is good.

The new cluster has multiple specifications, which may change due to pricing and other factors: Three types of nodes (Standard, High-Memory, and GPU)… all nodes will feature Dual Skylake Gold 6126 processors, with very advanced optimization features.

Important to also note is that this cluster will now have a lifespan of five years (!) – one year more than the historical four-year lifespan.

Very excitingly, the new cluster will be named **Terremoto**, which means *earthquake* in Spanish and Italian. Initially 70+ names were submitted, and the collective community refined the list down to this final choice. The Committee expresses enthusiasm for this naming convention.

Assessing Post-Purchase Demand

Chris explains that when researchers have new grants, they typically do not come in right in May or June when CUIT typically announces its expansion round. It is feasible for the University to buy nodes then have them sold back to the researcher at later dates, but we need to know whether this occurs and how often. If anyone knows that researchers have high demand for node purchase throughout the year, please do send comments and suggestions to chris.marianetti@columbia.edu.

Marc notes that frequently Columbia will hire a senior faculty member who has substantial node requirements later in the year, which makes it especially difficult to establish. Bob says that a No Cost Extension agreement with vendors is standardly typical at Brookhaven National Laboratory (which is compliant with the Department of Energy's requirements). Chris suggests a policy that if someone needs to buy over "X" number of nodes, we can go through an NCE process – this does not allow users to change any specifications, and they would need to purchase the exact same kind of system as governed in the master contract.

Rob explains two distinct scenarios: Buying new hardware, and establishing a resale arrangement. For the latter consideration, this may be difficult for financial reasons. Chris suggests we ask Deans to purchase a few nodes apiece for a pilot program, then ascertain the program's utility. Rob also suggests we establish a No Cost Extension agreement as part of our current negotiations with the vendors (to Bob's point).

Chris notes that pre-purchased nodes – even if not bought back by new researchers – could thereafter be made available for public use as a way of bolstering the University's investment in data science. Bob suggests we develop the "pitch" for how these nodes would be used. Chris suggests we collect numbers on usage for the Free and Education Tiers, which may help craft an argument for this arrangement. Tom suggests we ask for a specific number and a specific financial commitment, which will stand a better chance of gathering the schools' agreements.

Bob additionally suggests a Junior Faculty/Graduate Student Computational Award, which new faculty can assign nodes to for demonstrated expertise. EVPR and RCS will speak further about developing such a competition, although the resources would need to be purchased by schools or CUIT.

Rob suggests that for this vendor agreement, we buy some extra switches in order to accommodate off-cycle purchasers, which would have to be paid for upfront. Chris agrees this will be beneficial to do now.

Chris believes that we should formalize an agreement soon in order to accommodate new requests. RCS and EVPR will speak further about how to do this.

Update on Training Subcommittee

The Subcommittee contains 15 members – many on the larger SRCPAC committee – and is tasked with identifying what currently exists at Columbia by way of training in data science, ascertaining graduate student demand, and developing recommendations for new supportive programs. The committee has met three times so far, will meet twice again in the Spring semester, and has distributed surveys to all Morningside graduate students and 12 Morningside departments.

As of today, 208 graduate students have participated in the survey, with 6 departments confirmed and 6 departments in-progress. Marley and Tian will speak further about increasing Statistics involvement in both surveys.

Students had interesting responses judging by preliminary analysis: By far, students want more training in Python; students express a wide variance in proficiency in Python, although very few people think they are experts. Similarly, very few people know anything about Cloud computing; people know quite a bit about Excel. It is therefore a problem whereby students are computationally-minded by way of Excel, but cannot breakthrough into other computational technologies.

Students were also asked about demand for formats of informal training, with a clear preference for pre-semester boot camps, and not much interest in mediated self-study. Most students (over 60%) were not aware of campus-based resources in teaching computation or data science.

Students were asked about competitor institutions that did well at teaching students in computation, and they broadly identified MIT, Berkeley, and Johns Hopkins. Gaspare has offered making connections with the Chief Information Officers at these schools.

Marc then reviews a few cherry-picked qualitative comments submitted by students, which can be found on the corresponding presentation slides.

Marc then explains that his Subcommittee has brainstormed a number of new programs to help meet expressed demand:

- A pre-semester boot camp for incoming graduate students
- Developing an institutional partnership with Software Carpentry for ongoing workshops
- Having a Distinguished Lectures series from industry representatives
- Curating online training modules
- Help desk for research computing (Gaspare likes this option)
- A full-time staffer to coordinate this effort

Marc then asks whether SRCPAC has other suggestions for programmatic offerings; anyone with ideas should email mspieg@ldeo.columbia.edu or mb3952@columbia.edu.

Tian notes that peer institutions have "peer office hour" offerings, whereby graduate students in data science make themselves available to meet with other students to offer advice, perhaps for pay. The University of Washington has asked Data Science faculty to shift their office hours into the Data Science space and make themselves available for all students.

Jochen suggests that we can gather further interest from SIPA and/or Business in developing novel funding mechanisms for these new efforts.

CUIT Updates

George and Sander begin by sharing the status of the Secure Data Enclave. The SDE was initially sponsored by the Population Research Center, and is now in the final security audit stage before being rolled-out for general consumption. Historically, SDEs required physical access to one room and a computer without internet access; now, we have a screen share to allow individuals to use the service remotely without requiring physical in-person presence. Jochen asks the data quantity for typical use cases, to which Sander responds that the typical amount is in the gigabyte-stage; the default is 60GB, but this can be expanded due to project needs but with an increased storage cost.

Globus is a service used since Hotfoot, which allows data transfer between defined end-points; Columbia is expanding its licensing to include features requested by researchers, such as publishing via Globus' web service for many TBs of data. Jochen will reach out to Sander offline to ask for more information.

Maneesha Aggarwal announces a newly-formed Emerging Technologies group, which meets the third Friday of every month, to share internal activities by way of emerging technologies such as AI, VR, machine learning, block chain, etc. Because we are so decentralized, it is difficult to communicate resources; the group has grown to over 100 members (faculty, students, and staff across all campuses). The group also hopes to engage with corporate partners such as Intel, Microsoft, and HP. These industries offer stipends to encourage research in new emerging areas. Any interested individuals who wish to participate in this group can email mb3952@columbia.edu.

Peter Jorgensen then explains the University's enterprise agreement with Amazon Web Services, signed in November 2015. There are many benefits to using this agreement as opposed through establishing an independent contract (technically it is also required for compliance with the University's financial policies…).

There are linked accounts – for accounts already created – that can be transferred over to ARC-based billing, or delegate accounts for new requests. Root credentials are stored and maintained by CUIT, which provides advanced security, while researchers maintain authority and control of the account.

Amazon can hold Personal Health Identifying Information through the Business Associate Agreement, but this requires active opting-in through the account; not all AWS services can be covered. All relevant Columbia and CUMC policies remain in effect, such as RSAM and securities registration.

Direct Connect's contract was just signed (it took 20 months!), with many features getting piloted currently: This provides us with a direct connection to the US-East-1 Region, which allows us to directly access campus information from off-campus even without a public IP address; this is an important feature for individuals who have information that cannot be publicly disclosed, which allows for great security advancements.

The Global Data Egress waiver is being rolled-out with Amazon for research purposes: They will waive the cost of research data egress up to 15% of the monthly billings per account; they expect this to cover almost 100% of our data egress costs, unless we are doing something extremely heavy.

The two Amazon Web Services links will soon be added to SRCPAC's webpage, and soon will develop a large email announcing this agreement.

Publications Reporting

Any research publications emerging out of Habanero- or Yeti-hosted projects should please (please) a) acknowledge Columbia HPC in the manuscript, and b) report citation information to srcpac@columbia.edu. This information is critically important to appraising the RCEC of progress, and demonstrating the utility of further investments.