ARISTOTLE
CLOUD FEDERATION

CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation

Science Team Advisory Committee (STAC)

3/20/2017 Meeting Minutes

Adam Brazier, Aristotle science team lead
brazier@cornell.edu

Invited to 3/20/2017 Science Team Advisory Committee Meeting *(attendees italicized)*

Cornell University (CU):
James Cordes cordes@astro.cornell.edu
Adam Brazier brazier@cornell.edu
Angela Douglas aes326@cornell.edu
Nana Ankrah na423@cornell.edu
Brandon Elam Barker brandon.barker@cornell.edu
Sara C. Pryor sp2279@cornell.edu
Tristan Shepherd tjs346@cornell.edu
Patrick Michael Reed pmr82@cornell.edu
Julianne Dorothy Quinn jdq8@cornell.edu
Resa Reynolds rda1@cac.cornell.edu
Susan Mehringer shm7@cornell.edu
Paul Redfern red@cac.cornell.edu
David Lifka lifka@cornell.edu

University at Buffalo (UB):
Tom Furlani furlani@ccr.buffalo.edu
Varun Chandola chandola@buffalo.edu
Cristian Tiu ctiu@buffalo.edu
Dominik Roesch drosch@buffalo.edu
Brian A. Wolfe bawolfe@buffalo.edu

University of California, Santa Barbara (UCSB)
Rich Wolski rich@cs.ucsb.edu
Andreas Boschke andreas@cs.ucsb.edu
Kate McCurdy kate.mccurdy@lifesci.ucsb.edu

National Science Foundation
Amy Walton awalton@nsf.edu

#1541215

NSF - *Amy Walton, NSF program manager*
- The 1st NSF Data Infrastructure Building Blocks PI Workshop (https://dibbs17.org) was led by Cornell as a supplement to the Aristotle award and was a tremendous success. There was high energy on the part of the 72 participants and the resulting data and insights produced will be invaluable to NSF as we plan for the future of research and data cyberinfrastructure. The final workshop report will be available by the end of March. Thank you all!
- Your 18-month review is scheduled for the morning of April 25th. Focus is on a working prototype and showing you are well on the way to success. The goal for these projects was to provide a robust and shared CI capability. Plan on 60-90 minutes of top level presentations/demos followed by an extensive questions/answer and discussion period through late lunch. Cover 5 areas: (1) rationale for the project, (2) CI approach used and how the prototype is doing, (3) project management, i.e., who is doing what and how it is being measured, (4) what are the next stages of the project, i.e., if we fund you for years 3 and 4, what are you going to accomplish? (note: this could include the possibility of adjusting milestones for new innovations or broader impacts), (5) when this award is completed, what will the community have that it doesn't have now? One week prior to April 25th, NSF should receive a reference package on the project with the PowerPoints, a copy of the execution plan, any report updates. This will be distributed by NSF to the panel. Hit the highlights. Explain what the prototype is, what is working, how what you have is helping the scientific community. Provide a well-organized package. Including some key application examples seems to work for some projects.

Infrastructure Update – *Adam Brazier (CU)*
- We're moving to Eucalyptus 4.4. We were successful in getting HPE to build OAuth2 support into 4.4 and testing it. This will provide us much more functionality. You will be able to login using your own institutions' authentication as well as InCommon. If you're using XSEDE resources, this is what they are using too, so it will be like one ticket access to potential resources.
- Year 2 hardware installations and purchases are underway. The University at Buffalo has added 112 cores and is upgrading their network. Cornell is adding Ceph storage since that was a greater need than more cores. UCSB, on the other hand, has ample storage, so they're adding cores since usage of their current cores is maxed out. Heterogeneous platforms and software at the various sites is a plus since it provides users with more alternatives to address their needs.

Portal Update – *Susan Mehringer, portal team lead (CU)*
- The Aristotle portal has been up for quite a while and is located at https://federatedcloud.org. If you have comments or suggestions, send them to help@federatedcloud.org. A status graph for all 3 sites is on the portal which shows the availability of cores. Monthly, quarterly, and annual reports are also available. They are password protected, so contact us via "help" if you'd like access.
- OAuth2 has been implemented on the portal, and will be used for all sections and functionality that require authentication.
- The next major step is to integrate allocations and accounting information into the portal, by project and by person. Information will include, e.g., project members, usage, allocations balance, and storage usage.

Science Team Update - *Adam Brazier, science team lead (CU)*
- We're using Slack for communications which has been incredibly useful given the distributed nature of our project. If you are not a member of our Slack channel, email brazier@cornell.edu and ask to join. Our RT ticket tracking system is another valuable tool for science users (help@federatedcloud.org).
- Allocations are somewhat relaxed at this point; a formal allocations process will be implemented in the future.

#1541215

- We're close to completing the development of a virtual cluster capability for MPI and OpenMP (we just need to improve the disc I/O speed). The intent is to provide virtual cluster access for modest size jobs (not thousands of cores). On-demand access to virtual clusters in the cloud vs. waiting in a queue for a very large-scale system will improve time to science for users with modest needs.
- We applied for and we're approved for 2 million virtual core hours on Jetstream.
- Science teams will be asked to provide science highlights for our upcoming 18-month NSF review. Lifka emphasized the importance of highlighting how the Aristotle project has helped (or will help) the science teams do things differently or better than in the past or, in some cases, with faster time to science.

Individual Science Team Updates

Use Case 1: A Cloud-Based Framework for Visualization and Analysis of Big Geospatial Data - *Varun Chandula update (UB)*
- We plan to add more analysis tools to our framework, add more simulation data, and make the system more robust.
- Next steps will include discussions about how to submit/track jobs which will be important in opening up the capability to the broader community.
- We can provide a demo for the 18-month review.

Use Case 2: Global Market Efficiency Impact - *Dominik Roesch (UB)*
- We've set up a framework for the analysis of high frequency trading data (20TB+) and will provide this capability to 3 PhD researchers this summer with the future goal of opening up this capability to researchers at other institutions in the future.

Use Case 3: High Fidelity Modeling and Analytics for Improved Understanding of Climate-Relevant Aerosol Properties - *Tristan Shepherd (CU)*
- We're had a productive collaboration with CAC and are very close to running WRF-Chem (a parallel version Cornell built for the cloud) on multi-year, high-resolution data across North American.

Use Case 4: Transient Detection in Radio Astronomy Search Data – *Adam Brazier (CU)*
- We're preparing for cloud usage and will do early runs on Jetstream. Cornell and a global team of astronomers previously uncovered the source of a "fast radio burst." We plan to expand the analysis of that data beyond the previously analyzed data to explore and, hopefully, produce new insights.

Use Case 5: Water Resource Management Using OpenMORDM – *Julianne Quinn (CU)*
- We're working to benchmark and test the scaling of OpenMORDM and its equivalent in the Python equivalent Project Platypus. Scientific output will commence when the MPI virtual cluster work is completed.

Use Case 6: Mapping Transcriptome Data to Metabolic Models of Gut Microbiota – *Brandon Barker (CU)*
- Several large instances were created (Windows and Linux) and used for modeling and genetic analysis. This resulted in a paper which will be published in the *Journal of Bacteriology* and a presentation that will be delivered at the *Data-Driven Biotechnology Conference* in Copenhagen (May 7-11, 2017).
- Computational development continues on the modeling, and usage is expected to ramp up when we're ready to model a larger system based on the initial exploration studies currently underway.

#1541215

Use Case 7: Multi-Sourced Data Analytics to Improve Food Production – *Kate McCurdy (Sedgwick Reserve/UCSB)*

- Agricultural soil moisture monitoring and drought monitoring of oak trees will continue when soil sensor failures due to excessive moisture as remedied.
- The "Where's the Bear" project will continue when storm-impacted sensors are back up and running. This system is still using the Caffe framework and the Aristotle cloud to process all camera data from the field. The framework sorts images into species. We are still struggling with empty frames and frames with birds and will be honing this species recognition capability. We are using a citizen science site (Zooniverse) to pull data out of the Aristotle cloud and will be using volunteer crowdsourcing to verify the accuracy of the computer-generated data in the coming weeks. Ecology researchers are interested in using the first workflow to publish climate change impacts on wildlife. UCSB Computer Science is working with us to eventually launch our own repository site to manage the whole process of data acquisition, storage, and dissemination. Important note: this project has sparked a new scientific collaboration (not part of the original Aristotle proposal) with Sedgwick researchers who plan to perform latent species identification and analyze fishery health from images from SE Asia.

NSF
#1541215