ARISTOTLE
CLOUD FEDERATION

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

Science Team Advisory Committee (STAC)

4/1/2016 Meeting Minutes

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

Invited to 4/1/2016 Science Team Advisory Committee Meeting *(attendees italicized)*

Cornell University (CU):
James Cordes cordes@astro.cornell.edu
Shami Chatterjee shami@astro.cornell.edu
Adam Brazier brazier@cornell.edu
Angela Douglas aes326@cornell.edu
Bessem Chouaia bc335@cornell.edu
Nana Ankrah na423@cornell.edu
Brandon Elam Barker brandon.barker@cornell.edu
Sara C. Pryor sp2279@cornell.edu
Patrick Michael Reed pmr82@cornell.edu
Bernardo Carvalho Trindale bct52@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@ccrbuffo.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

Meeting Purpose - *Adam Brazier, science team lead (CU)*
Get feedback from the scientists to ensure that the Aristotle project team is responsive to their needs. Key aim: reduce time to science.

Introduction - *Amy Walton, NSF program manager*
Amy explained that Dave Lifka (PI) and Tom Furlani (Co-PI) are not at today's Science Team Advisory Committee meeting because they were at NSF this morning (4/1/2016) giving an invited talk on the Aristotle Cloud Federation. This project is of great interest to NSF. Program directors from multiple directorates and Coalition of Academic Scientific Computation (CASC) members attended the meeting and interest was very high. The talk included your 7 science use cases and why your participation is important in making sure the federation works. NSF made only 4 DIBBs awards and this one was funded because it addressed 2 important concerns of the NSF: (1) creating sustainable models for cyberinfrastructure facilities, (in this case, sharing resources), and (2) creating metrics so that researchers can quickly find and value the computing and data analysis resources they need. Adam Brazier, Aristotle Science Team lead, added that NSF is looking for progress in the science use cases in the project's monthly reports.

Infrastructure Update - *Resa Reynolds, infrastructure team lead (CU)*
- Cornell's year 1 hardware up and running. 168 cores (6GB/core) and 120TB SAN storage were added to our existing Red Cloud infrastructure. Images are being created by our science teams.
- UB's hardware is in place and they'll be installing the Eucalyptus cloud software stack in the coming weeks
- UCSB's hardware started to arrive and will continue to arrive over the next few weeks. By the end of month, it will hopefully be up and running.
- As a team, our goal is get the infrastructure up and get out of way of the science
- Each year we'll be adding more hardware. If science teams have specific requirements, e.g., more RAM to core, etc., they should let us know. Adam suggested using Slack (the recently launched science team communication platform) to communicate needs as well as lessons learned.

Portal Update - *Susan Mehringer, portal team lead (CU)*
- Purpose of the user portal is to get scientists the systems info they need, including how to get to different systems in the federation
- Features will include: allocations info, systems status (how busy each node is), science project info, user documentation/training and eventually OpenXDMoD and QBETs for advanced resource prediction
- PIs will have access to overall usage data, team member info, how much allocation is left, etc. Project team members will see a subset of this info.
- Current portal status: the federatedcloud.org site is up with a "coming soon"/latest news page. We added SSL. We will be adding InCommon for authentication. The database schema has been designed and built. UB provided us a rest API to share data between the federation sites. We're using a simple open source framework for the user portal so it's easy to reproduce and a model for others who wish to start their own federation or possibly join ours. Adam commented that the goal of the portal is to make the federation easy to use so researchers can focus on their science rather than administrative tasks.

Science Team Update - *Adam Brazier, science team lead (CU)*
- Science Use Cases are moving! Expect different rates as each project is different and also available effort is at different stages of readiness.

#1541215

- One element of Use Case 2, “Global Market Efficiency Impacts,” requires larger volumes than currently available on Lake Effect at Buffalo, so it will run at Cornell (first instance of cross-federation deployment)
- Cornell is committed to having a parallel MATALB capability and will make MATLAB Distributed Computing Server (MDCS) available on the CU node of Aristotle with 64 worker processes. This and other heterogeneity across the federation, in licensed software, installed software, and in hardware (heterogeneity may evolve over time), will be advertised to researchers.

Individual Science Team Updates
Note: science team project descriptions from the proposal were distributed to meeting invitees as a separate pdf

Use Case 1: A Cloud-Based Framework for Visualization and Analysis of Big Geospatial Data - *Varun Chandula update (UB)*
- Our goal is to create a framework and interface to handle all sorts of geospatial data and eventually remote sensing data. Users will submit analysis and visualization jobs to the cloud and at the back end requests will get translated into dynamically created cluster jobs that fire up, run, and are taken down after the users get their results.
- One of our students developed the app based on Spark. We're testing it now using UB's current cloud (called Lake Effect) so we will be ready to migrate to Aristotle as soon as it's available. Our interface is working well; we're focused on the back end plumbing now. We hope to have the first cut of our working use case in a month or two.
- We’re eager to share with and learn from the other use cases. The less we reinvent, the better. We can learn from each other.

Use Case 2: Global Market Efficiency Impact - *Dominik Roesch, Brian Wolfe updates (UB)*
- *Roesch:* We'd like to look at efficiency in stock prices and derivations from fundamentals across international markets, however we need access to international market data which will cost ~15K. We'll discuss with Tom Fulani to inquire about possible funding. Otherwise, our alternative plan is to bring high frequency data to UB this summer and develop a framework for it.
- *Wolfe:* Varun developed an image for us and we have access to a couple of working ssh ports. We're ready to roll. We just need some software to do image recognize on pdf images.

Use Case 3: High Fidelity Modeling and Analytics for Improved Understanding of Climate-Relevant Aerosol Properties - *Adam Brazier update (CU)*
- This project is focused on using NASA satellite data and requires WRF-Chem software. Researchers have had a variety of problems getting this software installed on various clusters around the world because it's large and complicated (really a "pig"). Cornell CAC consultant Steven Lee built a parallel version of WDF-Chem which we'll soon test running as a cluster of VM's on Aristotle. Cornell's large instance size (32 cores) may be good for this. A big advantage of the cloud is that once an image is built, we can spawn many duplicates for scalable processing as demand increases.

Use Case 4: Transient Detection in Radio Astronomy Search Data - *Shami Chatterjee (CU)*
- This is a rather hot research area; we just published a paper in *Nature*. We're looking for a single pulse in sea of noisy data that is very fast in time and frequency. We will try out a variety of approaches on Aristotle with our data which is local. Also, at the recent NSF NanoGRAV Physics Frontiers Center meeting, other institutions with radio telescope data expressed interest in using Aristotle. If we build it, they may indeed come.

#1541215

Use Case 5: Water Resource Management Using OPENMORDM -*Adam Brazier update*
- Bernardo Trindale is building an image now. Aristotle will be very helpful in developing their research apps and the potential to run full-size simulations (extremely large, thousands of cores) on AWS is particularly appealing because they may want to develop an AWS service for US municipalities.

Use Case 6: Mapping Transcriptome Data to Metabolic Models of Gut Microbiota - *Brandon Barker update (CU)*
- We plan to map transcriptome data to metabolic models using algorithms with widely varying run times. Advantages of Aristotle: (1) no need to deploy large capacity system for apps whose computing times aren't known up front, (2) we're able to spin up either Windows or Linux instances which is beneficial because in some cases the same software packages have slightly different features available in Windows vs. Linux. We have been investigating NixOS as a great way to develop reproducible scientific workflows, but will use Ubuntu and Windows for now. Bessem Chouaia said that eventually it may be helpful to have a very high memory and low CPU node.

Use Case 7: Multi-Sourced Data Analytics to Improve Food Production - *Kate McCurdy update (Sedgwick Reserve/UCSB)*
- We're busy working on our instrumentation--deploying cameras, agricultural sensors, Wi-Fi mechanisms to bait and spring traps, and testing drones. Cloud data collection will be important for projects such as irrigation management and control that uses sensing and analysis for native farm plants to better understand water use and the effect on yield. We're also doing deer population monitoring with a ground survey and we'd like to augment or offset that with image analysis.
- Our "Where's the bear" project will be trying to analyze 5 million digital photos and do some facial recognition or some kind of species recognition to monitor wildlife populations for predator protection. We're also using drones to monitor black birds (CA endangered species) with a fixed camera in a remote site which we'll try to tie into our mesh network. We're excited about the prospect of Aristotle helping us with our data and analysis needs.

#1541215