ARISTOTLE
CLOUD FEDERATION

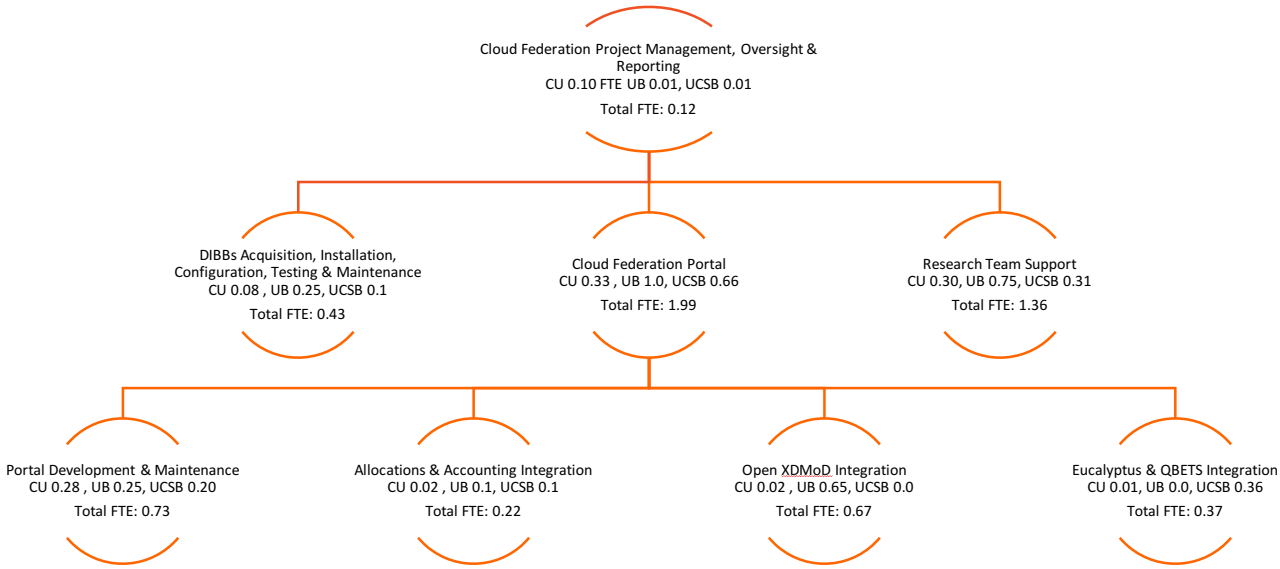
CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation

Monthly Report 1/31/2017

Report 16 of 18

Submitted by David Lifka (PI)
lifka@cornell.edu

This is the sixteenth required monthly report of the Aristotle Cloud Federation team. We report on plans and activities for each area of the project Work Breakdown Structure (WBS).



#1541215

Contents

NSF
#1541215

1.0 Cloud Federation Project Management, Oversight & Reporting Report

1.1 Subcontracts
All subcontracts are in place. Nothing new to report.

1.2 Project Change Request
No new project change requests were made this month.

1.3 Project Execution Plan
The Project Execution Plan (PEP) was approved by NSF on 12/18/2015. We are operating as planned and continuously updating our PEP on a monthly basis.

1.4 PI Meetings
Lifka was Chair of the 1st NSF Data Infrastructure Building Blocks PI Workshop (DIBBs17) January 11-12, 2017 in Arlington, VA. The workshop was organized and managed by Cornell as a supplemental to the Aristotle award.

1.5 Status Calls
1/17/2017 project status call topics:

- How to pursue optimal pricing for Globus endpoints.
- Reliability of Supermicro hardware (lose more discs per year with Supermicro; must RAID the head node if using Supermicro for storage).
- Use of Zooniverse as a Citizen Science repository.

2.0 DIBBs Acquisition, Installation, Configuration, Testing & Maintenance Report

2.1 Hardware Acquisition

- UB's 10G network equipment arrived but was the wrong equipment; the correct equipment is expected to arrive in 1-2 weeks. Four new node controllers were purchased from Ace Computers. These should arrive within a week when they will be installed, tested, and added to the UB cloud pool.
- Cornell will order Cepth storage in late January or early February, along with a new cloud controller and a 10G network switch to run the HPE Helion Eucalyptus 4.4. software stack with the Ceph storage backend.
- UCSB is assessing their compute and storage needs in preparation for year 2 purchases.

2.2 Software Installation, Configuration, and Testing

- Network bandwidth testing continues. Cornell opened a ticket with the XSEDE NOC and tests are underway between Cornell and UCSB, with troubleshooting assistance from the NOC.
- Cornell's test cluster is pulling down nightly Eucalyptus builds of version 4.4 which supports OAuth2. Both of the issues reported last month (i.e, login buttons and an outstanding bug) have been resolved. Testing of the early 4.4 release continues.

#1541215

There were no updates to the infrastructure planning table this month:

	Cornell (CU)	Buffalo (UB)	Santa Barbara (UCSB)
Cloud URL	https://euca4.cac.cornell.edu	https://console.ccr-cbls-2.ccr.buffalo.edu/	https://console.aristotle.ucsb.edu
Cloud Status	Production	Production	Production
Euca Version	4.2.2	4.3	4.2.2
Globus	Yes	Planned	Planned
InCommon	Yes	Yes	Yes
Hardware Vendor	Dell	Dell	Dell
# Cores	*168	**144	140
RAM/Core	4GB/6GB	up to 8GB	up to 9GB
Storage	SAN (226TB)	SAN (336TB)	Ceph (288TB)
10Gb Interconnect	Yes	10Gb inter-cluster; 1Gb external, 10Gb external planned	Yes
Largest Instance Type	28 core/192GB RAM	24 core/192GB RAM	16 core/16GB RAM
	* 168 additional cores augmenting the existing Red Cloud (376 total cores)	** 144 additional cores augmenting the existing Lake Effect Cloud (312 total cores)	

2.4 Potential Tools

- **CloudLaunch**

  The Cornell team continues to work on deploying a virtual cluster in Red Cloud with a generic compute node image for functional testing, including running sample jobs.

- **HPE Helion Eucalyptus**

  The Cornell team worked closely with the HPE Helion team to test their OAuth2 implementation; the 4.4. release is scheduled for mid-February.

- **Supercloud**

  Nothing new to report this month.
  .

#1541215

3.0 Cloud Federation Portal Report

Content updates to the project are ongoing: https://federatedcloud.org.

The usage graph (https://federatedcloud.org/using/federationstatus.php) was completed last month; it shows basic early usage data from all 3 sites. For ease of conformity between federated sites, UB has provided an updated REST API code that now reports time in UTC/GMT; this has been implemented at one site, and is expected to be implemented at all three sites next month

There were no changes to the portal planning table this month:

Portal Framework			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 3/2016	**4/2016 – 12/2016**	**1/2017 - End**	**1/2017 - End**
Gather portal requirements, including software requirements, metrics, allocations, and accounting. Install web site software.	Implement content/functionality as shown in following sections. Add page hit tracking with Google Analytics, as well as writing any site downloads to the database.	Implement content/functionality as shown in following sections. Add additional information/tools as needed, such as selecting where to run based on software/hardware needs and availability.	Release portal template via GitHub. Update periodically.
Documentation			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 3/2016	**4/2016 – 10/2016**	**11/2016 – End**	**1/2017 - End**
Basic user docs, focused on getting started. Draw from existing materials. Available through CU doc pages.	Update materials to be federation-specific and move to portal access.	Add more advanced topics as needed, including documents on “Best Practices” and “Lessons Learned.” Check and update docs periodically, based on ongoing collection of user feedback	Release documents via GitHub. Update periodically.
Training			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 3/2016	**4/2016 – 12/2017**	**4/2017 – 12/2017**	**1/2018 - End**
Cross-training expertise across the Aristotle team via calls and 1-2 day visits.	Hold 1 day training for local researchers. Offer Webinar for remote researchers. Use recording/materials to provide asynchronous training on the portal.	Add more advanced topics as needed. Check and update materials periodically, based on training feedback and new functionality.	Release training materials via GitHub. Update periodically.

#1541215

User Authorization and Keys			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 1/2016	**2/2016 – 5/2016**	**6/2016 – 3/2017**	**1/2017 – End**
Plan how to achieve seamless login and key transfer from portal to Euca dashboard.	Login to the portal using InCommon.	Beta testing Euca 4.4 with Euca console supporting Globus Auth. Will deploy and transition to Euca 4.4. on new Ceph-based cloud.	Move seamlessly to Euca console after portal Globus Auth login.
Euca Tools			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 3/2016	**4/2016 – 12/2016**	**1/2017 – End**	**1/2017 – End**
Establish requirements, plan implementation.	No longer relevant since Globus Auth will let us interface with Euca web console	N/A	N/A
Allocations and Accounting			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 3/2016	**3/2016 –3/2017**	**3/2017 – 6/2017**	**6/2017 – End**
Plan requirements and use cases for allocations and account data collection across the federation. Design database schema for Users, Projects and collections of CPU usage and Storage Usage of the federated cloud.	Implement project (account) creation in the database and display on the portal. Integration hooks for user and project creation/deletion and synchronization across sites.	Automate project (account) creation by researcher, via the portal.	Report on usage by account, if the researcher has multiple funding sources. Release database schema via GitHub.
Metrics and Usage			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 7/2016	**7/2016 – 9/2016**	**10/2016 – 3/2017**	**10/2017 - End**
Buffalo team utilize Cornell scripts to design a REST API for basic cloud data and deploy at 3 sites and publish usage data to project portal (completed). Buffalo currently standardizing the API by using UTC across the sites and refactoring the code efficiency.	Provide documentation for installing XDMoD and SUPReMM at individual sites. Install Open XDMoD/SUPReMM at individual sites and begin data collection. This includes installation of SUPReMM and the data collection piece at the federation sites. Begin integration with federated authentication providers.	Federated data collection will ship data from XDMoD instances at the individual sites to a master XDMoD instance at UB where overall cloud data will be displayed. This is in alpha testing at UB with completion planned for 3/2017.	A prototype cloud realm using Euca data is planned for 10/2017. When completed, federated data from all 3 sites will be available at the master XDMoD instance. Release materials via GitHub. Update periodically.

#1541215

Buffalo also completed a redesign of the XDMoD data warehouse to support cloud metrics and is moving this into the testing phase.	Currently waiting for latest release of Open XDMoD (v.6.5.1) which will be available at year end at http://open.xdmod.org/. Note: this version does not support cloud metrics but will give sites an opportunity to get infrastructure in place for a future version that does.		

3.1 Software Requirements & Portal Platform
Work continued on implementing Globus OAuth2 authentication. Cornell worked with UB and Globus to get UB on Globus Authentication and also ran a successful test with a UCSB user to authenticate to the portal via Globus Authentication.

3.2 Integrating Open XDMoD and DrAFTS into the Portal
There were minor changes to Open XDMoD to correct bugs found during testing of the updated schema to support cloud data. Work will continue at a faster pace in February as resources have been freed to continue collecting Eucalyptus data for XDMoD.

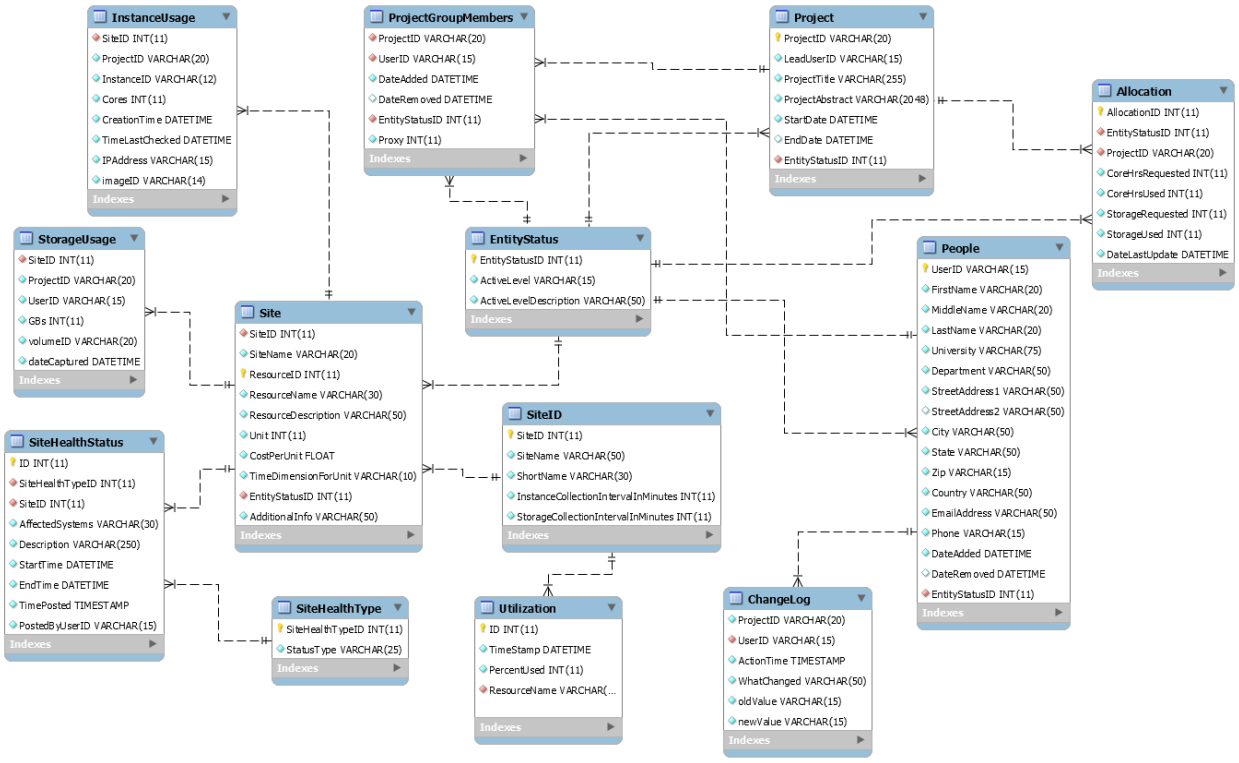
UCSB worked on improving the quality of DrAFTS predictions. Previously, they were seeing a random error rate of ~4%; that error rate has been reduce to under 1%. A more complete validation of these new results was submitted to *HPDC'17 (26th International ACM Symposium on High-Performance Parallel and Distributed Computing*) for publication consideration

In addition, a new set of experiments with the Globus Genomics group is also underway. Globus Genomics is developing a new scheduler that will use our DrAFTS prediction tool more directly. Based on results, this work may be submitted to the *SC17 Conference*.

3.3 Allocations & Accounting
Development on accounting and allocations is proceeding. The database and tables with test data are complete, and interface implementation is starting. Currently, we are working on php scripts to import project usage data from UB and UCSB using the REST API and adding the data to the database. We are also working on developing Stored Procedures for collating and reporting usage data; our planning includes handling unanticipated data gaps caused by unexpected downtimes, such as a power outage at one of the sites.

The database schema was updated this month with a SiteID table:

#1541215



4.0 Research Team Support

4.1. General Update

- Evaluation of specific requirements for the Jetstream Use Cases is underway. The first Use Case is likely to be Transient Detection in Radio Astronomy Data (see Use Case 4 below). Contact was established with Marlon Pierce (Indiana University) regarding XSEDE ECSS support.
- Work on MPI continues, particularly by Cornell Cloud Systems Engineer Bennett Wineholt. The goal is to complete a script that can make a cluster and run MPI with a workable level of performance in February 2017.

4.2 Science Use Case Team Updates

Use Case 1: A Cloud-Based Framework for Visualization & Analysis of Big Geospatial Data
UB created the second iteration of the user interface for the "Machine Learning for Sustainability" framework which is undergoing testing (e.g., running a Gaussian Process-based change detection algorithm on climate simulation data). The system will track usage through user accounts. UB will gradually open this framework up to the broader scientific community.

#1541215

Use Case 2: Global Market Efficiency Impact
UB finance researchers finished aggregating the tick-by-tick data for one of their projects. They will begin analyzing this data in February/March. Currently, they are doing outreach to other finance faculty and PhD students to assess interest in learning about how to use this framework.

Use Case 3: High Fidelity Modeling and Analytics for Improved Understanding of Climate-Relevant Aerosol Properties
MPI work continues as described above. A new postdoc, Tristan Shepherd, has arrived at Cornell and will be contributing to this work effort beginning in February.

Use Case 4: Transient Detection in Radio Astronomy Search Data
Cornell astronomers Cordes and Chatterjee plan to start processing with the Jetstream allocation in February; grad student Robert Wharton will be the point person. The goal is to process the whole of a campaign at the Very Large Array (VLA) to achieve localization of Fast Radio Bursts (FRBs) in order to look for additional FRBs in the whole field of view. Software will be built for both Aristotle and Jetstream.

Use Case 5: Water Resource Management Using OpenMORDM
No updates this month.

Use Case 6: Mapping Transcriptome Data to Metabolic Models of Gut Microbiota
Cornell PI Angela Douglas's NIH grant proposal was funded; as a result, the Aristotle project will benefit from 40 hours of effort which will be used to train students in modeling software and best practices, as well as algorithmic issues.

Use Case 7: Multi-Sourced Data Analytics to Improve Food Production & Security
Where's the Bear: Based on the positive results from the "Where's the Bear" experiments with TensorFlow, the Google open source software library for the machine learning, the UCSB science team is now establishing a collaboration with Zooniverse (https://www.zooniverse.org). Zooniverse is a citizen science project repository focused on research that enables citizen scientists to inform and forward efforts that rely heavily on digital imagery. The collaboration between the science team and Zooniverse will encourage citizen scientists to participate in the Sedgwick Reserve ecology effort (e.g. the quarterly deer survey) and also serve as validation for the automatic identification of species using the Where's the Bear software infrastructure. Aristotle will serve as the primary image repository serving images to Zooniverse and other interested researchers.

Agricultural Food Security Project: UCSB/Sedgwick Reserve is using the dormant season to plan a new, more water efficient irrigation infrastructure. Currently, the grape vineyard is watered (using a SmartFarm moisture monitoring system) from a well that also fills the Sedgwick ecological water reserve. It is possible to separate the agricultural water usage from the ecological usage with the addition of a new irrigation infrastructure. Aristotle personnel are participating in the planning since the intention is to use SmartFarm IoT (Internet of Things) technology to implement automatic irrigation scheduling based on real-time moisture sensing. The Aristotle cloud will host the analysis and schedule software infrastructure necessary to gather and process sensor information and provide actuation directives for the irrigation pumps and valves.

NSF
#1541215

Finally, based on this agricultural technology research, the science team has established a new collaboration with the Lindcove Research Extension Center (LREC - http://lrec.ucanr.edu). As part of this collaboration, the team will look at automatically analyzing data that is generated by a citrus "packline" operated at the center. LREC's primary function is to study citrus agriculture. To do so, it hosts a large variety of citrus trees that it uses to conduct experiments (e.g., pesticides, water usage, genetic grafting, etc.) and the automated "packline" machine is used to determine the effects (positive or negative) on fruit. The collaboration will investigate how packline data can be gathered, processed using analytics, and delivered to the test orchards in real time. Currently, packline analysis is done after-the-fact and manually. Thus, the project is investigating how IoT can be used to speed the decisions which growers make in the field by integrating existing technology that is otherwise unable to be used for this purpose.

5.0 Outreach Activities

5.1 Community Outreach

- 72 people (DIBBs PIs, co-PIs, and 7 NSF directors) attended the 1st NSF Data Infrastructure Building Blocks PI Workshop (DIBBs17) organized and chaired by Cornell (https://dibbs17.org) in January in Arlington. 37 posters on significant DIBBs success/innovations were featured and 37 white papers were submitted identifying DIBBs project challenges and solutions. NSF program director Amy Walton considered it an "outstanding" PI Workshop for the DIBBs community: "Your efforts created an energetic – highly productive – environment for the workshop. The activities encouraged involvement, collaboration, and contribution. Participants were *excited* about what they were doing."

  Cornell is currently conducting a post-workshop survey. The NSF DIBBs17 Workshop Report will be released by Cornell in April 2017.