ARISTOTLE
CLOUD FEDERATION

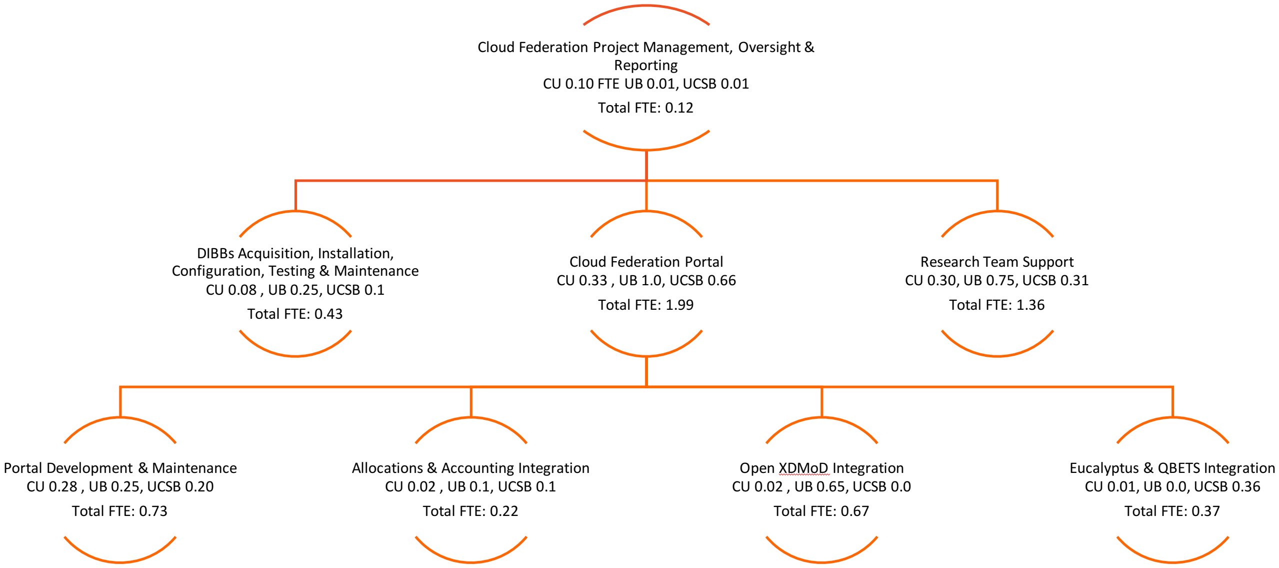
CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation

Monthly Report 2/26/2016

Report 5 of 18

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

This is the fifth 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).


Cloud Federation Project Management, Oversight & Reporting
CU 0.10 FTE UB 0.01, UCSB 0.01
Total FTE: 0.12
DIBBs Acquisition, Installation, Configuration, Testing & Maintenance
CU 0.08 , UB 0.25, UCSB 0.1
Total FTE: 0.43
Cloud Federation Portal
CU 0.33 , UB 1.0, UCSB 0.66
Total FTE: 1.99
Research Team Support
CU 0.30, UB 0.75, UCSB 0.31
Total FTE: 1.36
Portal Development & Maintenance
CU 0.28 , UB 0.25, UCSB 0.20
Total FTE: 0.73
Allocations & Accounting Integration
CU 0.02 , UB 0.1, UCSB 0.1
Total FTE: 0.22
Open XDMoD Integration
CU 0.02 , UB 0.65, UCSB 0.0
Total FTE: 0.67
Eucalyptus & QBETS Integration
CU 0.01, UB 0.0, UCSB 0.36
Total FTE: 0.37

#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

The first quarterly Aristotle External Advisory Committee (EAC) meeting was held on 2/12/2016. In attendance:

Name	Affiliation	Project	Email
Amy Walton	NSF	Aristotle	awalton@nsf.gov
Adam Brazier	CAC	Aristotle	brazier@cornell.edu
Susan Mehringer	CAC	Aristotle	susan@cac.cornell.edu
Paul Redfern	CAC	Aristotle	red@cac.cornell.edu
Jamie Kinney	AWS	SciCo	jkinney@amazon.com
Dmitrii Calzago	HPE	Eucalyptus	dzc@hpe.com
Craig Stewart	IU	Jetstream	stewart@indiana.edu
John Towns	NCSA	XSEDE	jtowns@illinois.edu
Rick Wagner	SDSC	Comet	rpwagner@sdsc.edu
Ian Foster	UC	Globus	foster@cs.uchicago.edu
Steve Johnson	WCM	NIH CTSC	johnsos@med.cornell.edu
Tom Furlani	UB	Aristotle	furlani@buffalo.edu
Rich Wolski	UCSB	Aristotle	rich@cs.ucsb.edu

Lifka, Furlani, and Wolski provided a project overview including WBS elements, governance, and a status update. Amy Walton commented on the significance of this project as a model for the NSF community, highlighting the importance of focusing on improved time-to-science and the role of advanced metrics in demonstrating that goal. In future calls, we intend to spend more time asking the EAC for advice and guidance. There was a desire from the reviewers to follow up with various Aristotle collaborations—specifically QBETs and Open XDMoD for cloud infrastructure. Furlani and Wolski agreed to do so. In addition, after the EAC meeting, Ian Foster put Rich Wolski (UCSB) in touch with the Globus Genomics team regarding QBETS capabilities. The teams decided that there might be an easy path to integration and are working on it. They think they'll be ready to run a test in March 2016.

1.5 Status Call with NSF Program Manager

Our monthly status call with NSF program manager Amy Walton was held on 2/10/2016. Topics of discussion included:

- Updates on Supercloud testing on Jetstream.
- Possibility of a NSF DIBBs workshop in conjunction with the fall 2016 CASC meeting.
    - Awaiting direction from Amy Walton.
- Update on XDMoD and QBETS early integration efforts and successes.
- Proof of concept activity moving basic science VMs from Cornell to Aristotle partner resources to ensure a consistent user experience. Next step will be to do an all-to-all test where each partner tests a local VM on the other partner sites.
- Continued progress with science teams getting their required software and workflows working.
- Preparing for a Science Team Advisory Board meeting now that the first quarterly EAC meeting has been completed.
- Amy Walton pleased with quality of Aristotle project reports to date.

1.6 Project Planning and Preparation

Project planning and preparation by the Aristotle team continued in February 2016. Extensive effort was put into the allocations and accounting system; in particular, how to transition allocation and accounting practices from early science team testing to production computing. CU and UB are completing installation and testing of their first year hardware installment from Dell. UCSB has completed vendor selection and has decided on Dell hardware. All of these efforts are described in more detail in this monthly report.

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

2.1 Federation Resource Status Updates

- **CU**
  The CU team's hardware is installed and is undergoing stress testing prior to its March 2016 deployment.

- **UB**
  The UB team's hardware is installed with a base OS and plans are to install the HPE Helion Eucalyptus stack the first week in March 2016.

- **UCSB**
  The UCSB team has selected Dell as the vendor for computation gear and quotes should be processed soon. They have made design decisions for the storage back end and will be deploying CEPT for EBS and RiakCS for S3. In addition, they are finalizing the design for joining the UCSB and Aristotle cloud.

The CU/UB/UCSB infrastructure planning table has been updated below:

	CU	UB	UCSB
Cloud URL	euca4.cac.cornell.edu	ccr-cbls-2.ccr.buffalo.edu	TBD
HPE Helion Eucalyptus Version	4.2.1	4.2.1	4.1.2
Migrate to 4.2.1	X	X	3/1/2016
Globus	Yes	Planned	Planned
InCommon	Yes	Planned	Planned
Hardware Quotes	New hardware will be deployed in March 2016.	New hardware will be deployed in March 2016.	Processing quotes.
Hardware Vendor	Dell	Dell	Dell
# Cores	168	112-140 (target)	TBD
Ram/Core	4GB/6GB/8GB	6GB	TBD
10Gb Interconnect	Yes	Yes	Yes

2.2 Industry Influence: Eucalyptus

No updates this month.

2.3 Potential Tools: CloudLaunch & Supercloud

No updates this month.

3.0 Cloud Federation Portal Report

No changes were made this month to the portal planning table below:

Portal Framework			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 3/2016	**4/2016 – 10/2016**	**11/2016 - 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.	Update materials to be federation-specific.	Add more advanced topics as needed, including documents on “Best	Release documents via GitHub. Update periodically.

		Practices" and "Lessons Learned." Check and update docs periodically, based on ongoing collection of user feedback.	
Training			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 3/2016	**4/2016 – 10/2016**	**11/2016 – 3/2017**	**4/2017 - 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 and materials to provide training asynchronously 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.
User Authorization and Keys			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 1/2016	**2/2016 – 5/2016**	**6/2016 – 9/2016**	**10/2016 – End**
Plan how to achieve seamless login and key transfer from portal to Euca dashboard.	Login to the portal using InCommon.	Get 4.2.1 federated key after InCommon login.	Move seamlessly to Euca console after portal InCommon 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.	Implement minimal set of Euca Tools to bridge portal to Euca console.	Add/refine/update, based on ongoing collection of user feedback.	Release via GitHub. Update periodically.
Allocations and Accounting			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 3/2016	**3/2016 – 5/2016**	**6/2016 – 9/2016**	**10/2016 – 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 – 3/2016	**2/2016 – 5/2016**	**6/2016 – 10/2016**	**11/2016 - End**
Implement graphs of basic usage data, including % utilization, available resources, and user balance, using scripts from Cornell CAC for basic data collection. Provide documentation for installing XDMoD and SUPReMM at individual sites.	Install Open XDMoD/SUPReMM at individual sites and begin data collection. This includes the installation of SUPReMM and the data collection piece at the federation sites. Begin integration with federated authentication providers.	Federated data collection across sites. Ship data from the individual sites to UB. We can summarize data remotely and send the summarized data or collect all raw data and summarize locally. Other job information will be federated as well using the prototype model under development with OSG. Display federated metrics in Open XDMoD at UB.	Release materials via GitHub. Update periodically.

3.1 Software Requirements & Portal Platform

Planning continued on the portal. It is currently populated with a "Coming Soon" page and news links at http://www.federatedcloud.org and will soon have an SSL certificate. In order to make the portal easy and free to recreate, a Bootstrap web framework was chosen to handle formatting. We expect to implement the framework in March 2016 and to begin to add documentation.

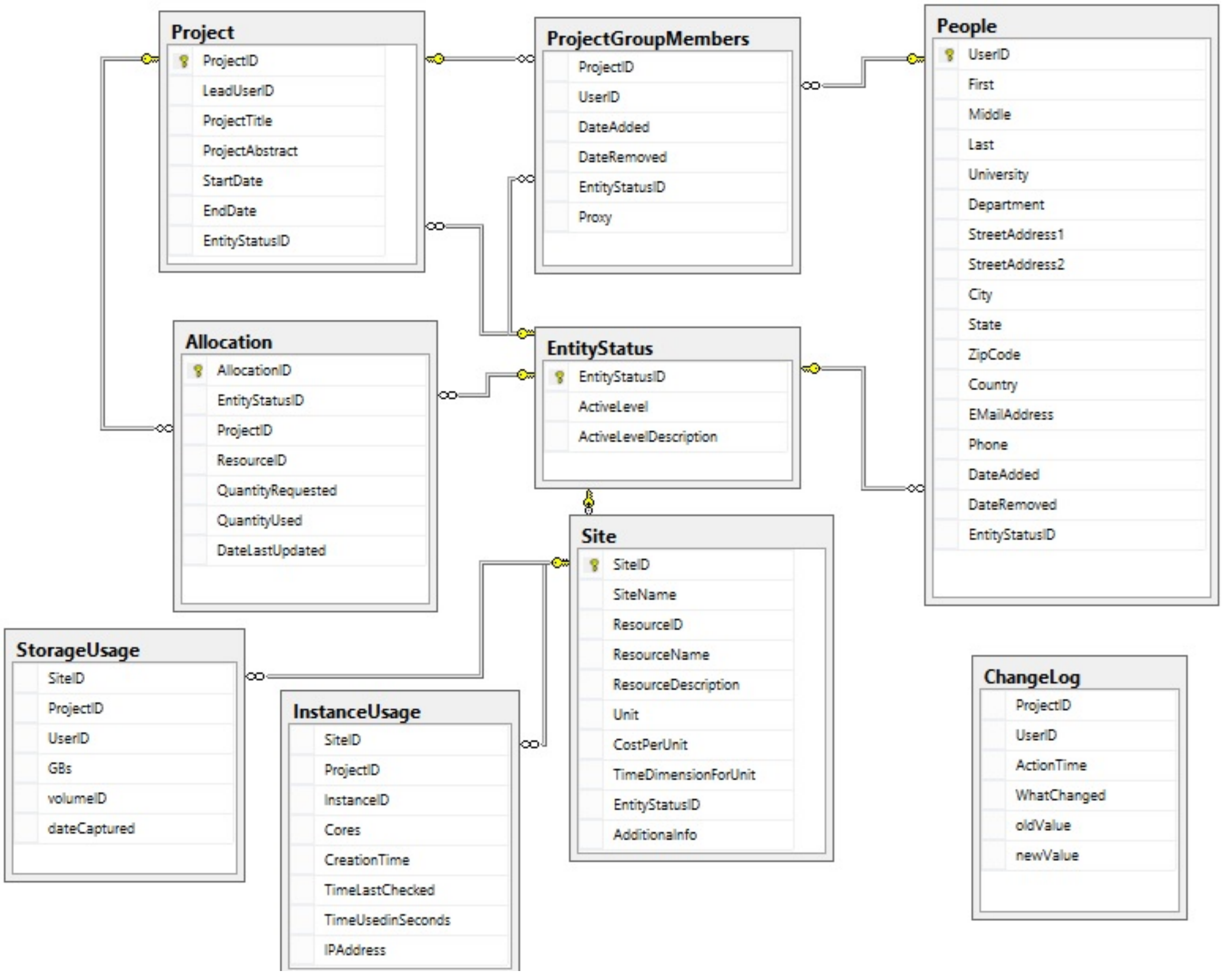
3.2 Integrating QBETS into Open XDMoD

On 1/11/2016 Steven Gallo, Joseph White, and Robert DeLeon (UB) had a conference call with Rich Wolski (UCSB). Following this conference call between the Open XDMoD and QBETS teams, the UB group ran simulations with QBETS using XSEDE HPC data to explore the use of QBETS as a wait time prediction tool. Nominally, QBETS can be used to choose which XSEDE resource to run on given a class of job. This type of analysis can also provide insight into choosing cloud resources. As a simple example, wait time was modeled for jobs employing a commonly run application—NAMD—which is frequently run on a various XSEDE resources. To provide the best statistics, wait times were compared on TACC's Stampede and SDSC's Comet where NAMD is most frequently run. In principle, the user could classify their job by application, nodes and any other parameters that are known pre-execution, run the QBETS prediction at the time of job submission, and choose the XSEDE resource with the smallest wait time. With a very small effort, the estimated run times of the jobs could also be factored in and the user expansion factor ([wait_time + run_time]/run_time) could be predicted to optimize the user's choice of resource. QBETS predictions were made for four classes of NAMD jobs: (1) 2 node Stampede jobs, (2) 2 node Comet jobs, (3) 4 node Stampede jobs, and (4) 4 node Comet jobs. Interestingly, through most of the time range, the 2 node NAMD jobs ran faster on Comet than they did on Stampede; this is the reverse of the finding for 4 node NAMD jobs. Overall, the results indicated that QBETS is a useful tool for predicting batch HPC job wait times. The details of the application of this QBETS technology to the Aristotle project will depend on exactly how the cloud federation is implemented.

3.3 Allocations & Accounting

We continued our biweekly meetings to plan allocations and accounting for compute and storage across both the federation and different funded projects. This month we finalized our plans for data collection and reporting for compute time, and drafted our plans for storage usage. Defined use cases include researchers who have an allocation, who have purchased time, who want to run across the federation, and all combinations thereof. To accommodate this, we will require that federation ProjectIDs follow one convention and are unique across the federation, and that drawing from multiple accounts must be done under separate ProjectIDs. This will allow us to accurately account and report usage. UB has developed a REST API that will expose Eucalyptus accounting data based on the scripts provided by Cornell. This will be made available shortly.

No changes were made this month to the allocations and accounting database schema below:



4.0 Research Team Support

4.1 General Update

Slack was chosen and is now in use for general science team communications. Contact with the project scientists is well-established by now. At Cornell, we have more science use cases and consequently less time for science support, so we are prioritizing those efforts.

Following discussions with CU's Susan Mehringer and Lucia Walle, project accounts are being created for science teams to test and configure cloud instances. Each PI will create a project account which will be the ongoing account used when full Aristotle accounting begins.

Accounts for researchers on the UB CCR LakeEffect cloud have been created. The development of this cloud will be subsequently mirrored on the Aristotle cloud.

4.2 Science Use Case Updates

Use Case 1: A Cloud-Based Framework for Visualization & Analysis of Big Geospatial Data
The UB geospatial analytics use case team has created a virtual machine (VM) which runs the visualization and analysis interface (iGlobe) on the UB cloud. We have also created scripts to automatically start and terminate Hadoop and Spark clusters on the cloud. The scripts will be integrated with the visual interface (iGlobe) to execute analytics in the cloud.

Use Case 2: Global Market Efficiency Impact
The UB finance use case team has defined specifications for their use case. The necessary software (SAS) has been acquired and a VM has been created for the application. The next step is to run the use case on the VM.

Use Case 3: High Fidelity Modeling and Analytics for Improved Understanding of Climate-Relevant Aerosol Properties
Brazier met with Sara Pryor's CU research team (3 faculty/1 graduate student) to determine next steps and usage patterns. As a result of that meeting, an instance with an MPI-enabled version of WRF-Chem was built and documented by Steven Lee and will be tested by the Pryor team in March 2016.

Use Case 4: Transient Detection in Radio Astronomy Search Data
No updates this month.

Use Case 5: Water Resource Management Using OpenMORDM
CU's Patrick Reed identified his use case team. He will create a project account in early March.

Use Case 6: Mapping Transcriptome Data to Metabolic Models of Gut Microbiota
In order to investigate better configuration management—a key issue with cloud deployment of the science use cases—Brandon Barker (CU) created a NixOS image and is configuring it for the team. NixOS is a cloud-focused OS with the entire configuration of the system being specified in a declarative fashion, making it ideal for reproducible science.

Use Case 7: Multi-Sourced Data Analytics to Improve Food Production

The Sedgwick Reserve is studying the effects of the California drought on agriculture and animal activity. As part of that effort, they have installed a set of soil moisture sensors in a vineyard that is being farmed for table crops. The sensor data is automatically captured and sent to the cloud at UCSB for analysis. While the full soil moisture sensor installation is not yet complete, the first set of sensors are now installed and functioning (data available from http://128.111.84.220:22001/).

In addition, the team has installed a number of monitoring programs to determine the reliability of the moisture sensing network. This reliability and subsequent analysis appear to be necessary to ensure data integrity, although the need for this reliability analysis (and its nature) are new to this project.

Sedgwick plans to complete the installation, and then combine the moisture data with camera trap data to document animal behavior as the El Nino season continues in California. The camera trap image analysis application (called "Where's the Bear?") is beginning its on-boarding process. There is an application running that will transition onto the cloud that is analyzing the current image archive. In March 2016, we will start working on the data acquisition part so that the images can be ingressed automatically.

5.0 Outreach Activities

5.1 Media Outreach

Cornell contacted the Editor of *Scientific Computing World* (circulation 80,000 across all media) and pitched including federated clouds as part of a February 2016 cloud story. Subsequently, Lifka was interviewed by Editor Tom Wilkie and the article "Will the cloud change scientific computing" is now out at: http://www.scientific-computing.com/news/news_story.php?news_id=2781. It has been retweeted to > 25,000 followers to date. Amy Walton and Irene Qualters were pleased with the article.

5.2 Industry Outreach

Cornell was interviewed by a 5-person HPE Helion Eucalyptus team in early February. The information shared will serve as content for future communications on federated and hybrid clouds by HPE, i.e., articles, case studies, blogs, etc.