ARISTOTLE
CLOUD FEDERATION

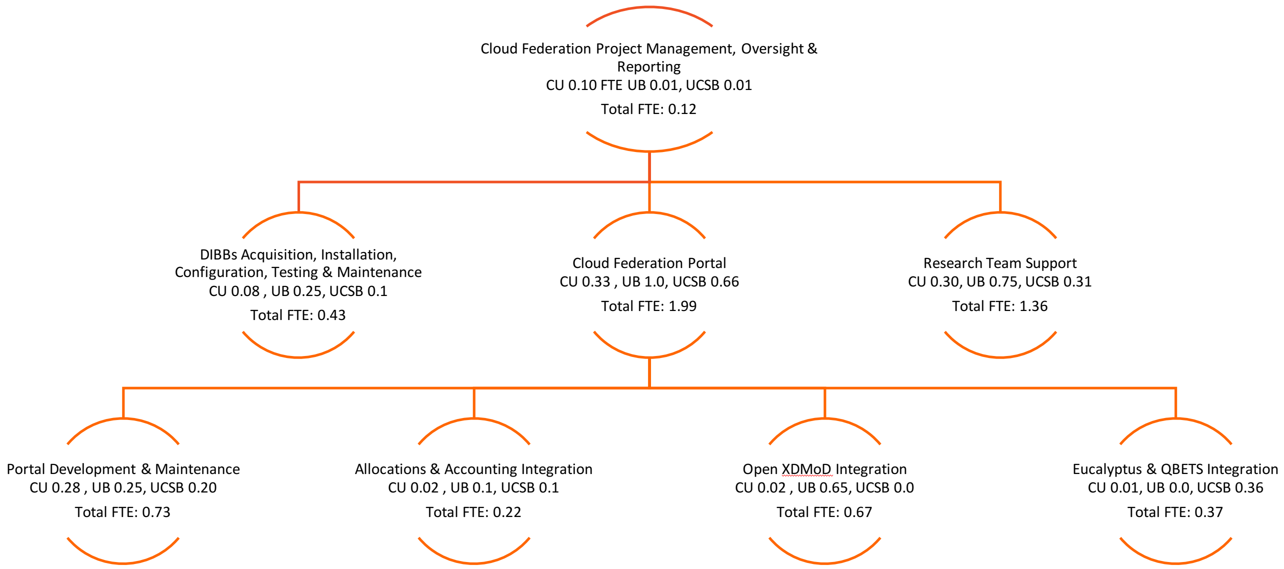
CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation

Program Year 1: Quarterly Report 2

3/30/2016

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

This is the “Program Year 1: Quarterly Report 2” 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

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 quarter.

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

As follow-up to Lifka's Cycle Computing meetings, CU is preparing to send an improved version of CloudLaunch to Cycle Computing so that they can review it and consider a partnership with AWS to have CloudLaunch queues available to the community on AWS resources. We expect to continue these conversations in Q3 2016.

Lifka had a call with the AWS Scientific Computing (SciCo) team supporting Cornell and Aristotle. There continues to be great interest in our project and in broader interactions with the NSF HPC community. A project update was given and we discussed how we might work together in the coming months. We also discussed the potential for allocating AWS time to XSEDE and leveraging CloudLaunch with Cycle Computing and Supercloud from Cornell. These potential activities may facilitate making the transition to cloud easier for the science community.

Lifka arranged a call with George Turner, an IU Jetstream team lead, and Cornell Professor Hakim Weatherspoon (http://www.cs.cornell.edu/~hweather), Research Scientist Robbert van Renesse, (http://www.cs.cornell.edu/Info/People/rvr/) and their team about partnering with Cornell CS on their Supercloud project which is jointly funded by NSF CISE (program director Amy Apon) and NIST. This technology has the potential to make it easy to burst between Aristotle and other clouds including Jetstream, Chameleon, CloudLab, and the major public cloud providers. It removes the burden of requiring researchers to have different virtual machines for each cloud software stack. The meeting went very well and there is definite interest on the part of the Jetstream team. IU will provide Weatherspoon and van Renesse's team with Jetstream accounts for testing. Once testing is complete, the Cornell CS team will do a live demonstration of Supercloud bursting to and from Aristotle resources for the Jetstream team. Assuming a successful demonstration, Cornell CS will then give a demonstration to the Aristotle PIs and project leads. Rich Wolski's depth of cloud technology knowledge will really help us understand the benefits and limitations of Supercloud. Lifka also discussed this technology with XSEDE PI John Towns who is also extremely interested. After our in depth evaluation of Supercloud, we hope to provide a demonstration to our External Advisory Committee to better understand how we might move this technology from the lab into the hands of the national community. We will continue to report on this effort.

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

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 and there was an effort to integrate QBETS into Globus Genomics to better predict the AWS spot market. The initial results were impressive; almost “too good to be true.” Further testing seems to confirm the value of QBETS for predicting the AWS spot market. The XDMoD team took the standard QBETS distribution and tested it against the current NSF Service Provider system queues. Again, initial results were impressive and further interactions, particularly around its use for Comet at SDSC, will continue.

Discussions regarding allowing the UB and UCSB science teams to get started on Red Cloud at Cornell also took place. At UB, users are using their local UB cloud (“Lake Effect)” to develop and run scaled down versions of their use cases. Once the UB Aristotle cloud is available, these applications will be migrated. Cornell will provide the UB global finance science team access to Red Cloud in order to accommodate their high volume storage requirements.

Significant effort was put into scheduling the first Science Team Advisory Committee (STAC) meeting. This meeting is now scheduled for 4/1/2016. Lifka and Furlani will be in transit from the CASC meeting, so this meeting will be led by Adam Brazier, Susan Mehringer, and Resa Reynolds. Lifka will help set the agenda.

1.5 Status Calls with NSF Program Manager

The first monthly status call of the quarter with NSF program director Amy Walton was held on 1/11/2016. Topics of discussion included:

- Aristotle project may have a pathfinder role in helping NSF map out what roles clouds (NSF, public, etc.) might play in future NSF cyberinfrastructure.
- Possibility of a NSF DIBBs workshop in conjunction with the fall 2016 CASC meeting.
- Potential NSF reviewers for next round of DIBBs proposals: Dave Lifka or Susan Mehringer.
- Hewlett-Packard Enterprise (HPE) Helion Eucalyptus team considering building InCommon support into the Eucalyptus portal.
- Key project deliverables to keep in mind: Can others pick up Aristotle components and use/ customize them? Are lessons learned documented?
- Collection of cloud metrics will be added to Open XDMoD in CY2016.
- A new version of allocations and accounting database schema was developed last month.
- We are starting to get science teams prepared, e.g., Brazier got WRF-Chem atmospheric chemistry observation and modeling software running on a Red Cloud virtual machine, we started to identify queries that will stress astronomy data in new ways, etc. Some use case scientists know how to make their codes run in the cloud; others don't know what their workflows are. We will take an agile approach. Understanding how to easily containerize codes will be an important lesson learned.
- Amy Walton pleased with quality of Aristotle project reports.

Our next monthly status call with 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.

On 3/7/2016, Lifka had a conference call with ACI director Irene Qualters. Irene was pleased to hear about the project's early results. The main points of discussions included cloud usage paradigms represented by the Aristotle science teams and interoperability between cloud platforms (Eucalyptus, OpenStack, AWS, Azure, and Google).

Lifka was invited by NSF to present an overview of Aristotle Q1/Q2 progress in Arlington on 4/1/2016.

On 3/11/2016, our last monthly status call of the quarter occurred. Topics of discussion included:

- Updates on Supercloud testing on Jetstream.

- Further discussions regarding the possibility of holding a NSF DIBBs workshop in conjunction with the fall 2016 CASC meeting.
    - Awaiting direction from Amy Walton.
- Updates on XDMoD and QBETS early integration efforts, QBETs success to date with Globus Genomics, and potential use of QBETS by NSF Service Providers.
- Proof of concept activity moving basic science VMs from Cornell to Aristotle partner resources to ensure a consistent user experience. This will be followed by an all-to-all test where each partner tests a local VM at each of the partner sites. These tests are still being prepared.
- Continued progress with the science teams in getting their required software and workflows working.
- Preparing for the Science Team Advisory Board meeting now that the first quarterly External Advisory Committee meeting has occurred.
- 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 January 2016. Requirements analysis for the allocations and accounting system included how using Eucalyptus availability zones may allow UB and/or UCSB to partition their local cloud installations between local university users and Aristotle-funded users. We discussed and gained a better understanding of Eucalyptus provided federation capabilities and how we might leverage them. CU and UB placed orders for their first year hardware installment from Dell and prepared for installation and testing. UCSB negotiated potential hardware purchases with HPE, Dell, and other vendors. There was also continued User Portal planning and development, and ongoing work with the science teams.

February 2016 project planning and preparation included extensive effort on the allocations and accounting system; in particular, how to transition allocation and accounting practices from early science team testing to production computing. CU completed installation and testing of their first year hardware installment from Dell. UB received their first year hardware and began installation. And, UCSB completed vendor selection and ordered Dell hardware.

March 2016 project planning and preparation continued with a focus on the allocations and accounting system. UB acquired the public network space for Aristotle and will be installing the Eucalyptus stack in April. UCSB began to receive their servers.

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

2.1 Federation Resource Status Updates

- **CU**
  The CU team installed and stress tested its year 1 hardware. We added 168 cores (6GB/core) and 120TB SAN storage to the Red Cloud infrastructure.

- **UB**
  The UB team cabled their new Aristotle servers, storage, and network switch and installed a base OS. In addition, UB upgraded the Center for Computational Research's (CCR) existing cloud (called "Lake Effect" - https://ubccr.freshdesk.com/support/solutions/folders/5000273002) to HPE Helion Eucalyptus version 4.2.1. This upgrade experience will be beneficial in the installation of the UB Aristotle cloud software. UB's Central IT has acquired the public network space

necessary for the Aristotle cloud and the UB team will be installing the Eucalyptus software in April 2016.

- **UCSB**
  The UCSB team investigated a variety of storage and compute configurations and negotiated with HPE and other vendors. We selected Dell as our vendor for computational gear and decided to deploy CEPT for EBS and RiakCS for the S3 storage back end. Our first Dell servers have arrived. We are finalizing the design for joining the UCSB and Aristotle cloud.

The three sites briefly discussed using HPE Helion Eucalyptus availability zones and that possibility is still under consideration by UB and UCSB as an easy way to separate usage accounting while still allowing resource sharing. UCSB plans to evaluate separate AZs in either EDGE or MDO networking setups.

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	1/1/2016	N.A.	TBD
Globus	Yes	Planned	Planned
InCommon	Yes	Planned	Planned
Hardware Quotes	Hardware deployed. 168 cores added to existing Red Cloud. 376 total cores.	Hardware installed. Working through network reqs. with Central IT. Install Euca stack next.	Hardware ordered.
Hardware Vendor	Dell	Dell	Dell
# Cores	168*	112-140 (target)	140 (target)
Ram/Core	4GB/6GB/8GB	6GB	8GB
10Gb Interconnect	Yes	Yes	Yes

* 168 additional cores augmenting the existing Red Cloud.
**UB and UCSB installing Aristotle as new cloud (not integrating with existing clouds)

2.2 Industry Influence: Eucalyptus

The CU team had a couple of meetings with HPE's Chris Grzegorczyk to discuss using InCommon with HPE Helion Eucalyptus. Specifically, we wanted to share Aristotle requirements for InCommon integration with Chris so that he could in turn let us know how HPE might be able to assist.

Chris offered two possibilities:
1) Add InCommon support to HPE Helion Eucalyptus' current cloud federation so that a user could log in via InCommon credentials and run on any cloud in the federation using the same account ID, the same keys, and subject to the same IAM policies;
Or,

2) Allow users to log into the web console using their InCommon credentials. The user's account and permissions will need to be administered separately in each cloud.

Given the decentralization of infrastructure in academe, we chose the second approach. HPE Helion Eucalyptus will add InCommon login support on the web console and create tools for cloud administrators to associate InCommon DNs to HPE Helion Eucalyptus users.

2.3 Potential Tools: CloudLaunch & Supercloud

Steven Lee of the CU infrastructure is massaging the CloudLaunch software stack (https://www.cac.cornell.edu/technologies/CloudLaunch.pdf) before sending it to Cycle Computing for evaluation by their engineers.

The CU infrastructure team also met with Cornell's Supercloud developers and Indiana's Jetstream team to facilitate future testing of Supercloud, a Cornell CS tool that enables users to easily run images across cloud platforms. Supercloud's Xen-Blanket wrappers allow images created under HPE Helion Eucalyptus to run on an OpenStack cloud, AWS, Azure, etc. Indiana provided Cornell access to Jetstream for a Supercloud proof of concept and it is in early test phase. The first test will be to migrate a VM from Jetstream to Red Cloud and back. We plan to report results next month.

3.0 Cloud Federation Portal Report

The format of the portal planning table (below) was modified this quarter to allow much more detail to be added. Dates were also updated to reflect gating factors and additional process steps.

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 Practices" and "Lessons Learned." Check and update docs periodically, based on ongoing	Release documents via GitHub. Update periodically.

		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

The basic portal platform was built on a cloud instance, running Centos7 with the LAMP stack installed, and populated with a "Coming Soon" page at http://www.federatedcloud.org. The SSL certificate was also installed. Next, we will add InCommon authentication. In order to make the portal easy and free to recreate, a Bootstrap web framework was chosen to handle formatting. The first draft implementing the framework and the portal design has been completed. Next quarter we expect to implement the framework and start populating content.

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 (USCB) on integrating QBETS into XDMoD and Aristotle. We discussed how QBETS works, how it could be combined with data from XDMoD, and potential uses for it in modeling Aristotle usage, e.g., predicting the value of sharing between different clouds within the federation and bursting to AWS. Wolski subsequently prepared a stand-alone version of QBETS with documentation and test data and put it in a GitHub repository accessible to UB personnel.

In February 2016, UB 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.

In March 2016, Wolski revived a clustering version of QBETS which allows for tighter batch job wait time predictions based on using clustering algorithms to predict restricted classes of job wait times. The code can cluster jobs based on requested maximum wait time, number of processors, or a product of the two. The clustering version of QBETS was downloaded from GitHub and compiled by UB Center for Computational Research (CCR) personnel. We ran the code in all of the different clustering modes using CCR Rush batch job data. Typically, 1-3 clusters were found for either clustering by requested maximum run time or by processor. In some cases the bounds on the predicted wait times within the various clusters were tightened compared to the non-cluster single QBETS prediction. We also ran it successfully using XSEDE TACC Stampede batch job wait data. The next step will be to run clustering QBETS on other XSEDE resource data sets and compare the predicted wait times between the resources. Unfortunately, most other XSEDE resources do not currently report requested wait time to the TGCDB. We are working on obtaining the requested wait time directly through the SUPReMM data pipelines. When this data has been obtained, we will run clustering QBETS on the wait time data and compare between XSEDE resources to see if clustering QBETS would be a viable batch job wait prediction tool.

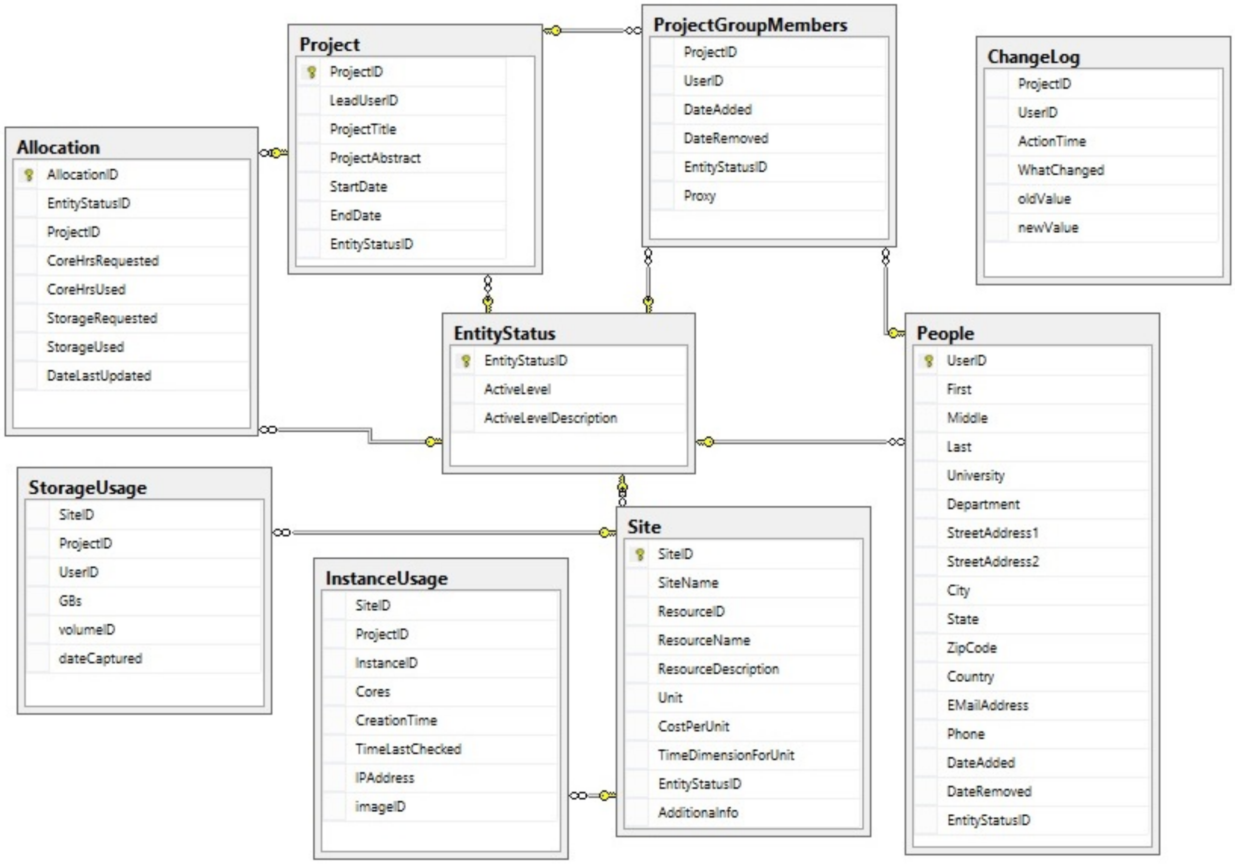
3.3 Allocations & Accounting

This quarter the team began biweekly meetings to flesh out use cases for allocations across both the federation and different funded projects. A researcher may fall into a very simple case of using resources from one allocation at one site, while another may draw from several funding sources and work on multiple nodes in the federation. Our goals are to define all use cases, then define requirements for users, data collection and reporting, and finally, design a phased implementation plan.

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. We established a convention for federation ProjectIDs that 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.

The database schema for allocations and usage data has been updated (see the new schema the top of page 12). The database has been created. This month, we began implementing this plan; four project groups have been created at Cornell, with project, user, and usage data, enabling the researchers to begin setting up instances.

UB has developed a REST API that will expose Eucalyptus accounting data based on the scripts provided by Cornell. This month Cornell began testing and implementing this API for use within the federation.



4.0 Research Team Support

4.1 General Overview

We are compiling a list of packages of general utility for a standard base image, package list, or scripted installation. This work is in progress and subject to review; so far, the list includes:

wget	mlocate
gcc-fortran	csh
tcsh	perl
time	redhat-lsb
redhat-lsb.i686	compat-libstdc++-33.x86_64
compat-libstdc++-33.i686	

Additionally, a brief list of optional items, such as X11, will be produced to help users who may be more familiar with desktop Linux installations which are more software-rich.

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.

Trello was chosen for Science Team Support task management, along with GitHub for code management and related issues. Both Trello and GitHub were integrated into Slack channels for ease of reporting and providing information for project management.

The first quarterly meeting of Scientific Team Advisory Committee (STAC) will occur on 4/1/2016 at 12:00 noon EST.

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.

We deployed the base webglobe server on http://128.205.11.214/wglobe/Index.html. The server is currently running on the UB Lake Effect cloud and will eventually be migrated to the UB Aristotle cloud. The analysis stack currently setup in the cloud consists of:

- An HDFS with climate data stored as distributed NetCDF files.
- iGlobe server which analyzes the distributed NetCDF data using the SciSpark API which runs on top of a Spark cluster. The iGlobe server creates the Spark cluster "on-demand" and destroys the cluster after the analysis is done.
- Web server that hosts the webglobe client. The client allows users to interact with the underlying climate data through their browsers.
- For the next step, we will deploy more advanced distributed analytic codes to interact with the data. The web interface will also be developed further to allow users better interaction with the climate data.

Use Case 2: Global Market Efficiency Impact
UB science team lead, Varun Chandola, is currently installing, testing and running his VM environment on the existing UB CCR cloud and will migrate to the UB Aristotle cloud as it becomes available. This will allow UB use case scientists to begin leveraging the Aristotle cloud almost immediately.

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. We developed and tested the capability to create and destroy VMs with SAS installed on the machines. The primary researcher has been able to execute the basic codes as part of the use case. Currently, more extensive jobs are being executed.

Use Case 3: High Fidelity Modeling and Analytics for Improved Understanding of Climate-Relevant Aerosol Properties
Sara Pryor's graduate student received training on WRF-Chem and will assist in the installation of the distributed-memory version of WRF-Chem. 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 April 2016. Instances (single-process and distributed-memory) were migrated from a development project to a newly-created Aristotle project for testing and development.

Use Case 4: Transient Detection in Radio Astronomy Search Data
Work has begun on creating an instance, with necessary software installed, for the astronomy search data use case. Brazier met with Cordes (PI) and Chatterjee and representatives of two other radio astronomy projects: the Long Wavelength Array (LWA-New Mexico) and the Murchison Widefield Array (MWA-Australia) to discuss the possibility of adding additional data sets. Cordes, Chatterjee and Brazier planned basic software architecture and functional requirements.

Use Case 5: Water Resource Management Using OpenMORDM
CU's Patrick Reed identified his use case team. His team created a project account in early March. They plan to build instances for OpenMORDM in April 2016.

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.

Barker communicated with PI Douglas on possible additional requirements, as well project setup on Aristotle. He is investigating the use of Supercloud as a possible means to scale out to other clouds. Initial versions of libSBML and COBRApy packages for Nix were built and software documented.

A Windows instance is preferred by this science team for MATLAB. This was created and user documentation is being added, including edits pertaining to how to create a Windows instance. We set up an NFS server for serving application files and for investigating MATLAB Linux network client installation to NFS volume. A Samba server was set up, tested and documented to allow file sharing across Windows and Linux. We also documented scalable file access for cross-cloud or many-instance file/app access and for starting MATLAB from a shared filesystem.

Use Case 7: Multi-Sourced Data Analytics to Improve Food Production
This quarter's work on the UCSB-led use case included:

- Installed the IRROMETER soil moisture monitoring system and currently integrating data acquisition with Aristotle.
- Recruiting a student researcher for the "Where's the Bear?" photo image analysis project.
- Presentation and recruiting opportunity presented at freshmen seminar on ecological science and sustainability (Jan. 12, 2016).
- IRROMETER data acquisition not functional yet. January 15, 2016 debugging field trip unsuccessful; scheduled vendor to provide on-site service on February 3, 2016.
- On-site data acquisition "weigh station" hardware is on order from Sedgwick.

- Segdwick found a tenant for a test farm. Planting of peppers will be scheduled after the first set of soil moisture sensors are sited.
- Next quarter plans include a Citizen Science deer survey and a soil electrical conductivity (EC) survey which measures soil properties that affect crop productivity.

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 was installed and functioning. Unfortunately, the soil moisture sensing project experienced a hardware failure in March 2016. Sedgwick is meeting with the vendor in early April to get new equipment. They have about a month's worth of data, but it is intermittent due to hardware instability in the sensors. A new site will come online as soon as the hardware issues are resolved. 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. The image processing survey is proceeding. Currently, the camera trap data is downloaded manually from each camera once a quarter or so. The latest data will be uploaded in early April. There are plans to ingest the images automatically in the coming months.

5.0 Outreach Activities

5.1 Media Outreach

A January 9, 2016 *Cloudhostingzine* article on "Cloud Computing for Academic Research" featured the Aristotle Cloud Federation.

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" was published at: http://www.scientific-computing.com/news/news_story.php?news_id=2781. It has been retweeted by 42 outlets to >25,000 followers to date. Amy Walton and Irene Qualters were pleased with the article. This article was subsequently featured by *InsideHPC*.

5.2 HPE Helion Eucalyptus Outreach

Cornell had several calls with Colby Dyess, HPE cloud marketing director, and worked with the HPE cloud referencing team (Sean Garcia and Tracy Roberts) to capture the Aristotle Cloud Federation story and lay the groundwork for future outreach activities, e.g., case studies, blog posts, etc. A case study will be publically available in Q3 2016. Cornell also introduced the HPE Eucalyptus manager to the AWS Scientific Computing manager to help build a federated cloud ecosystem.