ARISTOTLE
CLOUD FEDERATION

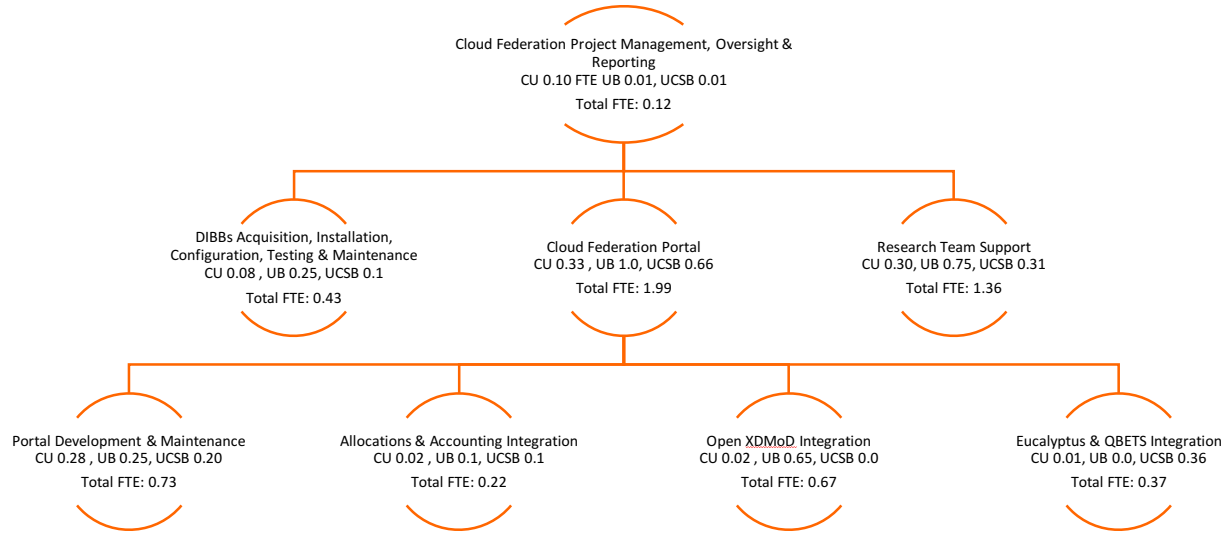
CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation

Monthly Report 11/20/15

Report 2 of 18

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

This is the second 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

#1541215

1. Cloud Federation Project Management, Oversight & Reporting Report

UCSB has identified Andreas Boshke as their science team lead supporting Science Use Case 7.

Co-PI Wolski proposes a project change request. He would like to employ UCSB technical staff for the support and maintenance functions associated with Aristotle to better support the science efforts. The research functions will be carried out by Wolski and a Graduate Research Assistant that are currently in the budget, but funds originally allocated for a postdoctoral researcher (without a change in funding level) would support technical staff at UCSB instead.

The reason for this change is two-fold. First, the funding level available for a postdoctoral researcher is insufficient (in the West Coast job market) to attract a qualified candidate. Wolski has explored other creative options to augment the salary (e.g., split finding a position with another project) but the salary level is just too low. The level was set by the project so that there would be parity across sites (i.e., UCSB did not spend substantially more on personnel costs) and it is an unfortunate circumstance (but a confirmation of intrinsic value) that it is proving almost impossible to woo a postdoctoral researcher with the appropriate research interests and skill set from industry.

Secondly, the UCSB responsibilities are split between what are essentially production-support functions and research functions. UCSB staff (like the staff funded by the project at Cornell and the University of Buffalo) are experienced in providing the support and maintenance functions. Thus, engaging UCSB technical staff will make the administrative staffing at UCSB resemble more closely the staffing employed at CU and UB. At the same time, the research functions (surrounding QBETS and load federation) will remain with the Co-PI and the Graduate Researcher. Like CU and UB, UCSB staff also has experience in transitioning research products into user-community use.

We believe that this staffing change better suits the project's goals and maximizes the chance of success given the challenges it will address.

PI Lifka is supportive of this change request as it does not change the project requirements, deliverables or time frames for activities for UCSB. As explained above, it also provides better staffing parity with CU and UB.
Lifka seeks approval from NSF Program Director Amy Walton.

Lifka has completed necessary subcontracts paperwork with the Cornell Office of Sponsored Programs. UB has confirmed their subcontract is in place. We are still awaiting confirmation of the UCSB subcontract.

Lifka and Furlani have responded to NSF Program Director Amy Walton's inquiry regarding the report status for another NSF grant Furlani is participating in. The report has been submitted by UB and is awaiting NSF approval involving multiple directorates. The grant PI (Kofke) contacted the NSF PO and was told the final report was approved. CU's cooperative agreement and project execution plan (PEP) cannot be approved by NSF until this is resolved by NSF.

A great deal of planning has been done by the team in November. A Google shared folder with all project planning documents has been established. Team members are tasked with updating those documents as part of the monthly reporting process. Mehringer and Reynolds have been working on a detailed multi-phase plan for all Aristotle components and capabilities required for a targeted January 2016 go-live. Mehringer has lead discussions on portal requirements. Initial requirements are outlined later in this

#1541215

report. Lifka, Furlani and their team members met at SC15 to discuss the phased implementation of Open XDMoD. UB and UCSB have begun initial conversations around QBETS integration into Open XDMoD.

2. DIBBs Acquisition, Installation, Configuration, Testing & Maintenance Report

The CU team obtained updated Dell quotes for the first hardware purchase. CU also upgraded Red Cloud to Eucalyptus 4.2 and worked with HP to obtain a hotfix for 4.2 that will be included in 4.2.1.

The Buffalo team has been working with their Central IT department and NYSERNet to secure a /22 block of IP addresses. The campus currently does not have anything greater than /24 blocks. They are working on rolling out their existing Eucalyptus cloud into a production state and have integrated it into their FreeIPA and Foreman/Puppet infrastructure. Buffalo has also worked on refreshing the previous hardware quotes with some slight changes based on feedback from Eucalyptus during the installation of their first cloud.

The UCSB team apprised the campus NOC of the scope and scale of Aristotle's bandwidth and addressing needs and researched current 10G switch offerings. They also worked with vendors on updated hardware quotes based on the Dell quotes submitted with the proposal.

Reynolds created the following planning doc in the Aristotle Google shared folder. She is actively working with the team to identify target dates for all capability roll outs and resource specifications.

	CUI	UB	UCSB
Cloud URL	euca4.cac.cornell.edu		
EUCA Version	4.2 with hotfixes	4.1.2	4.1.2
Migrate to 4.2.1	As soon as it's available	?	3/1/16
Globus	Yes	?	?
InCommon	Yes	?	?
Hardware quotes	Waiting for updated quotes	?	?
Hardware vendor	Dell	?	?
# cores			
ram/core			
10gb interconnect	Yes		

CloudLaunch: continues to be tested and hardened at CU. Lifka had discussions with Cycle Computing at SC15 around Cycle Computing hardening and supporting CloudLaunch while continuing to make it available to the community as open source. Their goal would be to have AWS login nodes running CloudLaunch that would allow anyone to login and submit jobs to AWS. This would be ideal for the Aristotle Federation. Users could simply login to a different login node to submit jobs that require more hardware resources than are available in the Federation. Lifka will follow up with Cycle Computing to see if we can move this idea forward.

RT (Request Tracker): Aristotle incidents will be tracked using CU's RT ticketing system. Initially, researchers will send email to Aristotle-help@cornell.edu to report problems, request support, or contact the team. Once the Aristotle portal is "live", we plan to add an alias of help@federatedcloud.org. Researchers will then be able to use either email to submit requests. Brazier, the science lead, will assign science-related tickets and Reynolds, the infrastructure lead, will assign infrastructure-related tickets as necessary.

NSF
#1541215

For clarification:
Aristotle-team@googlegroups.com – to be used for team-wide discussions.

Aristotle-help@cornell.edu – to be used for reporting issues/problems/resolutions.

Discussion items can go to the Aristotle-team list for discussion, but problems/questions should go to the ticket system for a number of reasons:

- For project reporting purposes, we need it to keep project problem tracking in one place
- It is the best way to track open problems
- Researchers will get a response even when their personal contact is out of touch
- Consultants can find project history

If a researcher persists in writing to the consultant directly, the consultant should make that contact into a ticket (retraining).

3. Cloud Federation Portal Report

Portal work this month focused on planning. As a team we agreed on what each phase of "Aristotle Production" would include. Phase 1 was the primary and obvious focus. We expect to refine Phase 2 and beyond as we progress. Each step is focused on providing full federation capability while making each of the components stand-along ready and redistributable. Phase 1 largely leverages existing components in use at each site. Our current component plan is as follows:

Component	Phase 1	Phase 2	Phase 3
	Now - Jan 2016	Jan 2016 - 18 month mark	18 month mark - ??
Allocations & Accounting	Allocations: Fair division of resources across three sites and projects based on project readiness. Accounting: Implement the accounting and tracking systems currently used on Red Cloud. UCSB & UB report the same data back to CAC, i.e., poll data and send reports to CAC.	Move portal & database to AWS.	Make available as download from GitHub.
Documentation & Training	Create basic user docs, focused on materials that will get users started. Draw from existing Red Cloud docs and the user project requirements.	Move the docs into a repository for the federation to draw from.	Make available as download from GitHub.
Usage & Status	Show % utilization graphs. Show available resources. Show user balance.	Incorporate Open XDMoD.	Incorporate QBETs (via Open XDMoD). Make available as download from GitHub.
User Authorization & Keys	Login to the portal using InCommon.	Get 4.2.1 federated key after InCommon login.	

#1541215

"Euca Tools"	Identify common euca portal tasks to be embedded in the portal via a button to a script. Identify which images should be created.	Create a repository to give back to euca.	
Systems	Get Globus running on all sites. Order and install hardware all sites. Determine software requirements for portal and accounting elements (see details in Aristotle spreadsheet).		

Software requirements for the portal were defined; we will be using open source software that meets requirements for the key functionality elements of authentication and Open XDMoD. Timeline planning for incorporating Open XDMoD was planned in more detail, first for incorporating Open XDMoD at each site, then for all sites on a federated portal, and later incorporating QBETs. We determined that the first phase of the project portal template will include project and user info, basic system status, documentation, and utilize InCommon authentication. Development of the portal template is planned to begin on Dec 1.

Portal software requirements are as follows:

- Ubuntu 12.04 or CentOS 6
- Apache
- MySQL 5.1 or 5.5
- PHP 5.3+
- Java
- PhantomJS
- Cron
- Logrotate
- MTA with sendmail compatibility (e.g. postfix, exim or sendmail)
- User docs, team contribution: MediaWiki -or- draw user docs directly from GitHub
- Log in: InCommon
- Styles: Bootstrap (optional; implementation sites can style as they wish)

(The software requirements are largely driven by Open XDMoD. See details here: http://xdmod.sourceforge.net/software-requirements.html

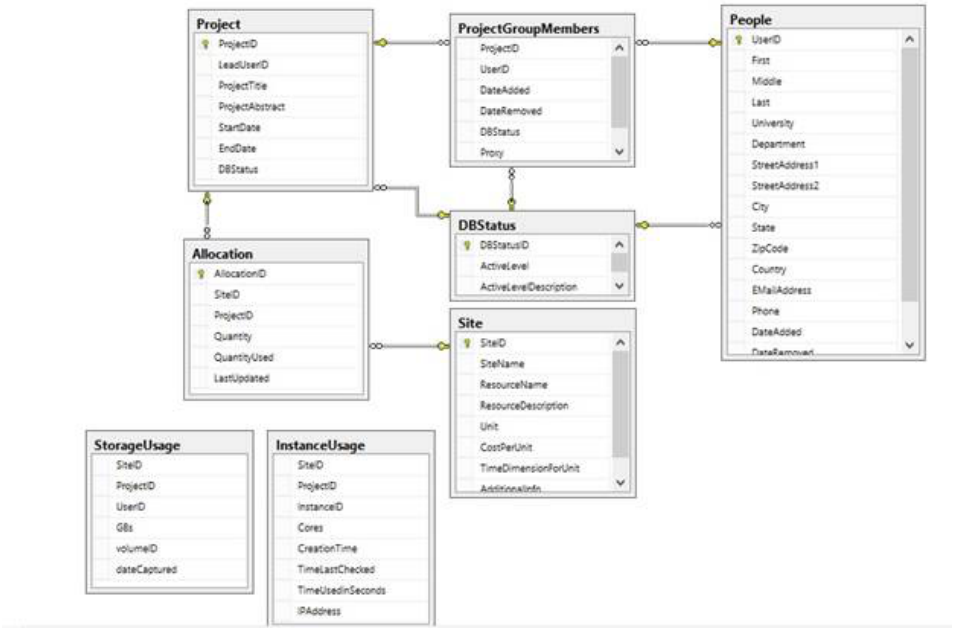
3.1. Allocations & Accounting

Allocations and Accounting components are also being planned in a phased approach. Initially, we will leverage the CAC Red Cloud Accounting and Allocation software we've been running in production for close to five years. We will help the partner sites implement usage accounting that reports back to CU and measures usage of all federation resources. At the same time, we are actively developing a longer term plan for breaking out the allocations, accounting and all critical federation components into redistributable building blocks that others can use to deploy their own cloud federation or to join the Aristotle Cloud

#1541215

Federation. Our goal (part of the project plan) is to have this plan completed, agreed upon and reviewed by a subset of the External Advisory Committee before the end of the calendar year.

A proposed schema for a project, user, accounting and usage database has been reviewed by the team. The goal is to keep this simple yet have all the information we will need for adding and removing projects and users while being able to report usage and allocations. This essentially provides the essential components of the CAC accounting system for cloud allocations and accounting. The next step will be to create the database and begin building stored procedures and API calls to update and record information.



4. Research Team Support

In order for the Aristotle Cloud Federation to succeed, it must support the 7 proposed science use cases effectively and, ultimately, provide options/paths for faster "time to science." In this section, we provide a progress report for each of the 7 science teams. Initially this involves getting project plans in place with each science team and understanding how they currently do their science and their requirements, and to help them understand what their initial allocations are likely to be.

We have created a requirements template and documentation for each team's requirements below.

Use Case 1: A Cloud-Based Framework for Visualization and Analysis of Big Geospatial Data

Hardware

16 node cluster. Each node with 8 cores and 28GB memory.	1 16 core node with 1 high-performance NVIDIA GPU (1,536 CUDA cores and 4GB video memory)
25K CPU core hours	1TB storage

Programming

Java	Python

Software infrastructure stack

Apache Spark on cluster	CUDA on single node

#1541215

Use Case 2: Global Market Efficiency Impact

Roesch: Hardware

1 node cluster with 16 cores and 128GB memory.	40K CPU core hours (depends on how fast it is to start/stop VM)
10TB (but depends on what data will be available)	

Roesch: Programming

OneTick (Proprietary software)	Perl
R	MySQL
GNU parallel	Git

Roesch: Software Infrastructure Stack

Maybe none. OS, assuming Linux for now.

Tiu: Hardware

1 node/>=8 cores, 64GB memory	CPU cycles required: unclear on CPU cycles required
>= 2TB storage for now	

Tiu: Programming

MATLAB (w/ certain toolboxes)	R
MySQL	Julia

Tiu: Software Infrastructure stack

Assuming Linux

Wolfe: Hardware

1 node, 8 cores, ? GB memory (typically 32 or 64GB)	CPU cycles - not sure
Storage - 5TB	

Wolfe: Programming

Python	Perl
7zip	OCR software TBD
SAS	

Wolfe: Software Infrastructure Stack

Linux	Bash

Use Case 3: High Fidelity Modeling and Analytics for Improved Understanding of Climate-Relevant Aerosol Properties

WRF-Chem is needed. Brazier is working to install this and will then go back and iterate with Sara Pryor on the necessary hardware; also we have opened channels to talk to AWS about this, as they have WRF installed as Docker containers which maybe an interesting way to provide scalability using a modern technology vs. more traditional methods involving MPI or OpenMP.

Use Case 4: Transient Detection in Radio Astronomy Search Data

Hardware

○ About 30TB of data, growing at 5TB/year	4GB per core.
Processing takes 1.5-2 core-hours per beam, about 210 000 beams. About 500 000 core-hours per complete reprocessing.	

Programming

PRESTO pulsar package and dependencies	

Software Infrastructure stack

Linux	Windows Server
SQL Server	ASP.NET MVC
Python 2.7	

#1541215

Use Case 5: Water Resource Management Using OpenMORDM

The Reed team's biggest usage has been 526,000 cores. They are a likely candidate to work with AWS. We will work to make the connections with Reed's team and the AWS SciCo Team. We believe this would be an exciting capability, demonstrating value on an important science problem: drought management. We think AWS may consider supporting this use case because their research product is in part for municipalities that could in principle pay for the necessary computing for the product in operations mode.

Hardware

32 compute nodes	512 core, Dual 8-core E5-2680 CPUs @ 2.7 Ghz
128GB of RAM (8GB/core)	10GB ethernet and InfiniBand QDR full interconnect
8TB /home (NFS mounted across 10GB ethernet to all compute from head node)	900GB local /tmp on each node
95TB Lustre /scratch space	

Programming

Intel Compilers (including MKL)	Openmpi 1.6.5
Mathematica	MATLAB
SAS	Boost
cmake	eclipse
hdf5	netcdf
valgrind	visit
zlib	acml
R	BLAS, LAPACK libraries

Software Infrastructure stack

Cloudlaunch or similar, to launch and manage cluster for MPI runs then take it down again	Linux

Use Case 6: Mapping Transcriptome Data to Metabolic Models of Gut Microbiota

Brandon Barker and Angela Douglas continue to discuss requirements.

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

Hardware

8 to 32 nodes, at least 4 cores per node (8 better)	At least 4GB memory per node (8 is better)
At least 50GB storage (100GB better) for every 4 cores in ephemeral disk	5TB in object storage and/or volumes (split yet to be determined)
Core Hours: Probably 96 hours/per week/per node (e.g., we run an ensemble for 96 hours/week). This is really a guess at this point.	

Programming

R	MPI
PostgreSQL	MySQL
Hadoop/Pig/Hive	Spark
Storm	Zookeeper
Puppet	MATLAB

Software Infrastructure Stack

Linux (Centos and Ubuntu)	Python 2.7
Apache	Tomcat
AppScale	

In addition, they are using an open source sensor network framework called GSN, but we have had to fork it for them to make it build in the cloud. Not sure in which category GSN goes -- probably software stack. Also -- there isn't now, but there may need to be, Windows images.

#1541215

5. Outreach Activities

We generated national, international, and local media coverage across 21 media outlets including major IT magazines (CIO, Campus Technology, Government Technology News, NetworkWorld) as well as HPC verticals (ACM TechNews, HPCwire, InsideBIGDATA, Scientific Computing Magazine). A strong social media presence (Twitter, LinkedIn, etc.) was generated as well in partnership with NSF, AWS, Cornell, and our other partners. The federation was also featured at the Cornell SC15 exhibit and Aristotle news flyers were distributed at our exhibit and the University at Buffalo's exhibit.

National Media Coverage

ACM TechNews
Cornell Leads New NSF Federated Cloud Project
http://technews.acm.org/archives.cfm?fo=2015-11-nov/nov-06-2015.html#824843

Campus Technology
Cornell to Lead NSF-Funded Cloud Federation for Big Data Analysis
https://campustechnology.com/articles/2015/11/04/cornell-to-lead-nsf-funded-cloud-federation-for-big-data-analysis.aspx

Cloudwards
Aristotle: Academic Focused Cloud Funded
http://www.cloudwards.net/news/aristotle-academic-focused-cloud-funded-10917/

Cloud Strategy Magazine
Cornell Leads New National Science Foundation Federated Cloud Project
http://www.cloudstrategymag.com/articles/85924-cornell-leads-new-national-science-foundation-federated-cloud-project

Data Center Talk
Cornell to Head the $5M Federal Cloud Computing Program
http://www.datacentertalk.com/2015/11/cornell-to-head-the-5m-cloud-computing-program/

CIO
University researchers get $5M grant to build 'Aristotle Cloud'
http://www.cio.com/article/3003079/iaas/university-researchers-get-5m-grant-to-build-aristotle-cloud.html

Global Wireless Research
University researchers get $5M grant to build 'Aristotle Cloud'
http://globalwirelessresearch.com/news/university-researchers-get-5m-grant-to-build-aristotle-cloud/

Government Technology News
Wisdom of the Clouds: Aristotle Cloud Federation
https://gcn.com/articles/2015/11/05/aristotle-cloud-federation.aspx

HPCwire
Cornell Leads New NSF Federated Cloud Project
http://www.hpcwire.com/off-the-wire/cornell-leads-new-nsf-federated-cloud-project/

InsideBIGDATA
Cornell to Lead Aristotle Cloud Federation for Research
http://insidebigdata.com/2015/11/03/cornell-to-lead-aristotle-cloud-federation-for-research/

NetworkWorld
University researchers get $5M grant to build 'Aristotle Cloud'
http://www.networkworld.com/article/3002615/iaas/university-researchers-get-5m-grant-to-build-aristotle-cloud.html

Next Generation Communications
Cornell Leads New NSF-Sponsored Aristotle Cloud Federation Project
http://next-generation-communications.tmcnet.com/topics/nextgen-voice/articles/412634-cornell-leads-new-nsf-sponsored-aristotle-cloud-federation.htm

Scientific Computing
Cornell to lead $5M NSF Federated Cloud Project
http://www.scientificcomputing.com/news/2015/11/cornell-lead-5m-nsf-federated-cloud-project

Web Host Industry Review
National Science Foundation Sponsors $5 Million Federated Cloud Project
http://www.thewhir.com/web-hosting-news/national-science-foundation-sponsors-5-million-federated-cloud-project

International Media Coverage

CIOL (India)
Cornell University to develop federated cloud
http://www.ciol.com/tag/aristotle-cloud-federation/

ComputerWorld (Australia)
University researchers get $5M grant to build 'Aristotle Cloud'
http://www.computerworld.com.au/article/588491/university-researchers-get-5m-grant-build-aristotle-cloud/

Primeur Magazine (UK)
Cornell leads new National Science Foundation federated Cloud project
http://primeurmagazine.com/weekly/AE-PR-12-15-26.html

TechWorld (Australia)
University researchers get $5M grant to build 'Aristotle Cloud'
http://www.techworld.com.au/article/588491/university-researchers-get-5m-grant-build-aristotle-cloud/?fp=16&fpid=1

Local Media Coverage

Cornell Chronicle
Aristotle: A federated cloud for academic research
http://www.news.cornell.edu/stories/2015/11/aristotle-federated-cloud-academic-research

Cornell Daily Sun
Cornell to Lead $5M Federal Cloud Computing Program
http://cornellsun.com/2015/11/04/cornell-to-lead-5m-federal-cloud-computing-program/

Ithaca Journal
Cornell to head effort to streamline data flows
http://www.ithacajournal.com/story/news/local/2015/11/05/cornell-head-effort-streamline-data-flows/75229980/