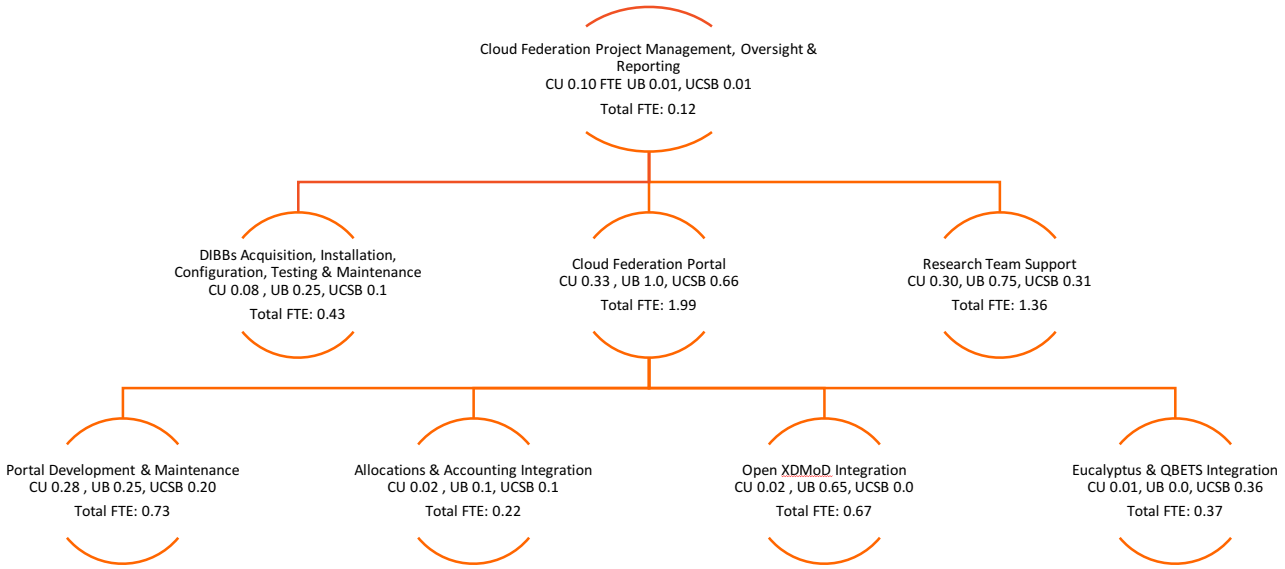
CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation

Program Year 2: Quarterly Report 2

3/31/2017

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

This is the Program Year 2: 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).



ARISTOTLE
CLOUD FEDERATION

Contents

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 award to the Aristotle project (see Section 5.1 for details).

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.

1/31/2017 call topics:

- Desire to add REU students to Aristotle project at UCSB and Cornell and potential benefits to the projects and to the students.
- Integrating OAuth2 into Open XDMoD; Cornell to provide sample Python code as reference for the UB team.
- Cornell plans to build Ceph storage in year 2 for the backend of Eucalyptus 4.4.
- Troubleshooting network speeds between Cornell and UCSB. I2 is not seeing an issue. Tuning the kernel and power management on the node controller is helping.
- UB is testing Ace Computers hardware to see if Ace is viable as a vendor
- How to pursue optimal pricing for Globus endpoints.
- Use of Zooniverse as a Aristotle Citizen Science repository.

2/14/2017 call topics:

- At times, the accounting data is incorrect due to loss of data feeds from one or all of the cloud systems. We plan to devise a system with automatic notification, so a "human in the loop" is not required to identify the loss of accounting data. Monitoring will be added in 2 locations and UB will add Nagios monitoring to check the validity of the data being provided by the API. Cornell will add automated sanity checking to the portal to confirm the data coming from the API looks reasonable.

- Longer term, hopefully by September, XDMoD will become the Database (DB) of record for Aristotle cloud utilization. The beta version is expected at the end of August. For this to work, all sites must commit to follow an upgrade schedule for Eucalyptus software and stay current with updates, subject to appropriate testing. This is the same expectation we will have that all sites running federated XDMoD will be running the same version of XDMoD. Sites must agree to configure and install XDMoD once the beta version is released.
- Note: federated XDMoD will not be ready in August as this will follow cloud beta version testing. If Cornell still requires an API to access raw accounting data to present in the existing portal, UB will need to allocate staff to work on extending the existing XDMoD API to provide this information.
- It would be helpful for Cornell to provide the list of metrics they are using in the portal from UB's current API so that they can confirm they will be available in XDMoD. If there are differences, migrations plans will need to be discussed.

2/28/2017 and 3/14/2017 project status call topics:

- Planning is underway for the April 25th 18-month review. Amy Walton participated in this month's Aristotle Science Team Advisory Committee (STAC) meeting and provided information on the review. She also complemented the team for organizing and leading the 1st NSF Data Infrastructure Building Blocks PI Workshop (DIBBs17) which was regarded as highly successful by NSF as well as the workshop participants.
- All project reports (annual, quarterly, monthly, and supplemental) will be available online for the 18-month NSF review committee at https://federatedcloud.org/reports/. STAC meeting minutes will be available at: https://federatedcloud.org/science/advisorycommittee.php.
- The Blue Waters Workload Analysis carried out by the UB XMS (XD Net Metric Services) team at the request of NSF from July 2016-July 2017 delayed the Aristotle federated XDMoD target release date from 12/2016 to 7/2017.
- UCSB is troubleshooting a problem with slow snapshotting. Snapshotting a 400GB volume in the UCSB cloud took hours and sometimes resulted in error. While the snapshot error was fixed in the official Eucalyptus 4.4 release, slow performance was attributed to uploading the snapshot to object store. Cornell implemented a workaround for its single availability zone by setting *<cloud>.storage.shouldtransfersnapshots* to false so snapshots are not uploaded to object store. However, uploading to the Ceph cluster in the UCSB cloud is significantly slower than Cornell's (8.3MB/s vs. 50MB/s). UCSB and Cornell are comparing notes on their configurations of the Eucalyptus cloud and Ceph cluster to see what could attribute to the differences in performance.
- The RT help desk accounts are being set up for UCSB and UB (central RT is at Cornell).
- Different software capabilities at the 3 sites is a plus; it will provide Aristotle users with more diverse resources to choose from, e.g., finance software at UB, MATLAB cluster software at Cornell, etc.
- Launch of the new MATLAB capability at Cornell has occurred. Users can now fire up 28 MATLAB Distributed Computing Server (MDCS) workers. This capability will be expanded in the future.

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

2.1 Hardware Acquisition

- UB received their 10Gb Dell network switches and is configuring them with UB Campus switches to provide 10Gb cloud connection to the public network. Four new node controllers were purchased from Ace Computers which increased the capacity of the cloud by 112 cores.
- Cornell ordered Ceph storage, 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. The Ceph storage is on back order due to a worldwide SSD shortage; we expect it to arrive in early April. This hardware will allow Cornell to build their production Eucalyptus 4.4 cloud with Ceph storage.
- UCSB ordered their year 2 hardware from HPE and Dell and resolved failed hardware components on Eucalyptus NC and Ceph OSD servers. Specific details on UCSB hardware will be available soon.

2.2 Software Installation, Configuration, and Testing

- Cornell upgraded their test cloud to the official Eucalyptus 4.4 release, pulled down nightly builds, and developed and tested configuration scripts for Globus single sign-on. Users can now log into their Aristotle accounts on the Eucalyptus Console via Globus Auth. Once logged in, they can access cloud resources as well as generate AWS-style access and secret keys to be used with command line tools or AWS API calls. Cornell is writing instructions for deploying these scripts and will work with UB and UCSB to configure their clouds to support Globus single sign-on so that Aristotle users can access all 3 clouds in the federations using their Globus credentials. Cornell is also installing and testing XDMoD on both the test and the production cloud. In addition, 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.
- UB updated their cloud infrastructure to Eucalyptus version 4.3.1 and applied CentOS security updates where necessary. They are investigating Ceph configurations and will purchase hardware using the remainder of year 2 funding (supplemented with UB Center for Computational Research funds). With the impending removal of support for SANs in future versions of Eucalyptus, all Aristotle sites plan to migrate to Ceph.
- UCSB is planning their upgrade to Eucalyptus 4.4 and installing XDMoD.

The infrastructure planning table was updated this quarter:

	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	4.3.1	4.3
Globus	Yes	Yes	Planned
InCommon	Yes	Yes	Yes

Hardware Vendor	Dell	Dell/Ace	Dell
# Cores	*168	**256	140
RAM/Core	4GB/6GB	up to 8GB	up to 9GB
Storage	SAN (226TB)	SAN (336TB)	Ceph (288TB)
10Gb Interconnect	Yes	Yes	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)	** 256 additional cores augmenting the existing Lake Effect Cloud (424 total cores)	

2.3 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 team to test their OAuth2 implementation. Version 4.4 has been released and implementation plans are underway at all 3 sites.

- **Supercloud**
  Nothing new to report this quarter.
  .

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 this quarter; 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 UB and UCSB, and will be implemented at Cornell when the latest version of Eucalyptus is installed on Cornell's production hardware.

The usage graph is being monitored to ensure data is being collected consistently from all sites. We are investigating software solutions.

Changes were made to the portal planning table this quarter:

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 and after implementation in Science Use Cases, 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 science group visits.	Hold 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.
User Authorization and Keys			
Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
10/2015 – 1/2016	**2/2016 – 5/2016**	**6/2016 – 3/2017**	**4/2017 – 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 – 3/2017	**3/2017 –5/2017**	**6/2017 – 10/2017**	**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.	Display usage and CPU hours by account or project 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. Buffalo also completed a redesign of the XDMoD data warehouse to support cloud metrics and is switching XDMoD to use this implementation.	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. Open XDMoD (v.6.5.0) has been released and is available 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.	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. Federated XDMoD is being updated to support the new cloud schema.	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.

3.1 Software Requirements & Portal Platform

In addition to 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

Planning for completion and cloud beta implementation of Open XDMoD is detailed in the table below with an August target.

		Feb		March				April				May				June				July				August			
Task	**Order**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**
Update cloud schema docs	1																										
Implement static dimension tables	2																										
Document JSON import schema	3																										
Documentation Updated	3M		X																								
Define file-based API	4																										
ETL config file references	7																										
Subquery support	8																										
Improve start/end/number handling	9																										
Action date chunk handling	11																										
ActionState	12																										
ETL Required Additions	12M										X																
XDMoD cloud event mapper	13																										
Updates to StructuredFile ingestor	14																										
Build single record ingestor	15																										
Build directory crawler	16																										
Aggregator	17																										
Testing	18																										
Eucalypus Data Ingested	18M															X											
Migrate from modw.jobfact	21																										
Implement realm & metrics	22																										
Implement drilldowns	23																										
Testing	24																										
Cloud Beta (Accounting)	24M																								X		
Summarization Infrastructure	25																										
Develop summarization tools	26																										
Summary docs in mongodb	27																										
Ingest summaries into XDMoD	28																										
Display perf data via existing Job Viewer	29																										
Cloud Beta (Performance)	29M																								X		
ETLv2 Housekeeping																											
Cloud Data																											
Cloud Realm																											
SUPReMM																											
Milestone																											

Modifications to the XDMoD schema to support cloud data were finalized. Improvements to the Eucalyptus log scraper were made to improve coverage of desired events such as detaching a volume from an instance. Preliminary modifications to the XDMoD ETLv2 process are underway to support

ingesting cloud data.

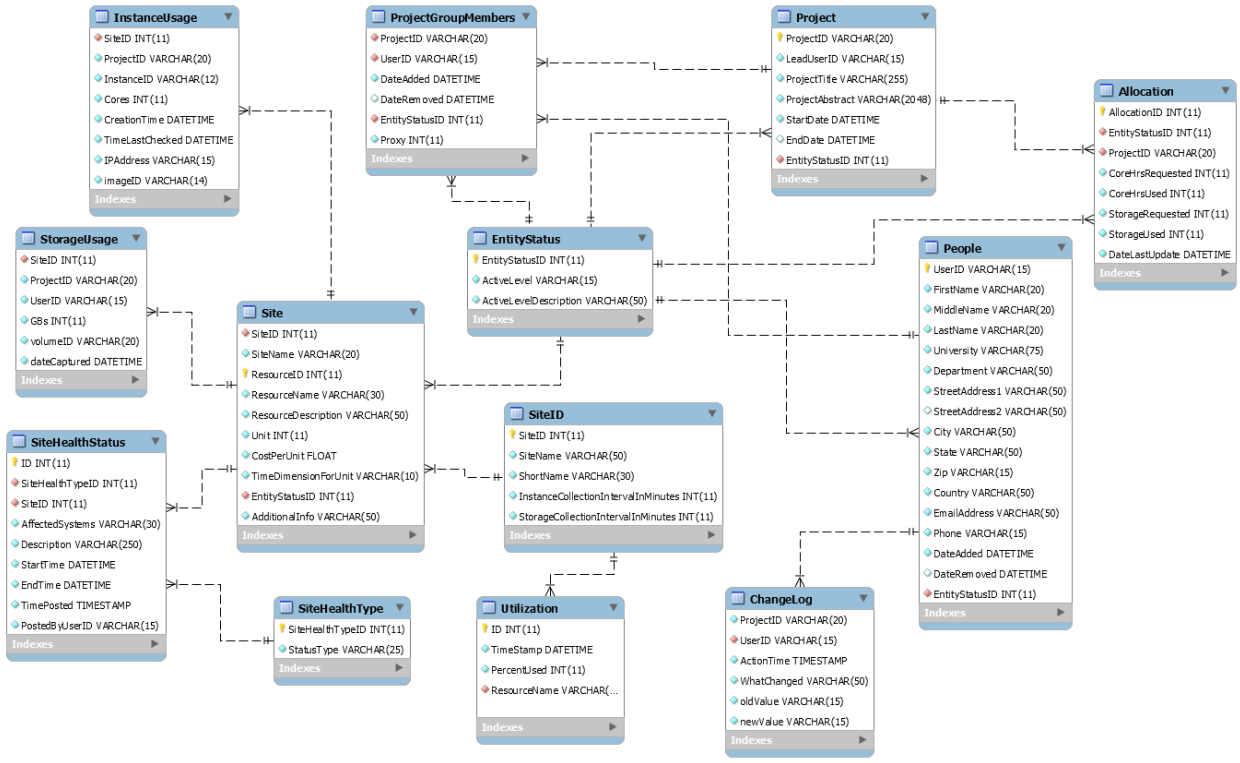
UCSB worked on improving the quality of DrAFTS predictions. Previously, they were seeing a random error rate of ~4%; that error rate has been reduced to under 1%. A more complete validation of these new results was submitted to *HPDC'17 (26th International ACM Symposium on High-Performance Parallel and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

In addition, a new set of experiments with the Globus Genomics group is also underway at UCSB. 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 was started. 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. At the same time, we are developing the portal dashboard, which will display usage data on a project level to science team members.

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



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.
- Bennett Wineholt and Brandon Barker (Cornell) continued their work on MPI. We now have a working virtual cluster of 8 nodes of 8 CPUs each, on which MPI jobs run. We have run multiple Docker Swarm clusters that execute MPI with a workable level of performance. We are currently scripting cluster setup to be more automatic and seeking opportunities to improve system performance.
    - The underlying mechanism for distributing images is based on Docker and, once performance testing and characterization is complete, the next tasks are scripting the creation and deployment of virtual clusters, and documenting the work. This is a key milestone for our science support.
    - We have made clear to the science teams that this is not a replacement for large-scale HPC, but is aimed at being a quick way of getting on-demand science results using code which requires moderate- or small-scale MPI.
    - The major characterization goal is evaluating which jobs can be run well on a virtual cluster of this sort, compared to ones which require a traditional HPC cluster.
    - The latest advances include the use of s3fs for high-speed parallel writes for code output.
- An Aristotle REU supplemental proposal (Cornell and UCSB) was completed and submitted for NSF review.
- An Aristotle Science Team Advisory Committee meeting was held March 20th (see meeting minutes in Section 4.3).

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. The Big Geo team is working on interface testing and tuning. Experiments are currently running to get scalability numbers for the change detection algorithm on the UB Lake Effect cloud and will be submitted as a paper to the ACM GIS conference. Additionally, we are finalizing the webGlobe website for the next month's demo and for future release to scientists.

Use Case 2: Global Market Efficiency Impact
UB finance researchers finished aggregating the tick-by-tick data for one of their projects. They are beginning to analyze this data and are doing outreach to other finance faculty and PhD students to assess interest in learning about how to use this framework. The finance researchers imported all NYSE Trade and Quote (TAQ) data up 2015. TAQ contains intraday transactions data (trades and quotes) for all securities listed on the New York Stock Exchange. This data will be used to compute the first of four measures of efficiency. Up to 30 more stock pairs have been integrated, extending the original research sample.

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

MPI work continues as described above. Brazier, Barker, and Wineholt met with Sara Pryor's new postdoc, Tristan Shepherd. A very productive meeting covered software and hardware requirements, benchmarking, and progress and future plans. To our knowledge, WRF-Chem (i.e., the chemistry-enabled version of WRF) has never been successfully run on a cloud. This is particularly true for clouds such as Aristotle and Jetstream that are not configured out-of-the-box to operate with massively parallel computational codes, and are default to work on a single computational node. Consequently, we encountered a number of technical challenges in running the Weather Research and Forecasting (WRF-Chem) model on Aristotle. For these and other reasons, we are currently building capacity using the physics only version of WRF. This allows us to fully and comprehensively evaluate the simulations conducted on Aristotle and to work with a more stable and well documented version of WRF (i.e. the physics only version) while addressing very important questions of relevance to climate science.

Our modified science objectives for the near-term are to:
(1) Comprehensively evaluate the platform dependence of climate simulations. We will evaluate the reproducibility of climate simulations derived from operation of WRF on different platforms (on a cloud relative to a traditional HPC platform, i.e., the DOE NERSC Cray system Cori). This will be conducted under separate funding.
(2) Evaluate the fidelity of the climate simulations using a suite of remote sensing and in situ observations.
(3) Undertake a novel study of the impact of wind turbines on downstream climate using a module within WRF that simulates the momentum extraction and turbulence introduction downstream of major wind turbine deployments.

Progress: The WRF Preprocessing System (WPS) and WRF v3.8.1 were installed and compiled with parallel NetCDF on Aristotle. After the Cornell CAC team completed troubleshooting the network file server, benchmarking was completed in March. The WRF wind turbine input files have been prepared, along with the input meteorological files. We anticipate starting the production science case in April. The first case is a simulation of the year 2008, at 12 km resolution. This will serve as both a test run of the capabilities of running WRF on a cloud-based system, and as a means to assess the fidelity of the regional climate.

Future Plans: After running the 12 km simulation for the entire calendar year of 2008 over the U.S. east of the continental divide, we plan to re-run the simulation with a high-resolution nest of 4 km over the state of Iowa (a state with high wind energy penetration). This simulation will include all wind turbines in the state and will examine the degree to which the flow field is modified downstream (i.e., in the state of Illinois). This will address a key science question: how does this valuable renewable energy source reciprocally interact with the atmosphere?

Use Case 4: Transient Detection in Radio Astronomy Search Data

Cornell astronomers Cordes and Chatterjee plan to start processing with the Jetstream allocation in by May; 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
Brazier and Barker met with the Reed group for their weekly in-person and Zoom group meeting. Barker demonstrated the Aristotle MPI cluster and we discussed the best way forward. The Reed group has provided three test software suites covering the range of functionality for their initial work. Of those, the Platypus software has been built and tested in the virtual cluster. The remaining two batches of software will be built and tested, including benchmarking against a cluster that Cornell CAC maintains for the Reed Group. The file system work for MPI jobs is nearly complete. We will then run benchmarks of Project Platypus (the Python version of OpenMORDM) to investigate scaling on science cases.

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. The whitefly paper has been resubmitted to the *Journal of Bacteriology* after addressing reviewers' comments. The research team plans to submit another modeling paper on 3 other insects (xylem feeders). Computational development continues on the modeling and when the exploratory studies are completed, a larger system will be modeled.

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. The Zooniverse image release is still in the works. There is a substantial amount of data processing that needs to be completed. In March we decided to begin to work on a new version of the TensorFlow model that is trained "from scratch." While the current implementation uses transfer learning, the team hypothesizes that better accuracy can be achieved from a fully trained model. To do the training will require substantial Aristotle capacity which is currently on order.

Agricultural Food Security Project: UCSB/Sedgwick Reserve used 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. Work on the new irrigation plan will commence when the weather clears. We identified a couple of undergraduates for the REU projects that the Aristotle team proposed.

We are working on deploying version 3.0 of the sensor platform. The current version (which is attempting to use weather information to control the sensing rate) is able to survive to approximately 10:00 p.m. each evening before the battery expires. We have a new battery discharge model and it seems to be functioning properly. Readings from the following site http://169.231.235.16 which is hosted on Aristotle show the current soil moisture reading taken from the UCSB-developed sensor platform. Note that the time frame

spans a 24-hour period. This is a substantial improvement over the previous version. The plan is to deploy this platform at several sites in the coming weeks to be able to track the full growing season. We will also continue to use this data to improve water usage.

Citrus Packline Data Analysis and Delivery (New Collaboration): 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.

4.3 Science Team Advisory Committee Meeting Minutes

Science Team Advisory Committee (STAC): 3/20/2017 Meeting Minutes
Adam Brazier, Aristotle science team lead

Invited to 3/20/2017 Science Team Advisory Committee Meeting *(attendees italicized)*

Cornell University (CU):
James Cordes cordes@astro.cornell.edu
Adam Brazier brazier@cornell.edu
Angela Douglas aes326@cornell.edu
Nana Ankrah na423@cornell.edu
Brandon Elam Barker brandon.barker@cornell.edu
Sara C. Pryor sp2279@cornell.edu
Tristan Shepherd tjs346@cornell.edu
Patrick Michael Reed pmr82@cornell.edu
Julianne Dorothy Quinn jdq8@cornell.edu
Resa Reynolds rda1@cac.cornell.edu
Susan Mehringer shm7@cornell.edu
Paul Redfern red@cac.cornell.edu
David Lifka lifka@cornell.edu

University at Buffalo (UB):
Tom Furlani furlani@ccr.buffalo.edu
Varun Chandola chandola@buffalo.edu
Cristian Tiu ctiu@buffalo.edu
Dominik Roesch drosch@buffalo.edu
Brian A. Wolfe bawolfe@buffalo.edu

University of California, Santa Barbara (UCSB)
Rich Wolski rich@cs.ucsb.edu
Andreas Boschke andreas@cs.ucsb.edu

Kate McCurdy kate.mccurdy@lifesci.ucsb.edu

National Science Foundation
Amy Walton awalton@nsf.edu

NSF - *Amy Walton, NSF program manager*
- The 1st NSF Data Infrastructure Building Blocks PI Workshop (https://dibbs17.org) was led by Cornell as a supplement to the Aristotle award and was a tremendous success. There was high energy on the part of the 72 participants and the resulting data and insights produced will be invaluable to NSF as we plan for the future of research and data cyberinfrastructure. The final workshop report will be available by the end of March *(see Section 5.1)*. Thank you all!
- Your 18-month review is scheduled for the morning of April 25th. Focus is on a working prototype and showing you are well on the way to success. The goal for these projects was to provide a robust and shared CI capability. Plan on 60-90 minutes of top level presentations/demos followed by an extensive questions/answer and discussion period through late lunch. Cover 5 areas: (1) rationale for the project, (2) CI approach used and how the prototype is doing, (3) project management, i.e., who is doing what and how it is being measured, (4) what are the next stages of the project, i.e., if we fund you for years 3 and 4, what are you going to accomplish? (note: this could include the possibility of adjusting milestones for new innovations or broader impacts), (5) when this award is completed, what will the community have that it doesn't have now? One week prior to April 25th, NSF should receive a reference package on the project with the PowerPoints, a copy of the execution plan, any report updates. This will be distributed by NSF to the panel. Hit the highlights. Explain what the prototype is, what is working, how what you have is helping the scientific community. Provide a well-organized package. Including some key application examples seems to work for some projects.

Infrastructure Update – *Adam Brazier (CU)*
- We're moving to Eucalyptus 4.4. We were successful in getting HPE to build OAuth2 support into 4.4 and testing it. This will provide us much more functionality. You will be able to login using your own institutions' authentication as well as InCommon. If you're using XSEDE resources, this is what they are using too, so it will be like one ticket access to potential resources.
- Year 2 hardware installations and purchases are underway. The University at Buffalo has added 112 cores and is upgrading their network. Cornell is adding Ceph storage since that was a greater need than more cores. UCSB, on the other hand, has ample storage, so they're adding cores since usage of their current cores is maxed out. Heterogeneous platforms and software at the various sites is a plus since it provides users with more alternatives to address their needs.

Portal Update – *Susan Mehringer, portal team lead (CU)*
- The Aristotle portal has been up for quite a while and is located at https://federatedcloud.org. If you have comments or suggestions, send them to help@federatedcloud.org. A status graph for all 3 sites is on the portal which shows the availability of cores. Monthly, quarterly, and annual reports are also available. They are password protected, so contact us via "help" if you'd like access.
- OAuth2 has been implemented on the portal, and will be used for all sections and functionality that require authentication.
- The next major step is to integrate allocations and accounting information into the portal, by project and by person. Information will include, e.g., project members, usage, allocations balance, and storage usage.

Science Team Update - *Adam Brazier, science team lead (CU)*
- We're using Slack for communications which has been incredibly useful given the distributed nature of our project. If you are not a member of our Slack channel, email brazier@cornell.edu and ask to join. Our RT ticket tracking system is another valuable tool for science users (help@federatedcloud.org).
- Allocations are somewhat relaxed at this point; a formal allocations process will be implemented in the future.
- We're close to completing the development of a virtual cluster capability for MPI and OpenMP (we just need to improve the disc I/O speed). The intent is to provide virtual cluster access for modest size jobs (not thousands of cores). On-demand access to virtual clusters in the cloud vs. waiting in a queue for a very large-scale system will improve time to science for users with modest needs.
- We applied for and we're approved for 2 million virtual core hours on Jetstream.
- Science teams will be asked to provide science highlights for our upcoming 18-month NSF review. Lifka emphasized the importance of highlighting how the Aristotle project has helped (or will help) the science teams do things differently or better than in the past or, in some cases, with faster time to science.

Individual Science Team Updates

Use Case 1: A Cloud-Based Framework for Visualization and Analysis of Big Geospatial Data - *Varun Chandula update (UB)*
- We plan to add more analysis tools to our framework, add more simulation data, and make the system more robust.
- Next steps will include discussions about how to submit/track jobs which will be important in opening up the capability to the broader community.
- We can provide a demo for the 18-month review.

Use Case 2: Global Market Efficiency Impact - *Dominik Roesch (UB)*
- We've set up a framework for the analysis of high frequency trading data (20TB+) and will provide this capability to 3 PhD researchers this summer with the future goal of opening up this capability to researchers at other institutions in the future.

Use Case 3: High Fidelity Modeling and Analytics for Improved Understanding of Climate-Relevant Aerosol Properties - *Tristan Shepherd (CU)*
- We're had a productive collaboration with CAC and are very close to running WRF-Chem (a parallel version Cornell built for the cloud) on multi-year, high-resolution data across North American.

Use Case 4: Transient Detection in Radio Astronomy Search Data – *Adam Brazier (CU)*
- We're preparing for cloud usage and will do early runs on Jetstream. Cornell and a global team of astronomers previously uncovered the source of a "fast radio burst." We plan to expand the analysis of that data beyond the previously analyzed data to explore and, hopefully, produce new insights.

Use Case 5: Water Resource Management Using OpenMORDM – *Julianne Quinn (CU)*
- We're working to benchmark and test the scaling of OpenMORDM and its equivalent in the Python equivalent Project Platypus. Scientific output will commence when the MPI virtual cluster work is completed.

Use Case 6: Mapping Transcriptome Data to Metabolic Models of Gut Microbiota – *Brandon Barker (CU)*
- Several large instances were created (Windows and Linux) and used for modeling and genetic analysis. This resulted in a paper which will be published in the *Journal of Bacteriology* and a presentation that will be delivered at the *Data-Driven Biotechnology Conference* in Copenhagen (May 7-11, 2017).
- Computational development continues on the modeling, and usage is expected to ramp up when we're ready to model a larger system based on the initial exploration studies currently underway.

Use Case 7: Multi-Sourced Data Analytics to Improve Food Production – *Kate McCurdy (Sedgwick Reserve/UCSB)*
- Agricultural soil moisture monitoring and drought monitoring of oak trees will continue when soil sensor failures due to excessive moisture as remedied.
- The "Where's the Bear" project will continue when storm-impacted sensors are back up and running. This system is still using the Caffe framework and the Aristotle cloud to process all camera data from the field. The framework sorts images into species. We are still struggling with empty frames and frames with birds and will be honing this species recognition capability. We are using a citizen science site (Zooniverse) to pull data out of the Aristotle cloud and will be using volunteer crowdsourcing to verify the accuracy of the computer-generated data in the coming weeks. Ecology researchers are interested in using the first workflow to publish climate change impacts on wildlife. UCSB Computer Science is working with us to eventually launch our own repository site to manage the whole process of data acquisition, storage, and dissemination. Important note: this project has sparked a new scientific collaboration (not part of the original Aristotle proposal) with Sedgwick researchers who plan to perform latent species identification and analyze fishery health from images from SE Asia.

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. This workshop was a supplemental award to the Aristotle project. 37 posters on significant DIBBs success/innovations were featured at the workshop 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."
- The "Final Report: 1st NSF Data Infrastructure Building Blocks PI Workshop" was released by Cornell on 3/28/2017 (see https://dibbs17.org/report/DIBBs17FinalReport.pdf).