ARISTOTLE
CLOUD FEDERATION

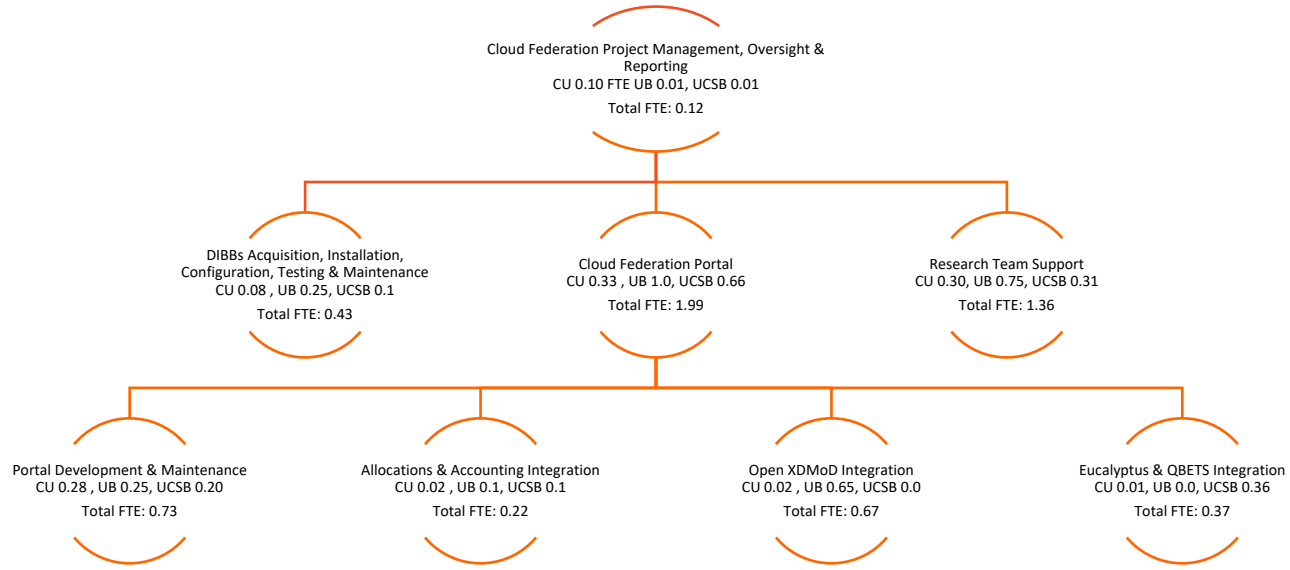
CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation

Program Year 3: Quarterly Report 4

9/25/2018

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

This is the Program Year 3: Quarterly Report 4 of the Aristotle Cloud Federation team. We report on plans and activities for each area of the project Work Breakdown Structure (WBS).



#1541215

Contents

#1541215

1.0 Cloud Federation Project Management, Oversight & Reporting Report

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

1.2 Project Change Request
No new project change requests were made this 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
- Lifka continued strategic discussions this quarter with the 3 major public cloud providers and academic leaders who are knowledgeable about cloud opportunities and challenges in preparation for an NSF-funded summit on "Creating an Open Cloud Marketplace" to occur in the Washington, DC area this Fall or Winter.
- In August 2018, Lifka met with Dartmouth VP/CIO Mitchel Davis, director of research computing George Morris and the Dartmouth Research Computing staff in Hanover, NH to discuss Aristotle and Dartmouth's interest in joining the federation. In addition to joining as a self-funded research institution, Dartmouth plans to facilitate Aristotle outreach to R3 doctoral universities (led by Cornell's deputy director Rich Knepper) to gauge their interest in participating in the federation with subscription-based consulting support.
- The National Academies of Science released "Opportunities from the Integration of Simulation Science and Data Science: Proceedings of a Workshop" this quarter. Aristotle co-PI Tom Furlani was a reviewer of this report and also participated in the workshop on a "Services, Usage Models, and Economics" panel with AWS and Microsoft**.** The report is available online at https://www.nap.edu/read/25199/chapter/1.
- Furlani and Lifka participated in the NSF DIBBs18 Workshop and presented a paper and poster: https://federatedcloud.org/papers/1541215%20Aristotle%20Cloud%20Federation%20DIBBs18%20poster.pdf.

1.5 Project Status Calls

7/17/2018 status call:

- Discussed the open cloud marketplace concept and plans for a "big 3" (AWS, Google, and Microsoft) summit with academe and NSF representatives discussing the opportunities and challenges of using public cloud from an NSF community needs perspective. Amy Walton requested that NIH participate as well. Discussion included location, hotel, logistics, etc. (to be handled by Fratkin Associates) and design of the event website/content (Redfern). Both NSF and the cloud providers agree that if doing long tail science and data analytics, resource-wise public cloud has superior technology solutions; however, there are challenges in terms of pricing transparency, usage tracking, usage limits, overall integration into the national cyberinfrastructure allocations framework, etc.

- Debriefing on the NSF DIBBs18 workshop which went well. It followed the same successful workshop framework that was developed by the Aristotle team for the DIBBs17 workshop. 76 cyberinfrastructure experts and scientists participated.
- Discussed preparing demos for PEARC18: Weather Research and Forecasting (WRF) in a container demo (Cornell) and CENTAURUS cloud service for k-means clustering demo (UCSB).
- Walle and Mehringer (Cornell) made improvements to the Aristotle portal so that the user dashboard loads much faster. Work continues fleshing out forms such as adding people to projects, etc.
- UB is verifying the data coming out of OpenStack logs for presenting basic cloud metrics. They are bringing some of that data into XDMoD for display so you can look metrics such as the number of VMs running, VM start/stop, memory used, cores used, and storage volumes. The goal is to see OpenStack data from all 3 sites once Cornell and UCSB have their OpenStack deployments in full production.
- An REU student is doing particularly well working on the development of a new radio astronomy data pipeline at Cornell. The pipeline will be containerized for use beyond the federation.
- A call is planned with Scala Computing. Scala develops on-demand, hyper-scalable HPC for complex and time-sensitive forecasting (such as the WRF-based use case).
- In general, domain scientists feel that software is hard to install in the cloud. Containers are the way to go. They're lighter and easier than a virtual machine and can run in lots and lots of places. They can also run on both MAC and Windows machines.

7/31/2018 status call:

- Further discussion on the "Creating an Open Cloud Marketplace" summit. It is important that decision makers as well as technologists who understand the provider's APIs participate. The goal is to try to convince the cloud providers to work closely with the community in order to help researchers navigate and use their resources.
- When you buy products from public cloud vendors, their products can change in ways that are hard to detect until a later time. For example, AWS recently changed their approach to the spot market. UCSB is working on a paper that analyzes this change (being written almost entirely by an REU student) and its impact. The simple analogy is, if you buy a Cray, you get a Cray, but if you buy a public cloud product, it may change midstream. The transparency of what you're buying is not as clear.
- UB is wrapping up a beta version of OpenStack cloud metrics so that local federation sites will be able to start collecting cloud data when they're ready.
- An infrastructure team call is scheduled to discuss plans for implementing federated authentication.
- The infrastructure team had a concall with Globus technologists about the capabilities of the new Globus AWS S3 Connector. Subsequently, Cornell tested this product on Euca. The AWS S3 Connector provides a way to transfer data directly to Cornell's cloud. Cornell is working on setting up a management method that allows scientists to use this slick capability.
- The radio astronomy data pipeline continues to mature and the development team are close to getting it on GitHub. The first target is to provide scientists with the pipeline framework in a container so that if they want to run the same pipeline that Laura Spitler used to discover FRB 121102, they can.
- Tristan Shepherd submitted an abstract that describes running WRF in a container on different underlying systems. This should raise our cloud and container visibility while identifying potential reproducibility issues.

#1541215

- The water management resources use case team continues to make progress on getting multi-instance MPI in a container operational. Even if it's not super-fast, it will be super portable.
- UCSB believes they can spawn workers from their CENTAURUS cloud service for k-means clustering in AWS Lambda serverless computing. Lambda enables you to run code without provisioning or managing servers; you only pay for the computer time you consume – there is no charge when your code is not running. UCSB said AWS Lambda can execute a function much cheaper with very independent lightweight workers. They don't see a lot of Lambda for science which is why they'd like to bury it in a service like CENTAURUS. They are currently working on a citrus frost prevention application with a Lambda backend and CENTAURUS. This app could demonstrate an incredibly cost-effective use of AWS. UCSB decided to use Flower (a web-based tool for monitoring Celery clusters) for CENTAURUS rather than XDMoD which is more of an accounting collecting tool than a real time reporting tool.
- The successful soil moisture monitoring for grapes project that demonstrated how SmartFarms can save water has ended. A new land remediation for ranching project will begin soon at the Sedgwick Reserve.

8/14/2018 status call:

- Discussed the challenge of scheduling scientists on short notice for demos at the 36th month review. Possibilities include a finance demonstration and SmartFarm applications.
- The compatibility issue between Docker and Singularity can be 5%-90% level depending on what you're doing; they're not really compatible unless you really know what you're doing.
- Brazier will apply for an extension to the Aristotle Jetstream allocation.
- UB built a Docker container with RStudio which the UB Statistics Dept. deploys on the UB Aristotle cloud and the university cloud for hundreds of statistics students. They are happy as clams with the container and are now trying to containerize all sorts of software. This is an example of how the Aristotle project is impacting the campuses at large (beyond the Aristotle project science use cases). A similar cross-institutional Aristotle impact has occurred at UCSB, increasing cloud visibility and acting as a catalyst for new cloud projects across the campus.
- Similarly, more than 30 faculty, staff, and student researchers affiliated with the Cornell Institute for Social and Economic Research (CISER) have shifted some of their workloads from clusters to Cornell's Red Cloud and find the on-demand resource effective and timely in meeting their computational and statistical analysis needs. Social sciences computing can be an ideal fit for the cloud considering statistical software is often used in a pleasingly parallel mode at moderate scale. Researchers especially like the availability of large memory instances: 28 core instances have 192GB RAM. Cornell CAC staff built 5 ready-to-use cloud images for the researchers: a Windows image with IBM SPSS, MATLAB, Mathematica, R, Rscript, SAS, and Stata/MP and 4 custom Linux images that feature Python, Miniconda, PostgreSQL, DataGrip, Kate editor, QIME2, and PICRUSt.
- Cornell deployed new nodes and has OpenStack and networking working right now on a test cloud. They will be deploying the OpenStack production cloud soon.
- The portal team talked about how to flesh out the portal dashboard and integrate what is going to be happening with Globus accounts across the federation at the same time.
- UB is getting ready to use Open XDMoD 8.0 and making sure that the cloud metrics are correct. OpenStack will need a patch which UB has developed to collect the logs that XDMoD needs. The patch will help provide access to command line tools so we can dump the proper log files for the projects as an administrator.

#1541215

9/11/2018 status call:

- Dartmouth participated in its first Aristotle call and will participate in future status calls. Research systems engineer Bill Hamblen will be their technical lead working with Steven Lee at Cornell to bring Dartmouth into the federation portal and accounting system. Lee will also help Dartmouth with authentication and Ceph. University of Michigan VP and CIO Ravi Pendse also expressed interest in joining the federation, but the project team prefers to focus on getting Dartmouth up and running before bringing other institutions into the federation. Dartmouth will enrich the federation with new science use cases, including digital humanities.
- Many use cases are containerized at this point. One container with multi-instance MPI is being tested on Jetstream because the water management use case scientists need lots of cores. This application shouldn't be too sensitive to latency in the cloud. Test results will be available soon.
- The Cornell portal team is working on getting the accounts and users created on each site. The scripts are written to do this. Walle will work with Mehringer to get this on the web site. She needs everyone to login to federatedcloud.org so she can get the users into the right projects and/or groups. UB can get the users and groups when the web form is available. UB will then test what Walle has.
- High availability is working at Cornell. The last step is to switch over our controlled clusters to our production Ceph. The connection is currently timing out to our production Ceph cluster. We believe there's a bug somewhere in the new version of Red Hat because it worked fine in a previous version.

1.6 Supplement

Cornell PI Lifka and the Aristotle co-PIs submitted a proposal to NSF this quarter titled "The Aristotle Marketplace Investigation (AMI): Supplement to the Aristotle Cloud Federation Project" to support the investigation of extending the Aristotle federation to include potential public cloud offerings. This supplement will support efforts to secure the cooperation of public cloud providers through a summit with Aristotle partners, public providers, the NSF, and the NIH, focused on identifying the gaps between research needs and public offerings. The supplement will also support the extension of the Aristotle portal to incorporate new entrants outside of research institutions and new tenders (Aristotle allocations, XSEDE allocations, vendor credits, or cash) which allow researchers to make use of multiple resources in order to secure cloud time. The $998,679 budget was approved.

Deliverables include:

- Leadership of summit activities for public cloud integration
- Development work on portal and database to allow the use of new types of resources and allocation formats.

Lessons learned in the process will be documented and disseminated to the cyberinfrastructure community.

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

2.1 Hardware Acquisition

- There were no hardware acquisitions this quarter.

NSF
#1541215

2.2 Installation, Configuration, and Testing

- Cornell reconfigured its network to create a network infrastructure to accommodate both the existing Eucalyptus cloud, Ceph cluster, and the new OpenStack controllers and compute nodes. The new network configuration allows OpenStack deployment to move forward and users to transition from Eucalyptus cloud to OpenStack cloud when the deployment is completed. Cornell completed its initial installation of OpenStack cloud. The infrastructure team is doing internal testing and the science team is identifying the first Aristotle projects to migrate to the OpenStack cloud.

- UB continues to migrate hardware from the Eucalyptus infrastructure to OpenStack. They are currently trying to address a single point of failure within their networking stack and trying to come up with a redesign of some of the networking. Specifically, they would like to use tagged VLANs to support multiple provider networks on the same network gear. This requires them to add some hardware to their development cloud and perform the testing there. They plan to have this figured out in October 2018.

- UCSB's OpenStack installation and configuration continues. UCSB has a small OpenStack installation that will ultimately grow to replace the Eucalyptus Aristotle cloud. There are essentially two issues that are being resolved. The first is with OpenStack networking. The "standard" demo networking configuration does not provide network isolation between OpenStack users. UCSB's security team allows this configuration for testing purposes but will not approve production usage. UCSB has worked out a networking configuration for OpenStack that is acceptable to the security compliance office but scripting this configuration so that production administration staff can implement it is still an ongoing activity. Secondly, the Eucalyptus Aristotle cloud has grown in scope to the point where the local UCSB AZs are larger than the NSF-funded AZ. The UCSB user community understands AWS and appreciates the AWS compatibility that Eucalyptus provides. Thus, migrating them to OpenStack has proved to be a labor-intensive challenge. UCSB Aristotle personnel have been working with Aristotle users to identify a schedule and priority for switch over to OpenStack. The initial non-NSF funded users will likely be the Statistics Department which uses Docker to implement sandbox environments for individual students.

2.3 Federated Identity Management

Cornell successfully configured OpenStack Horizon web console to use Globus Auth to authenticate users and documented the procedures. The infrastructure teams from each site worked together to define the identity management architecture for the federation and the user information format that will be published by the Aristotle portal to federation sites.

Cornell began working on the scripts that create the mapping files used by OpenStack Keystone from the user information received from the portal.

2.4 Cloud Status by Site

The chart below summarizes each site's production cloud status. All three sites are in the process of standing up OpenStack environments and all of the hardware will be transitioned from Eucalyptus to OpenStack. UB has a production OpenStack cloud; Cornell and UCSB are still Eucalyptus.

#1541215

	Cornell	Buffalo	Santa Barbara
Cloud URL	https://euca44.cac.cornell.edu	https://lakeeffect.ccr.buffalo.edu/	https://console.aristotle.ucsb.edu/
Status	Production	Production	Production
Software Stack	Eucalyptus 4.4	OpenStack	Eucalyptus 4.2.2
Hardware Vendor Year 1	Dell	Dell	Dell
Hardware Vendor Year 2	Dell	Dell, Ace	Dell, HPE
DIBBs Purchased Cores	*168	**256	356
RAM/Core	4GB /6GB	up to 8GB	9GB Dell, 10GB HPE
Storage	Ceph (1152TB)	Ceph (384TB)	Ceph (528TB)
10gb Interconnect	Yes	Yes	Yes
Largest instance type	28core/192GB RAM	24core/192GB RAM	48core/119GB RAM
Globus File Transfer	Yes	Planned	Planned
Globus OAuth 2.0	Yes	Yes	Planned
	* 168 additional cores augmenting the existing Red Cloud (488 total cores).	** 256 additional cores augmenting the existing Lake Effect Cloud (424 total cores).	***356 cores in UCSB Aristotle cloud (572 total cores, Aristotle is separate from UCSB campus cloud)

3.0 Cloud Federation Portal Report

Content updates to the project are ongoing (https://federatedcloud.org). Updates were made to many portal branches this quarter, including publications, partners, dashboard, and news and events.

We continue to monitor the Aristotle usage graph (https://federatedcloud.org/using/federationstatus.php) to ensure data is being collected consistently from all sites. We continue to implement software to verify that the data ingestion API is running. Nagios server at Cornell is now monitoring the usage report API at all 3 sites in the Aristotle federation. When the usage report API at a site becomes unreachable by the Aristotle portal, Nagios will alert the infrastructure team at Cornell to take appropriate corrective action.

The checks being performed have changed. The new API from UB gives a percentage for the graph; there is no longer a “Free” or “Capacity.” The format of the information being collected from OpenStack has changed and will require database modifications to accommodate the changes to provide core and storage usage.

The portal planning table was not updated this quarter.

#1541215

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.	Transition to OpenStack console with Globus Auth login.

#1541215

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 –5/2018**	**6/2017 – 10/2018**	**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. Note: due to OpenStack move, account creation across sites is delayed.	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.

3.1 Software Requirements & Portal Platform

No software changes were made to the portal platform this quarter.

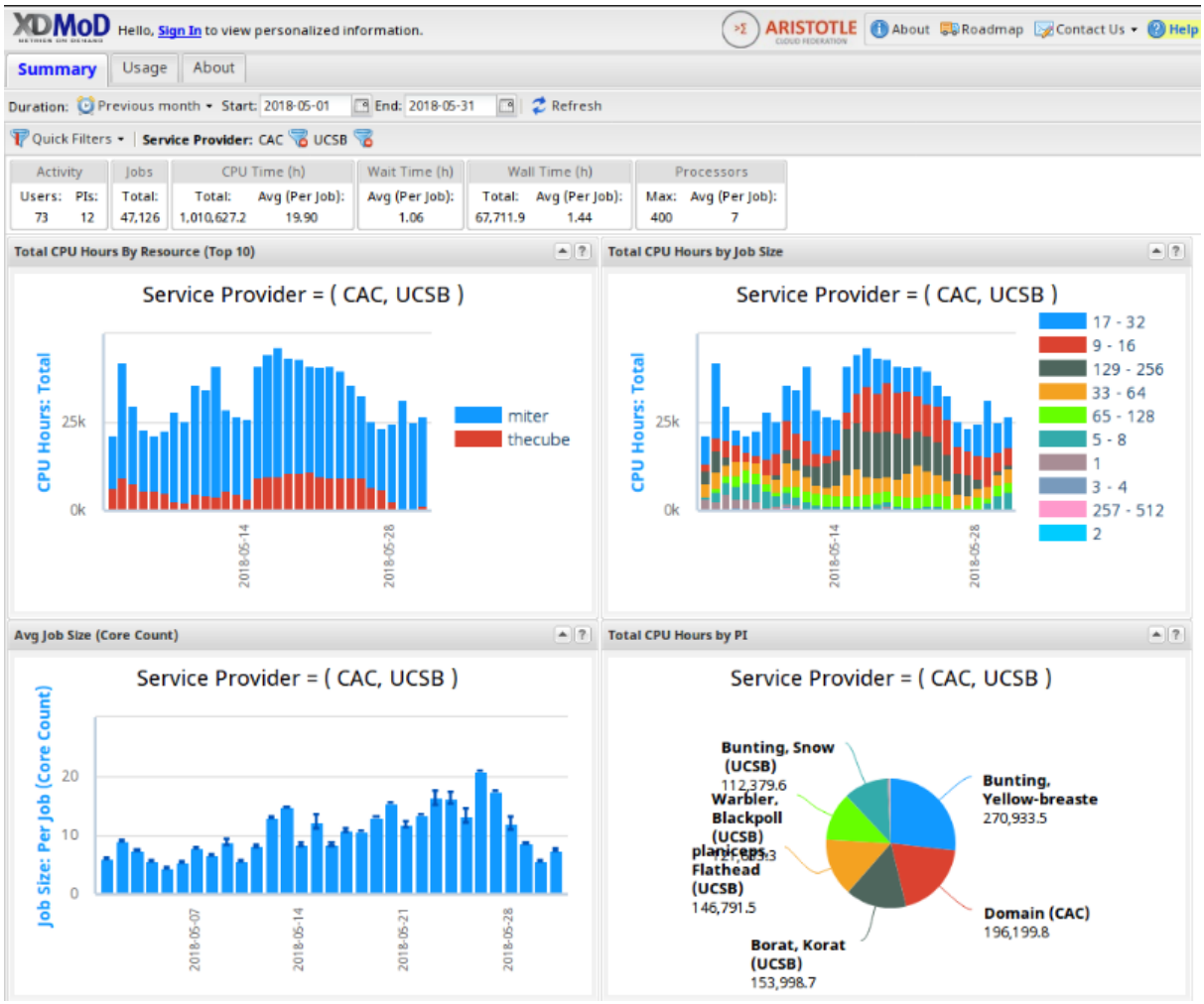
3.2 Integrating Open XDMoD and DrAFTS into the Portal

DrAFTS continues to function as a pricing tool for AWS but without the reliability guarantees. UCSB personnel met with AWS representatives to discuss the changes AWS made with respect to spot instance pricing and reliability. The AWS team claimed that there was no real change but Aristotle data seemed to indicate otherwise. After an extensive analysis, the team determined both that Amazon has raised the prices of spot instances and also reduced their reliability. Neither of these changes is documented nor were they part of the AWS announcement of "enhancements" to spot instance pricing. Thus DrAFTS has proved a useful tool for auditing and understanding the cost and reliability of cloud resources (even when the vendor fails to adequately announce service or pricing changes). We are exploring this new use of DrAFTS in the coming months as a way of determining reliability and price changes for volatile public cloud resources.

The UB team continues to test, verify, and refine cloud metrics. In addition to Eucalyptus, XDMoD 8.0 will support the ingestion of OpenStack log data for the generation of cloud metrics. Testing is underway for the 8.0 release targeted for the end of September 2018. Following this release, the UB team will begin work on including OpenStack data in federated Open XDMoD.

NSF
#1541215

This is a screenshot of the federated Open XDMoD page:



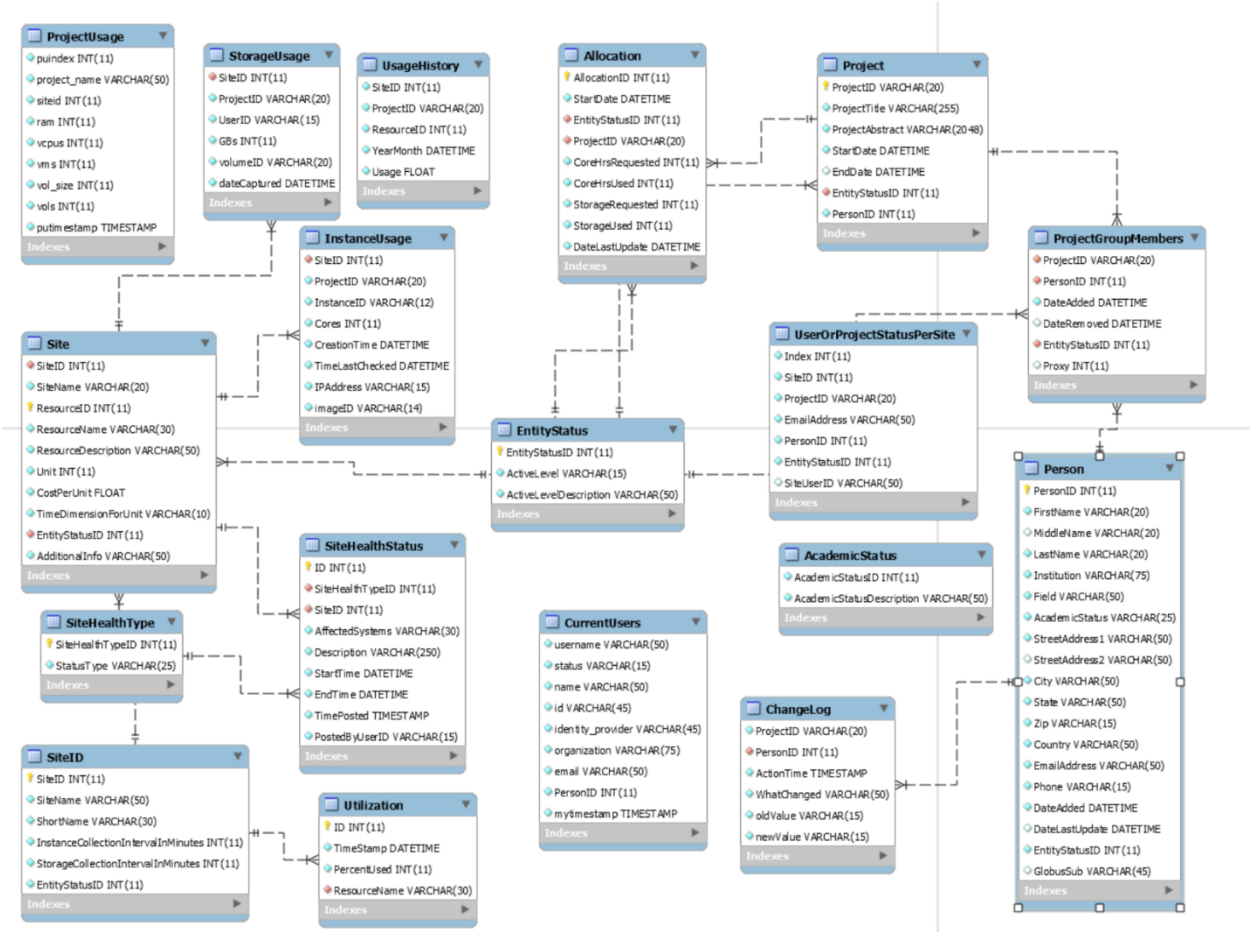
The Open XDMoD timeline is available online:
https://docs.google.com/spreadsheets/d/1KlBIWY8ntCC35_5v7o19rro_oOM0Cre8WER-pIlSxMI/edit?usp=sharing

3.3 Allocations & Accounting

- Created a new table called CurrentUsers.
- Created Python and php scripts to get Globus Authentication Information from a Globus sub to get a list of active users for OpenStack
- Called Python script from php script to get information from Globus into DB
- Wrote stored procedure to get project membership
- Wrote php code to call stored procedure to get to JSON for array of users on OpenStack
- Wrote stored procedure GetProjectListForUserRequest for the portal dashboard.

#1541215

This is the database schema:



4.0 Research Team Support

4.1 Science Use Case Team Updates

Science support research this quarter has focused on containerization, including building a pipeline infrastructure for radio astronomy (use case 4) and for MPI-based research (use case 5).

NSF-funded REU students had positive experiences at Cornell and UCSB and contributed to use case progress.

Use Case 1: A Cloud-Based Framework for Visualization & Analysis of Big Geospatial Data
Varun Chandola's UB team had a paper accepted in the Big Earth Data journal titled "Machine Learning for Energy-Water Nexus Challenges and Opportunities." The researchers used Aristotle cloud resources for comparative evaluations of various machine learning methods to better understand the Energy-Water nexus. The paper acknowledges the Aristotle grant.

#1541215

Additionally, a paper on the webGlobe tool (a browser-based, cloud-driven 3D user interface that allows scientists to upload, visualize, and analyze NetCDF data sets) will be presented at the BigSpatial 2018 workshop to be held in conjunction with the ACM SigSpatial conference in November in Seattle. A version of webGlobe was shipped to Chadola's collaborators at Oak Ridge National Laboratory to support their research activities in the area of climate data analytics.

Use Case 2: Global Market Efficiency Impact

Dominik Roesch is working with a new UB student who will help automate part of the finance data framework being hosted on Aristotle. This month Roesch will present a paper titled "Asset Pricing: A Tale of Night and Day" (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3117663) at two conferences: the 30th Annual Meeting of the Northern Finance Association and the 3rd Research in Behavioral Finance Conference. He will also submit a paper titled "The Impact of Arbitrage on Market Liquidity" (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2295437) to a journal in the coming weeks. The Aristotle project is acknowledged in all papers and presentation slides.

Use Case 3: High Fidelity Modeling and Analytics for Improved Understanding of Climate

Progress continues on the application of the Weather Research and Forecasting (WRF) model for climate-relevant simulations on the cloud project led by professor Sara C. Pryor and postdoc associate Tristan Shepherd.

Précis objectives of our current suite of simulations:

1. Quantify impact of resolution (to convective permitting scales) on near-surface flow (i.e., wind speed) regime fidelity
2. Examine scales of coherence in wind fields. Specifically, spatial scales of calms (i.e., wind speeds < 4 m/s), and spatial scales of intense wind speeds (i.e., wind speeds > the local 90th percentile value)
3. Quantify the platform dependence of wind simulations (i.e., quantify the differences in near-surface wind regimes from simulations conducted on conventional HPC and the cloud)
4. Examine inter-annual variability in near-surface wind speeds (can we simulate it, what is the source?)
5. Evaluate impact of large wind turbine (WT) developments on downstream climate (local to mesoscale)

We are addressing these objectives by conducting and analyzing the output from high-resolution numerical simulations with the Weather and Research Forecasting model (WRF, v3.8.1).

Focus of this quarter's research:

- Completing the analysis of output from our long-term simulation with the WRF model at 12km over eastern North America for 2001-2016 for the assessment of year-to-year variability in the wind resource. The scientific importance of this work is as follows: Inter-annual variability (IAV) of annual energy production (AEP) from wind turbines due to IAV of wind speeds from proposed wind farms plays a key role in dictating project financing but is only poorly constrained. This study provides improved quantification of IAV over eastern N. America using purpose-performed long-term numerical simulations. It may be appropriate to reduce the IAV applied to pre-construction AEP estimates, which would decrease the cost of capital for wind farm developments. If adopted by the wind energy financing industry, our recommendation could contribute to the long-term trend of decreasing levelized cost of energy from wind. The paper describing this research is now in press at the journal *Wind Energy Science*.
- Continuing to address the impact of (a) changes to the computational node on which our simulations with the WRF model are running and (b) the compiler used in terms of the net effect on the

#1541215

simulated wind climate. These matters are ongoing, since we still do not have a compilation of the WRF model using the Intel compiler, but initial results regarding simulation sensitivity to computing platform are published in a forthcoming book chapter.

- Continuing simulations to address the sensitivity of climate and wind farm wake simulations to the precise method used to describe the aerodynamics of wind turbines and thus their disturbance of the flow field. Progress of the simulations on the Aristotle node has been slow. The source is uncertain but the number of 'real days' per day of simulation appears to be significantly slower (by a factor of two) than in our previous runs. The Cornell Aristotle use case support team have found no problems with the environment and thus this slow-down remains something of a mystery.

Activities planned for next quarter:
It is our intention that work over the coming 3 months will focus on the latter two matters listed above in addition to continuing to perform statistical analyses focused on model validation and verification efforts relative to in situ measurements of wind speed. This work is contingent on getting Wrangler mounted to one of our instances on Jetstream, and is a matter we are actively working with the Cornell support team on.

Journal manuscript:
Pryor, S.C., Shepherd, T.J. & Barthelmie, R.J. (2018). Inter-annual variability of wind climates and wind turbine annual energy production. Wind Energy Science Discussions. Accepted for publication in Wind Energy Science. https://www.wind-energ-sci-discuss.net/wes-2018-48/wes-2018-48.pdf

Book chapter:
Pryor, S.C. & Hahmann, A.N. (2018). Downscaling wind: Forthcoming in Oxford Research Encyclopedia, Climate Science. Oxford University Press. Ed. Von Storch, H. (10,000 words/10 figures). In press.

Conference presentations submitted:
Shepherd, T.J., Volker, P., Barthelmie, R.J., Hahmann, A. & Pryor, S.C. (2019). Sensitivity of wind turbine array downstream effects to the parameterization used in WRF. 99th American Meteorological Society Annual Meeting (10th Conference on Weather, Climate, and the New Energy Economy). Abstract submitted.

Shepherd, T.J., Brazier, A., Wineholt, B. Barthelmie, R.J. & Pryor, S.C. (2019). Quantifying weather and climate simulation reproducibility in the cloud. 99th American Meteorological Society Annual Meeting (5th Symposium on High Performance Computing for Weather, Water, and Climate). Abstract submitted.

Use Case 4: Transient Detection in Radio Astronomy Search Data
Working with REU student Plato Deilyannis, former REU student Elizabeth Holznecht and Cornell Aristotle use case consultant Peter Vaillancourt, we have built a new flexible framework for running searches in radio astronomy data; data are read from their native format into NumPy Arrays, and the pipeline's routines are selected from a configuration file and include a friend-of-friends search implemented and tested by Deilyannis, and also allow the running of the pipeline developed by Laura Spitler who discovered FRB 121102. Holznecht is implementing a flood fill algorithm. Senior research associate Shami Chatterjee credited the Aristotle grant at the International Astronomical Union (IAU) General Assembly in Vienna in August, and in his talks at the IAU Symposium 344 ("The dwarf galaxy host of a fast radio burst"- https://astronomy2018.univie.ac.at/abstractsiaus344/#iaus344abstr16) and at the Division B meeting ("New results in radio astronomy: Fast radio bursts and transients" - https://astronomy2018.univie.ac.at/abstractsDivB/#DivBabstr2).

#1541215

The pipeline code is available on GitHub at: https://github.com/federatedcloud/FRB_pipeline

Use Case 5: Water Resource Management Using OpenMORDM
Recently we have containerized and published build scripts for the Lake Problem code from the Patrick Reed Research Group and successfully run it in MPI across multiple cloud virtual machines using Docker on Jetstream. This will allow distributed scientific software to be executed faster at cloud scale, both in existing institutional clouds, XSEDE resources, and public providers.

We also have initiated work on making the containers portable across base OS images (e.g., Ubuntu and Alpine Linux) and cloud providers, and have used the Nix package manager to achieve this portability as well as reproducibility in software builds. This portability also lends itself well to running the same software stack in a Singularity container, which we have already tested as a multi-container, single-node solution. Existing software is available at the Aristotle federated cloud GitHub repository for the Lake Problem (https://github.com/federatedcloud/Lake_Problem_DPS) and for base OpenMPI container images (https://github.com/FederatedCLoud/NixTemplates). Base Singularity image have been published to Singularity Hub (https://www.singularity-hub.org/collections/1220).

We are currently using host-network interfaces in Docker to minimize network overhead, and as a next step we will benchmark both Docker and Singularity to determine their performance characteristics. Future plans include testing reliability, effectiveness, and limits of scaling as well as automatic tasks for infrastructure provisioning and task execution.

Use Case 6: Mapping Transcriptome Data to Metabolic Models of Gut Microbiota
This quarter our project focused on the impact of gut microbial community composition and host gut environment on the metabolite release and exchange profiles. Our purpose is to use the insights from this *in silico* analysis to define precise hypotheses on metabolic interactions among gut microbes that can be tested by subsequent empirical analysis.

In this quarter, we:

- Constructed multi-species metabolic networks *in silico* of 2, 3, 4 and 5 bacterial species.
- Quantified predicted metabolite fluxes within and between species under multiple conditions with varied composition and bounds of the exchange reactions.
- Wrote a MATLAB function to restore exchange constraints to a multi-species model created from a single species.
- Used this function as part of a unit test to confirm that the creation of multi-species models does not influence the objective result when run with standard constraint-based analysis algorithms such as FBA, enabling us to confidently tailor exchange flux constraints for true multi-species models for use with SteadyCom.

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

Pond reclamation project for cattle grazing:
This project kicked off in Q3 at the Sedgwick Reserve. The project secured the power infrastructure (large-scale solar panels and batteries) necessary to deploy an array of cameras and sensors around one of the main ponds that will be used by the cattle. In addition, the ecological team determined that the cattle could be

#1541215

used, in part, to clear the reclaimed pond areas of native grasses and vegetation so they could be prevented from sinking into the mud. Thus, the project will deploy an array of moisture sensors and will use image processing to alert the land managers when cattle stray into areas that may be too "soggy" to support their weight. This effort represents one of the first to use cattle themselves to improve the land and water resources that they will themselves be using. The power infrastructure is being installed and tested for this use case.

California citrus frost prevention project (Lindcove Research and Extension Center):
Data acquisition for the Phase 2 deployment for the citrus orchard at Visalia, California began in Q3 but dust and ambient radio conditions have impeded the efficiency of the solar infrastructure. In particular, the science teams are seeing drop out from midnight until approximate noon in the Phase 2 deployment. Phase 1, however, continues to function correctly. The current hypothesis (which the team is investigating) is that the solar footprint has been calibrated too "tightly" for the local conditions (this calibration took place under controlled conditions). A refit for the Phase 2 deployment is planned to correct the problem. However, analytics research is now taking place that examines whether the data from the Phase 1 deployment can be used to alleviate the drop out in Phase 2. In addition, the local computer networking infrastructure (which is not designed to support instrumentation) is constraining further progress. The Aristotle science use case team is working with the local IT managers to develop a network configuration plan that can support the needs of the project.

Solar panels and sensors will help researchers know when to turn on windfans to protect citrus trees from frost.

Differential irrigation project:
Phase 2 of the differential irrigation project was deployed in Q3 at Fresno State almond orchard and immediately failed. Part of the issue is with sensor density. In the Phase 1 deployment, each low-power controller was capable of reading 3 moisture sensors and 3 temperature sensors. This density is too low to support deployment at scale. The Phase 2 deployment uses a new set of controller boards that are 5 times denser, but these boards appear to have higher power requirements than their designers first expected. Also, the quality assurance practiced by the manufacturers of these new boards appears inferior as 2 of the first 10 failed outright when deployed.

Both problems are being addressed by the science team and the local orchard managers. New boards, larger solar panels, and larger batteries are being tested for a full scale retrofit (planned for September). In addition, the farm managers have requested that the science team develop a warning system so that if the solar panels become damaged (due to farm operations or other problems) they will be automatically notified. The farm managers feel comfortable taking on maintenance responsibility for the solar infrastructure if they can be informed of problems electronically. The science team is now developing this alerting infrastructure.

NSF
#1541215

The differential irrigation project goal is to instrument trees within the test orchard over a full growing season to determine how much water can be saved by irrigating the different sides of the root stock in proportion to its dryness.



UCSB Aristotle researchers and grad students and faculty from Cal Poly San Luis Obispo and Fresno State work with farm consultants to instrument an almond test tree for differential irrigation. The goal is to see how much water can be saved by irrigating different sides (sunny and shade) of the root stock in proportion to its dryness

5.0 Community Outreach and Education

5.1 Community Outreach

- Aristotle co-PI Tom Furlani was a speaker at and a proceedings reviewer for The National Academies of Sciences, Engineering, and Medicine "Opportunities from the Integration of Simulation Science and Data Science" workshop. He participated on the "Service, Usage Models, and Economics" panel with Roger Barba (AWS) and Vani Mandava (Microsoft). His presentation was titled "Campus-Based Systems and the National Cyberinfrastructure Ecosystem." The workshop proceedings are now available at: https://www.nap.edu/catalog/25199/opportunities-from-the-integration-of-simulation-science-and-data-science
- PI David Lifka and Furlani represented the Aristotle project at the NSF-sponsored DIBBs18 Workshop in July with a paper and poster: https://federatedcloud.org/papers/1541215%20Aristotle%20Cloud%20Federation%20DIBBs18%20poster.pdf
- Co-PI Rich Wolski was on the Organizing Committee of the 3rd IEEE Cyberscience and Technology Congress.
- Two demos were developed for PEARC18 and presented informally in July to interested individuals: WRF in a container and CENTAURUS cloud service for k-means clustering.
- Rich Knepper was an author on a paper titled "Security best practices for academic cloud service providers" that was presented at the 2018 NSF Cybersecurity Summit for Large Facilities and Cyberinfrastructure: https://scholarworks.iu.edu/dspace/handle/2022/22123
- Lifka travelled to Dartmouth to meet with their CIO and Research Computing management and staff to discuss interest in the federation.
- News articles:
    - "UCSB SmartFarm uses cloud computing to help farmers increase sustainability" (*Santa Maria Sun*) - http://www.santamariasun.com/school-scene/17745/ucsb-smartfarm-uses-cloud-computing-to-help-farmers-increase-sustainability/
    - "Wireless smart farming to keep frost away from citrus" (*RCRWireless News*) - https://www.rcrwireless.com/20180717/internet-of-things/wireless-smart-farming-to-keep-frost-away-from-citrus-tag41

- "AgTech: The Success of Farming is in the Cloud" (*Challenge Advisory*) - https://www.challenge.org/knowledgeitems/agtech-the-success-of-farming-is-in-the-cloud/
- Cornell professors contribute to winning offshore wind energy alliance (*Cornell Chronicle*) - http://news.cornell.edu/stories/2018/07/cornell-professors-contribute-winning-offshore-wind-energy-alliance

5.2 Education

- During summer 2018, the Aristotle team provided five REU students training and the opportunity to participate in science use case projects and emerging cloud computing technologies.
  - Cornell REU student Plato Deilyannis worked with professor Jim Cordes, senior research associate Shami Chatterjee, graduate student Shen Wang, CAC consultants Adam Brazier and Peter Vaillancourt, and former Aristotle REU student (summer 2017) Elizabeth Holzknecht. Plato worked with Peter on implementing the flexible pipeline architecture and took responsibility for coding a friend-of-friends algorithm to examine the dynamic spectrum of high time-resolution radio observations, highlighting the best candidates from the observation. All of this work was done in Python and is available on GitHub. This code was run on several test data sets, including the discovery plot for the 121101 repeating FRB discovered by Laura Spitler.
  - Cornell REU student Cindy Wu was looking for trends and patterns in the growth of multiple species microbial communities in relation to the number of microbes in the community, working with five computational models of microbes that could be found in the gut microbiota of fruit flies. Coding was primarily done using COBRA toolbox, a linear programming toolbox that was added to MATLAB that contained scripts that could perform linear optimization on the fluxes of the computation models. All flux analyses were done to optimize the biomass equation in the models. Cindy wrote a script to construct multiple species models of any size from given individual models and after constructing the multiple species models, wrote code to analyze the exchange reactions of each individual model and all the possible combinations of different sized community models from the five given models.
  - Cornell REU student Peter Cook conducted several analyses of the operating conditions for wind turbines (WT) using 10-minute output from long-term numerical simulations using the Weather Research and Forecasting (WRF) model (Mar. 2001 – Dec. 2016) produced by the atmospheric sciences use case. The first of these analyses involved evaluation of the WRF output near WT hub-height (HH) relative to actual energy production data from the Energy Information Agency. Peter used EIA 860 and 923 data sets, by loading and manipulating Excel files in MATLAB, filtering out missing data and fixing incorrectly reported data prior to their use in the model evaluation. He then used the simulation output to examine the co-variability of wind speeds and wind energy power production at different temporal lags and aggregation. In order to assess the degree of coherence of wind speed variability at WT locations, Peter wrote several scripts to analyze WRF WS data using an array of statistical methods and learned how to employ parallel computing for efficiency. He also optimized non-parallel code by converting double-precision data to Boolean arrays and opting to perform vectorized and logical (or, and) operations over more complex mathematical (+,-,<,>) and elementwise ones. Peter visualized these results in a large number of figures.

#1541215

  - UCSB REU student William Berman completed his work with DrAFTS by designing a new interface for comparing spot instance prices. He graduated, completing his undergraduate degree and now works in industry as a professional developer.
  - UCSB REU student Gareth George joined the Aristotle team and, to date, has participated in three different projects. Initially, he worked with William Berman on the revamp of the DrAFTS data management infrastructure. When that work was completed and William graduated, Gareth transitioned to a project analyzing Amazon's new pricing mechanism for spot instances. His work showed that Amazon raised the prices and lowered the quality of spot instances without announcing these changes. This work has resulted in a paper that will be submitted to IEEE International Conference on Cloud Computing (IC2E) in October. Gareth then began focusing on supporting the science teams using Aristotle for IoT analytics. His current project is developing a portable version of AWS Lambda so that the science team applications which use Lambda can be executed in remote locations.
- Cornell is building a website to teach a focused workshop for Upward Bound students who will be visiting campus this fall. The workshop will teach the high school students introductory programming concepts using graphical interfaces, the Python language, and Aristotle Cloud Federation hosting.
- Aristotle use case scientist Chandra Krintz will be a keynote speaker at the 8th International Conference on the Internet of Things (IoT 2018) in October. The title of her presentation is "SmartFarm: IoT systems that simplify and automate agriculture analytics."

#1541215