ARISTOTLE
CLOUD FEDERATION

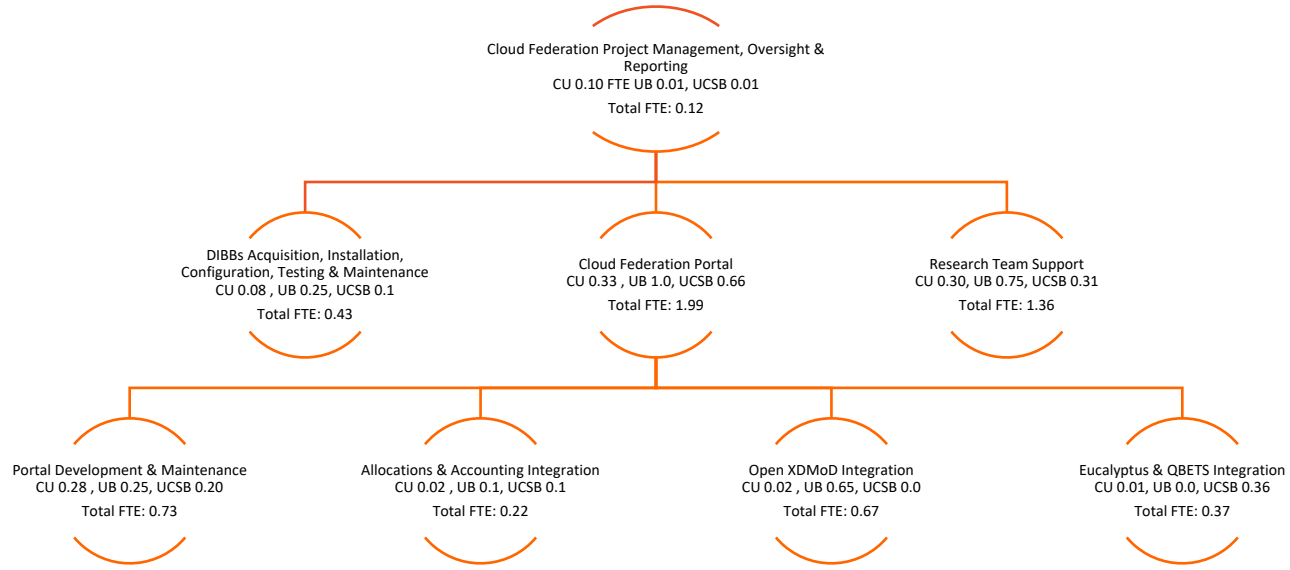
CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation

Program Year 3: Quarterly Report 3

6/26/2018

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

This is the Program Year 3: Quarterly Report 3 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 with the major public cloud providers.

1.5 Project Status Calls

4/10/2018 status call:

- Ten UCSB student teams in Wolski's CS293B Cloud Computing class are using a Jetstream allocation and Aristotle to solve large problems. These workloads (lots of small VMs) may help Jetstream better understand cloud workloads vs. HPC workloads.
- Aristotle has triggered a lot of demand for Aristotle-like resources at UCSB. Another college started its own OpenStack private cloud that mirrors Aristotle. UCSB administration is getting emails from faculty asking for cloud (administration does not have a clear understanding cloud or its costs). Aristotle has woken people up to the possibility of cloud. Faculty are productizing homework and assignments in Docker containers. Each student gets their own Docker sandbox at tremendous scale. This is moving the needle at UCSB. The question being asked is: what will it cost to make these services available campuswide?
- AWS has decided to hide their spot market algorithm which makes it significantly less useful to do scientific computing in the spot market. This decision by AWS is prohibiting us from using DrAFTS to make 99% reliable predictions of how long spot resources will be available before they are terminated. We will have a call with AWS to discuss this development.
- UB put together beta 8.0 RPMs for XDMoD which include support for federating HPC accounting data. All 3 sites will install this version so we can get started federating some HPC data, before federating cloud data and getting OpenStack working.
- All sites are working on their OpenStack installs. UB is lead installation and is sharing lessons learned with other federation sites.
- Wineholt is working with Patrick Reed's student (Cornell water resource management use case) to do multi-instance MPI on Jetstream. It will run between containers that are on different VMs; Docker swarm will eventually automate the process.
- Cornell use case scientist Sara Pryor is interested in whether, and to what degree, WRF outputs are impacted by hardware in the cloud. Wolski commented that they change the way containers do networking more often than you change your socks. He believes a researcher may get different readings when they pull the latest Docker and that may affect results. While it's the same container image, how it runs depends on what version of Docker is running.

#1541215

- UCSB is working on a new citrus orchard frost prevention project: the focus at this stage is real-time data acquisition and data cleansing on Aristotle; analytics are under development. The soil moisture monitoring of almond trees project continues, but users are struggling to understand just what the data is telling them. UCSB is analyzing whether this is a science or an infrastructure problem.
- UCSB is also developing an industry standard approach to analyzing irrigation management soil data that may have a distinct advantage over current Management Zone Analyst (MZA) software. Several papers are being developed to find out what the community thinks of this new approach.

4/24/2018 status call:

- UB installed Red Hat OpenStack and is in friendly user testing mode with cloud savvy users. They reconfigured Ceph and rebuilt everything with SSD drives. This solved UB Ceph performance issues and they're happy about that. They're now looking at some kind of internal communications problem involving container bridges in multiple AZs.
- Cornell continues work on its OpenStack installation. UB's sharing Puppet files with Cornell.
- Wolski taught a class at the UCSB College of Engineering on how to use OpenStack and how to use network topology to automate subnets so everyone gets basic network topology. At UB, users can create their own private network if they want to, but they don't have to. They use 2 provider networks: one public and one private. Users can pick either network, or use both. There isn't isolation unless a user creates a VPN.
- Cornell's Walle rewrote the stored procedure for accounting so it's more in line with the accounting system that Cornell's currently using. Core usage runs a lot faster, storage does not. Walle also created a new table and history in the database which may improve speeds. In addition, the Aristotle portal is now running with 2 processors instead of 1, which is faster.

5/8/2018 status call:

- Cloud usage and motivations to move to the cloud were discussed. Wolski said there's currently big budget pressure at UCSB (due in part to increases in overhead) and their CIO is interested in moving to public cloud. Furlani said lots of XSEDE jobs are running on 1 node or 1 core; ideally XSEDE resources should not be tied up for those types of jobs, they should be moved to the cloud. Lifka said the Dept. of Statistical Science at Cornell has 500 students each of which needs their own copy of R. It would be wonderful to provision that in the cloud. All the students would need are Chromebooks. 500 cores would not be needed because the students use the cores interactively. Another clear advantage of cloud is greater access to software and tools that may not be available on premise (e.g., each site in the Aristotle federation has some unique capabilities: UB has Thomson Reuters Tick History (TRTH) market data for financial engineering researchers, UCSB has machine learning tools, Cornell has MATLAB Distributed Computing Server, etc.).
- To take advantage of cloud capabilities and further motivate researchers to move to the cloud, experts (such as the Aristotle use case consulting staff) will be needed to show researchers how to onboard applications.
- There was discussion on how someone at Cornell would be allowed to run on UCSB's cloud: is that part of their allocation or is everyone allowed to run anywhere? It was decided that you join, you get a single sign-on that gives you access to resources across the federation, each site maintains local control, and you go use those resources that make the most sense to you.

#1541215

- Work is proceeding on a supplemental proposal that would assess how, and to what degree, public cloud provider platforms could be integrated with the Aristotle portal and contribute to the development of an open cloud marketplace.
- Building a Platform as a Service with Kubernetes (automated container deployment and scaling) might be valuable.
- OpenStack federated logins are working at UB. A separate authentication discussion is needed in order to decide plumbing to translate from Globus login to OpenStack accounting.
- Use case updates: Aristotle GIS use case team submitted papers to ACM GIS; the radio astronomy use case is making good progress building a much more flexible pipeline; Pat Reed's team is testing multi-instance MPI on Jetstream to see how the Python version scales on an OpenStack cloud; Sara Pryor is planning a paper on when a WRF simulation VM is moved from Aristotle to Jetstream to Azure, will researchers get exactly the same results; and, UCSB is acquiring TBs of data from a citrus packline machine because the California Citrus Board wants to know if they can use that data to analyze fruit quality, detect diseases, see drought effects, etc. UCSB is also preparing two papers on Aristotle-driven data science for soil EC (electrical conductivity measurements that measure soil properties that affect crops, i.e., soil texture, drainage conditions, etc.). They had an IoT software breakthrough that looks like an operating system for IoT and is very fast.

5/22/2018 status call:

- UB, UCSB, and Cornell infrastructure teams had a productive call last week on two topics: federation authentication, and OpenStack and hyperthreading. Development of Open XDMoD site info is progressing nicely. XDMoD was installed and data was ingested in all sites. We are starting to send data up to the hub at UB. It looks like the virtual host is not quite ready to go, but we are making great progress there. The UB team is ahead of UCSB and Cornell on OpenStack and hyperthreading. They thought about using virtual CPUs with OpenStack to tell node type and actually place instances on the nodes. The idea of having some nodes with hyperthreading on might be advantageous and exciting to researchers. We talked about federated authentication and, after a lengthy discussion, agreed that the portal is where the master Aristotle users and user associated sub will be kept. Each site will pull that information to create accounts on their local cloud sites. UB will diagram and share what the flow will look using Globus as the authentication process.
- UCSB would like to have an automated OpenStack deployment approach so it's easy to add new hardware and train people to use it. UCSB needs network isolation. The question is: how do they provide a default networking deployment for uneducated users for OpenStack? Their students have to know how to provision a private network on a public network; non-scientific users will need to know this as well. UCSB decided against using Ensemble after Engineering found Ensemble is not usable (even by their CS community) and Arts and Letters couldn't get Ensemble to work with Ceph. They think a product Red Hat trained them on, a container-based approach that sort of bootstraps OpenStack from bare metal will work and will automate the process so they can train other administrators to use it. UB and Cornell use Puppet and Foreman; UCSB may try that approach.
- UCSB is troubleshooting the running of Centaurus (a big parallel clustering service for soil EC data) on Jetstream and Aristotle. Centaurus purposely hides all complexity from the farmers using it. We discussed demoing Centaurus and/or WRF running across federation sites at PEARC18.

#1541215

6/5/2018 status call:

- Aristotle REU students are getting up to speed on their use cases and learning Docker.
- There was a blue sky discussion on gutting Pat Reed's MPI code and spinning elastic containers up and down to get scalability. This approach would be less costly than MPI which is always running.
- Discussions also occurred on how the federation exchange rate would work. It was decided to start simple with core hours and storage GBs, then come up with a formula, an equation of core, core speed, memory per core, or whatever features we decide on. Use data from XDMoD on memory, core, etc. with, perhaps, performance metrics that we agree on through some XDMoD kernel. Get the plumbing working right first, then augment it with multiple formulas. Look at a suite of kernels, then a ratio of compute power per platform. Try to correlate them in order to get efficient use of the resources.
- We're making progress on federated XDMoD: instances are up and running at all 3 sites and UB is bringing data in and validating it.
- Cornell is installing OpenStack. They're hooking up the compute nodes to the provider network so people can just request an external IP address. They also have a self-service network so users can define their own private network and they're making sure that works. Unlike Eucalyptus, OpenStack keeps an IP address reserved for an instance, even when it's stopped. Cornell could theoretically run out of addresses. If Cornell just goes with a self-service network, that's a lot of overhead of users; so, they'll use both, a self-service network and a provider network. UB is also running both. UCSB is just running a provider network, but they think they'll have to go to a self-service network and provide useful glue for less capable users.
- UCSB had a call with AppScale regarding their cloud pricing tool. AppScale said that the tool is more at a demo than product stage. We agreed to keep in touch with AppScale in case they decide to open source the pricing tool, in which case we might be able to work together with them on moving that forward.

6/19/2018 status call:

- UCSB and Cornell had a 6/18/2018 conference call with five AWS engineers and education representatives to explain that scientists need to have an adequate level of predictability in the spot market in order to get their time-sensitive papers, proposals, and research done with some level of guarantee. Unfortunately, their spot pricing algorithm is now hidden and AWS said that they will not expose the internals so that we can make predictions on how long jobs will run before they are terminated. Our prediction tool (DrAFTS) was providing predictions on 1-3 hour jobs typically used by scientists with 99% reliability. The AWS position seems to be that spot instances are not supposed to be predictable, so scientists should use their more expensive (~10x) on-demand instances. We are looking into what we can do with the new pricing model but, at present, it cannot support workloads that are interruption intolerant.
- A meeting of all Aristotle REU students is planned for next week.
- Cornell is looking at building Singularity images. Singularity will add an extra dimension to our bursting capability because it can run in our federation, on NSF clouds, and public cloud.
- Cornell updated the accounting database: a table to store core and GB usage by month was added.
- UCSB's OpenStack installation is in pre-production. They want to make sure everyone in their group knows how to rebuild the OpenStack cloud and they would like to change the networking. They're currently trying to migrate some instances from Eucalyptus to OpenStack. Globus identity

#1541215

provider is working with the OpenStack dashboard. If the portal could run like an LDAP server, that would be cool.

- Differences between Docker and Singularity (which is growing in popularity) were discussed. Docker runs as root; Stampede2 won't let Docker in a shared space machine. On a shared system, Singularity can run as SLURM; it can have 10 users in the same box and they can't talk to each other.

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

2.1 Hardware Acquisition

There were no hardware acquisitions this quarter.

2.2 Installation, Configuration, and Testing

Cornell: Installed OpenStack test cloud with 2 nodes; verified LDAP integration for authentication; entered the testing phase for the Red Cloud OpenStack cloud; and, worked with UB on Open XDMoD federation integration.

UB: Deployed/developed/tested OpenStack; integrated OpenID Connect with Horizon and CLI; integrated Globus Auth with Horizon (we are working on CLI with Globus Auth); migrating EBS and Instance-Store instances from Euca to OpenStack; worked with Cornell/UCSB on Open XDMoD federation integration.

UCSB: Created sandboxes for testing OpenStack deployments; testing TripleO tools for OpenStack deployment; worked with UB on Open XDMoD federation integration; attended OpenStack Summit.

Federation Authentication: Steven Lee started off the discussion by asking "How tightly coupled will the cloud sites be?" Regions are a possibility (as suggested by Wolski), but then how to manage accounts outside Aristotle? All agreed that this seems complicated. Steven suggested we use OAuth2 for federation login and described the Globus login process and how a user gets a "sub" – the portal then captures the sub and user's name/email. UB has installed OpenID in front of LDAP which allows for gathering more user information. After further discussion about how this might work, neither Cornell nor UCSB want to maintain OpenID installations. All three sites agreed that the federation portal will maintain the master "Aristotle user and sub list" that each site will use to pull into their environment. UB will send a format of a file that they need to import. Steven will send out a diagram of the existing Globus AUTH process.

OpenStack with Hyperthreading: This topic came into play when UB was migrating to OpenStack and needed additional cores. OpenStack allows placement of instances (Eucalyptus never did) making Vcpu and NoVcpu instances more attractive and maybe good to use on old hardware pools.

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. Buffalo has a production OpenStack cloud; Cornell and UCSB are still Eucalyptus.

#1541215

	Cornell (CU)	Buffalo (UB)	Santa Barbara (UCSB)
Cloud URL	https://euca44.cac.cornell.edu	https://lakeeffect.ccr.buffalo.edu/	https://console.aristotle.ucsb.edu
Cloud Status	Production	Production	Production
Euca Version	Eucalyptus 4.4	OpenStack	Eucalyptus 4.2.2
Hardware Vendors	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	28 core/192GB RAM	24 core/192GB RAM	48 core/119GB RAM
Global File Transfer	Yes	Planned	Planned
Globus OAuth 2.0	Yes	Yes	Planned
	*168 cores added to Cornell Red Cloud (488 total cores)	**256 cores added to Buffalo Lake Effect Cloud (424 total cores)	***356 cores in UCSB Aristotle cloud (572 total cores, Aristotle is separate from UCSB campus cloud)

2.3 Potential Tools

- Supercloud - nothing new to report.
- Red Hat OpenStack – we are currently transitioning from Eucalyptus to Red Hat OpenStack.

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

#1541215

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 are:

- First result is the most recent record from the database. Ensure this record is not more than 1-hour old, otherwise the API is likely down.
- Check all records to ensure Free >= 0 (should always be 0 or positive).
- Check all records to ensure Capacity > 0 (should never be 0).

The portal planning table was updated 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.

#1541215

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.
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

Using the new functionality of MariaDB 10.2, a new table was created to store core and GB usage by month. Each month this is updated to reflect the usage of core and GB by project and site.

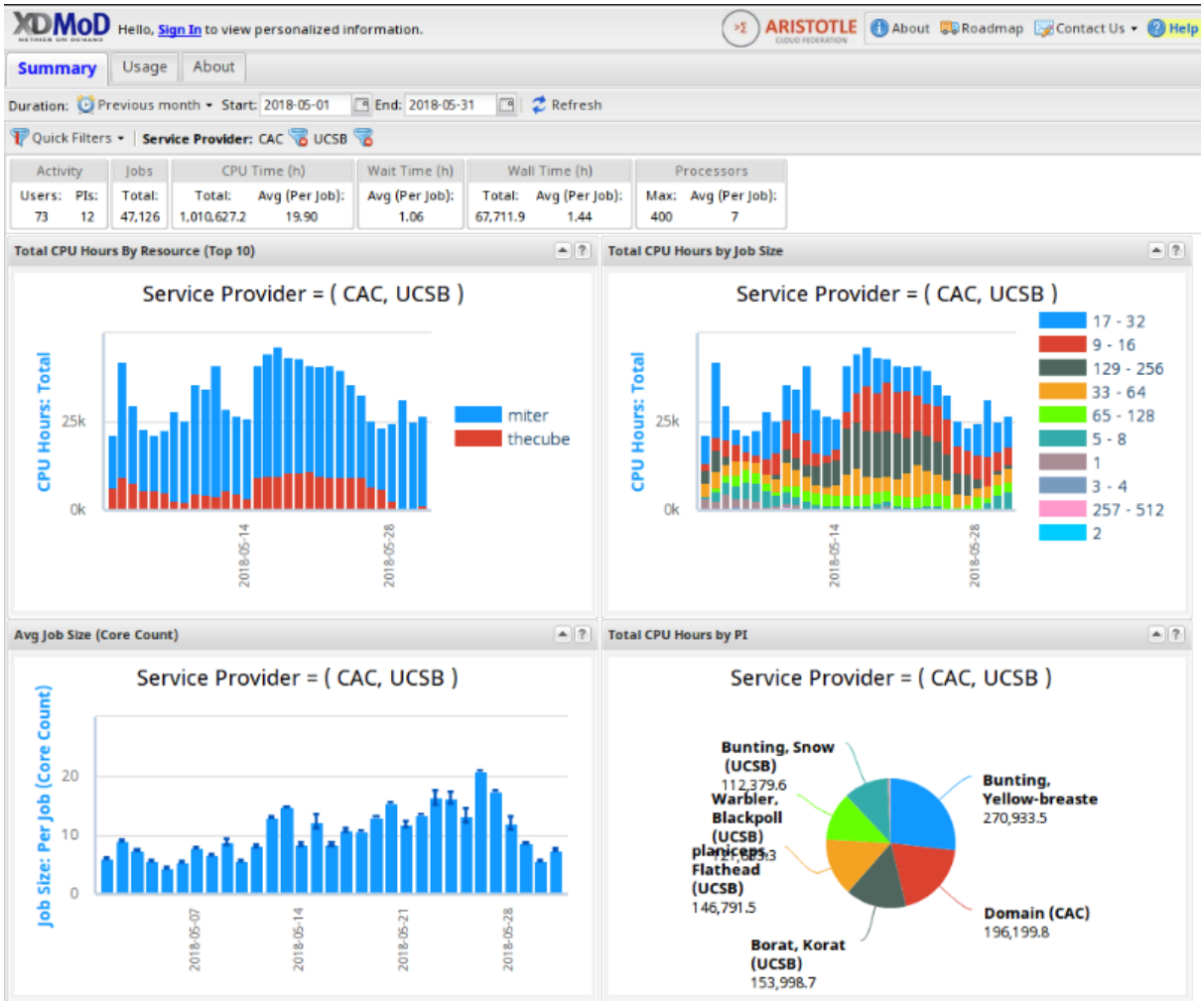
3.2 Integrating Open XDMoD and DrAFTS into the Portal

The DrAFTS team is testing their methodology against Amazon's new spot pricing policy which "smooths" price changes. In late January, Amazon introduced a new method for computing spot prices. While the exact algorithm remains hidden, the effect is to cause spot prices to remain stable over long periods of time. We are currently testing DrAFTS to determine the effect on DrAFTS bids. Under the old regime, DrAFTS could provide guarantees for up to 48 hours of durability. We suspect that under this new regime, the durability guarantees will be significantly longer, however we must allow long periods of time to elapse before we can validate this conjecture. This effort is taking place now; we are running jobs in the AWS spot market using DrAFTS bids to determine durability. The experiments to determine durability are on-going. To be statistically significant, they require several months to complete. In the meantime, we have contacted AWS and will meet with them to determine if there is a better way to make prediction in light of the change to the spot market.

#1541215

The UB team has made significant progress towards being able to include cloud-related metrics in Open XDMoD. In addition to the initial set of metrics developed with Eucalyptus data, we have implemented an ETL pipeline to ingest cloud event data from OpenStack log files into XDMoD and are testing the generation of the same metrics that were developed for Eucalyptus. Progress has also been made on the testing of Federated XDMoD. Both Cornell and UCSB have installed locally managed Open XDMoD instances and UB has installed a federation hub XDMoD instance and configured both the Cornell and UCSB Open XDMoD instances to replicate their local data to the federation hub, where it is aggregated to provide a global view of the federation. UB plans to add data from their own local XDMoD instance to the federation and enable Globus Authentication to allow logins from federation members. With the next release of Open XDMoD, we plan to include a beta cloud-metrics realm which will support an initial set of OpensStack metrics.

This is a screenshot of the federated XDMoD page:



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

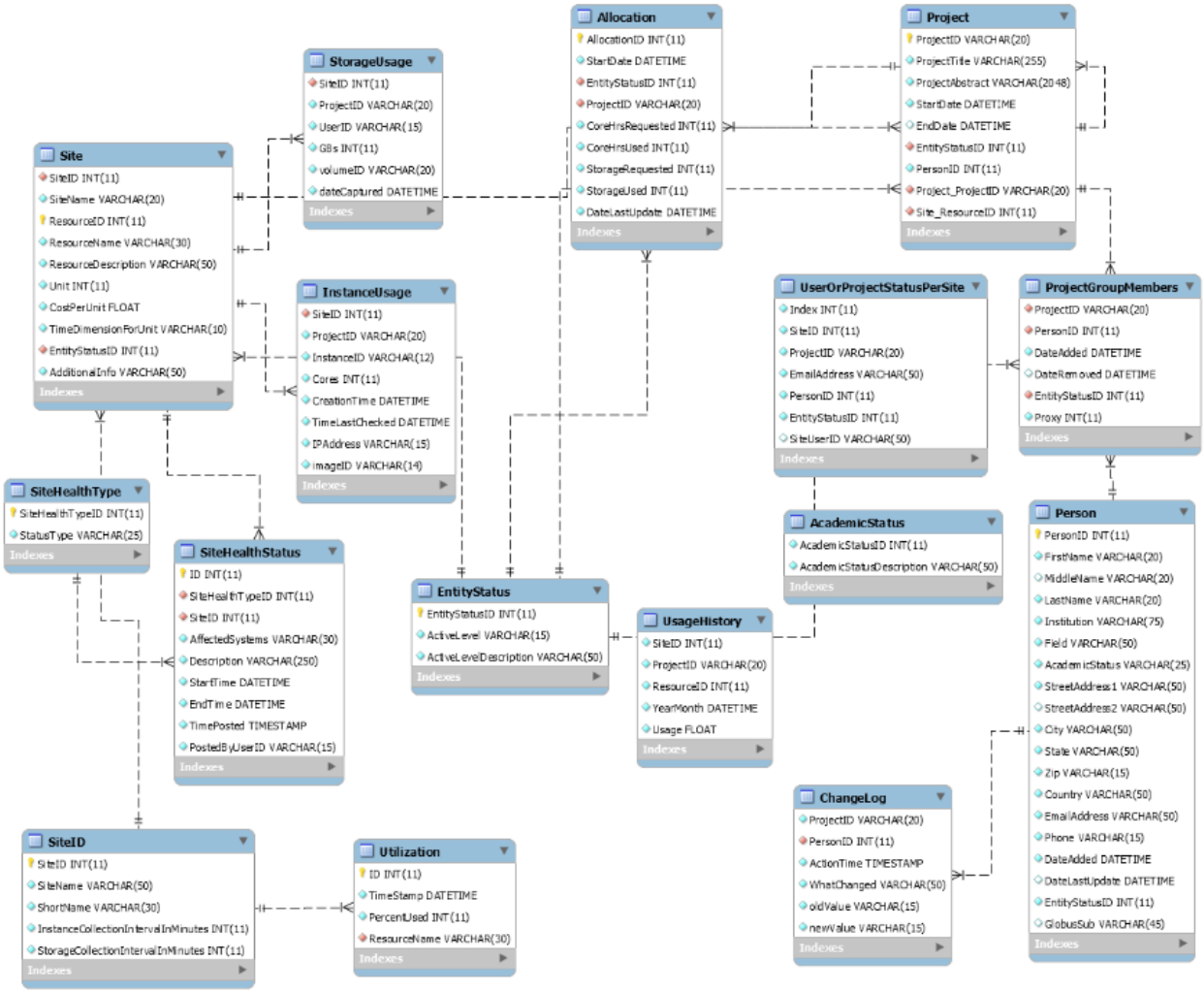
#1541215

3.3 Allocations & Accounting

Three additional tables were created:

- AcademicStatus – so that consistent information is provided into the DB about the users.
- UsageHistory – mentioned above
- UserOrProjectStatusBySite – to keep track of projects and users on each site.

Multiple new Stored Procedures were created for building the web portal functionality of adding users and projects. The updated database schema is below.



#1541215

4.0 Research Team Support

4.1 Science Use Case Team Updates

This quarter's core science use case support focused on building out infrastructure and documentation so that researchers can run their codes efficiently and at scale. Systems were launched in the Jetstream OpenStack domain and user documentation was provided for researchers to provision and connect more resources. Initial feedback indicated that more automation was needed to productively enable researchers to operate at scale, and we have been developing scripting tools to accommodate these requests. Additional details are provided in "Use Case 5: Water Resource Management Using OpenMORDM" described below.

Five Aristotle REU students have been selected and have commenced work with their respective science teams.

Additionally, we are working on 2 demos for the *PEARC18 Conference*.

Use Case 1: A Cloud-Based Framework for Visualization & Analysis of Big Geospatial Data
Varun Chandola's UB team had 2 new papers accepted by the *Big Earth Data Journal* that acknowledge the NSF DIBBs grant:

- "A Survey of Analytical Methods for Energy-Water Nexus Knowledge Discovery"
- "Machine Learning for Energy-Water Nexus: Challenge and Opportunities."

We submitted a third paper to the *ACM GIS Conference.* It describes the analytical capabilities of the *webGlobe* visualization and analysis tool running on Aristotle.

Use Case 2: Global Market Efficiency Impact
UB's Dominik Roesch is currently investigating:

- How price deviations (market inefficiencies) affect liquidity (the ease at which you can buy or sell). His new approach will be completed soon with finding submitted to a journal in 1-2 months.
- How the recent increase in the tick-size (the minimum price movement, of US stocks) affects liquidity. This is joint work with PhD student Albert Lee and Kee Chung. They plan to wrap up the first version by the end of July and will then submit their findings to several conferences.

Professor Roesch will teach 2 PhD students to use the OneTick framework and the underlying data hosted on Aristotle this summer.

A few weeks ago, we started a student project as part of Alan Hunt's CSE 611 course. The students' goal was to program a backend that would keep the Aristotle data up-to-date and create a financial dashboard for convenient data access and to showcase research results. Unfortunately, the student was not successful. Roesch discussed this with Hunt and they might try again with a different student next semester

Use Case 3: High Fidelity Modeling and Analytics for Improved Understanding of Climate
Postdoc associate Tristan Shepherd and Cornell professor/Aristotle use case lead Sara C. Pryor report that the précis objectives of their current suite of simulations are:

#1541215

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).

During this quarter, our team has placed an enormous amount of effort on addressing two important bottlenecks to making progress on our science objectives due to challenges in working on the cloud. These are articulated in Sections A & B below. We have also engaged in a third component of achieving our science objectives. This is described in Section C below.

Section A: Analyses of the impact of changes to the computational node (objective 3 above)
Background: We conducted a long-term simulation for 2001-2016.In the middle (i.e., at end of the simulation of 2007, Cornell applied a change to the node on which we were running).

Research question: Did those changes impact the WRF simulations of wind fields?

Model approach: Re-simulate a 3 month period done on the old node on the new node.

Analysis approach: Pairwise (i.e., every 10-minutes) comparison of wind speed from the two simulations; differences in 10-minute wind speeds from 3rd layer (close to typical wind turbine hub-heights; i.e., 83 m and at 10-m).

Results: Plots below illustrate: Spatial maps of the mean wind speed and the mean difference in each grid cell. Cumulative density functions of the sample of all mean differences from all grid cells and at two sample individual grids (i.e., time series of differences).

Implications: There is an impact from the changes applied to the node (Fig. 2 and 3). The impacts appear 'stochastic' i.e., are symmetric around zero (Fig. 4) and modest onshore, and are of modest magnitude compared to mean wind speeds (Fig. 1), and do not appear to scale with wind speed (Fig. 5). However, the magnitude of these effects is not negligible in the context of the wind energy industry, which uses a benchmark for uncertainty in mean wind speeds of < 0.3 m/s.

Inference: Fortunately our statistical analyses of the two (duplicate) simulations imply that the entire 16 year simulation can be treated as a consistent sample and used for the calculation of the variability of annual energy production from wind turbines given the 'disruption' is comparatively modest and not systematic. However, these analyses clearly demonstrates there needs to be stability of the node configuration during long-term simulations, otherwise it may inadvertently degrade the quality/continuity of the simulation. This challenge is unlikely to be encountered on traditional HPC because the simulation speed is so much faster

#1541215

on an HPC with many processors (we typically run with 128 or 256), relative to the cloud node with only a few (12 or 24).

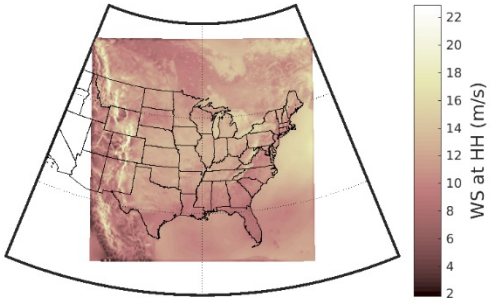


Fig 1. Spatial map of the mean wind speeds at wind turbine hub-height (i.e. approx. 83 m) for all 10-minute periods during Oct-Dec 2007 from simulations conducted on the original Aristotle node.

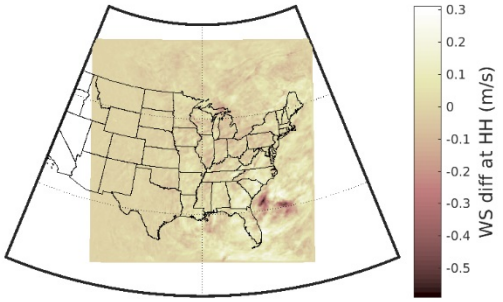


Fig 2. Spatial map of the mean difference in wind speeds at wind turbine hub-height arising from change in Aristotle node.

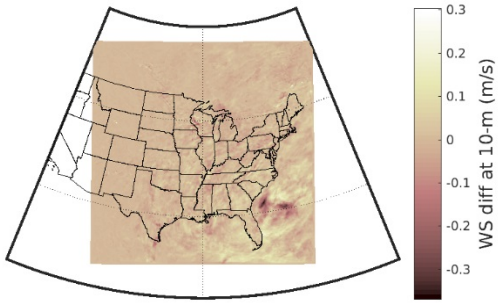


Fig 3. Spatial map of the mean difference in wind speeds at 10-m arising from change in Aristotle node.

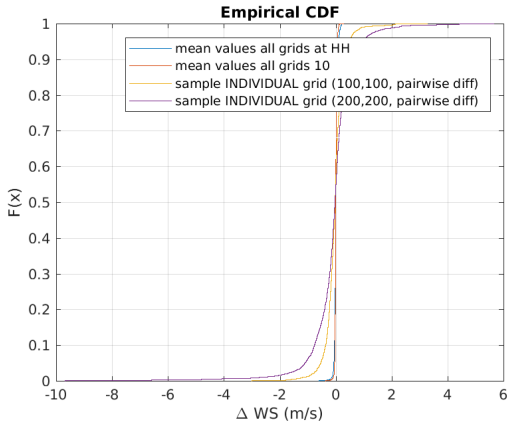


Fig 4. Cumulative density of mean difference in wind speeds at 10-m and wind turbine hub-height (all grid cells). Also CDF of the time-wise (every 10-min) differences in two specific grid cells (100,100 and 200,200) within the 319 by 319 grid cells.

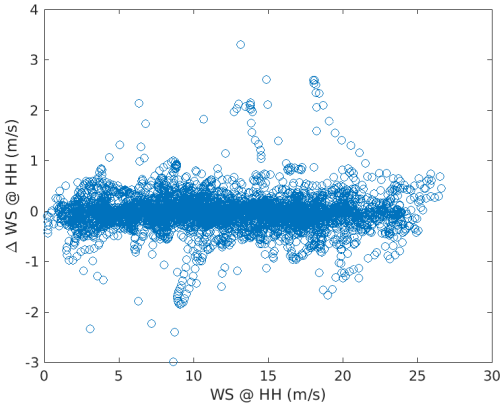


Fig 5. Scatterplot of 10-minute differences in wind speed due to changes to the node versus the wind speed at wind turbine hub-height.

#1541215

Section B: Attempts to compile WRF with Intel compiler to examine simulation sensitivity (objective 3)
Background: Our simulations with the WRF model on Aristotle have been conducted using the Fortran compiler. However, WRF is typically compiled using an Intel Fortran compiler. We sought to examine the degree to which our simulations differ according to the compiler used. It is plausible to postulate there will be an effect. Intel compiled versions of WRF are typically faster (information provided by our collaborators at the Danish Technical University). It appears this speed-up is related to use of certain approximations in Intel.

Research Question: Would use of a different compiler impact the WRF simulations of wind fields?

Approach: Over the last 3-6 months we have been attempting to build the WRF model with the Intel compiler. This has been a challenge and we have been unable to achieve a successful compile to date. We will be revisiting this problem in collaboration with Aristotle use case consultants.

Section C: Analyses of long-term variability in wind resources (objective 4)
Background: Inter-annual variability (IAV) of expected annual energy production (AEP) from proposed wind farms plays a key role in dictating project financing. IAV in pre-construction projected AEP and the difference in 50th and 90th percentile (P50 and P90) AEP derives from variability in wind climates. However, the magnitude of IAV in wind speeds at/close to wind turbine hub-heights is poorly constrained and maybe overestimated by the 6% standard deviation of mean wind speeds that is widely applied within the wind energy industry. Thus, there is a need for improved understanding of the long-term wind resource and the inter-annual variability therein in order to generate more robust predictions of the financial value of a wind energy project.

Approach: We have analyzed our long-term simulations with the WRF model (Feb 2001-Dec 2016) to examine IAV of annual mean wind speeds at/near to typical wind turbine hub-heights and found that IAV of mean wind speeds in these simulations is lower than is implied by assuming a standard deviation of 6%. We have also applied the power curve of the most commonly deployed WT in the domain to post-process the 10-minute wind speed output into estimated AEP.

Results: Rather than in 9 in 10 years exhibiting AEP within 0.9 and 1.1 times the long-term mean AEP (as implied by use of a standard deviation of 6%), our results indicate that over 90% of the area in the eastern USA that currently has operating wind turbines simulated AEP lies within 0.94 and 1.06 of the long-term average. Further, IAV of estimated AEP is not substantially larger than IAV in mean wind speeds. These results indicate it may be appropriate to reduce the IAV applied to pre-construction AEP estimates to account for variability in wind climates, which would decrease the cost of capital for wind farm developments.

Comment: These analyses were rendered possible by mounting a huge (100TB) hard drive to an Aristotle instance.

Activities planned for next 3 months:
It is our intention that work over the coming 3 months will focus on Section B assuming suitable support is available, and in conducting simulations to test the impact of WT on regional climate using two different WT parameterizations and a triple nested domain (objective 5). I am also supervising an REU student to work on spatial coherence of wind fields based on the long-term simulations (objective 2).

NSF
#1541215

Journal manuscripts:

Pryor, S.C., Barthelmie, R.J. & Shepherd T.J. (2018). The influence of real-world wind turbine deployments on regional climate. Journal of Geophysical Research: Atmospheres 123(11). https://agupubs.onlinelibrary.wiley.com/doi/epdf/10.1029/2017JD028114

Pryor, S.C., Barthelmie, R.J., Hahmann, A., Shepherd T.J. & Volker, P. (in press). Contemporary wind turbine deployments have a minor impact on regional climate. Journal of Physics: Conference Series.

Pryor, S.C., Shepherd, T.J. & Barthelmie, R.J. (in review). Inter-annual variability of wind climates and wind turbine annual energy production. Wind Energy Science.

Conference presentations given/abstracts submitted:

Pryor, S.C., Barthelmie, R.J., Hahmann, A., Shepherd, T. & Volker, P. (June 2018). *Contemporary wind turbine deployments have a minor impact on regional climate*. Presentation at the Science of Making Torque from Wind, Milan Italy.

Shepherd, T.J., Volker, P., Barthelmie, R.J., Hahmann, A. & Pryor, S.C. (June 2018). *Sensitivity of wind turbine array downstream effects to the parameterization used in WRF*. Presentation at WRF/MPAS User's Workshop, Boulder, CO.

Pryor, S.C., Barthelmie, R.J. & Shepherd, T. (April 2018). *Do current and near-term future wind turbine deployments have a substantial impact on regional climate*? Invited presentation at European Geosciences Union General Assembly 2018, Vienna, Austria.

Pryor, S.C., Barthelmie, R.J., Biondi, T. & Shepherd, T. (January 2018). *Improved characterization of the magnitude and causes of spatio-temporal variability in wind resources.* Poster presentation at 98th American Meteorological Society Annual Meeting (31st Conference on Climate Variability and Change), Austin TX.

Shepherd, T., Barthelmie, R.J. & Pryor, S.C. (January 2018). *Assessing the fidelity of the North American wind climate and impacts of wind farms using high resolution modeling.* Presentation at 88th American Meteorological Society Annual Meeting (21st Conference on Planned and Inadvertent Weather Modification), Austin TX.

Pryor, S.C., Barthelmie, R.J. & Shepherd, T. (September 2017). High-fidelity simulations of the downstream impacts of high density wind turbine deployments. Poster presentation at 4th Workshop on Systems Engineering for Wind Energy, Roskilde, Denmark.

Use Case 4: Transient Detection in Radio Astronomy Search Data

For the Radio Transient Detection Use Case, we have successfully implemented an improved reproduction of the PALFA2 pipeline for detecting single pulse candidates that may be Fast Radio Burst (FRB) sources. The improved pipeline includes utilization of PRESTO functionality, modulation index calculation, parameter customization, and the production of graphic output of the data for human inspection. Remaining steps for this version of the pipeline include improved graphic output, customization of graphing parameters, and reduction of candidate set based on modulation index threshold. This comprises the prototype.

This improved version of the PALFA2 pipeline serves as a starting point for the implementation of a new pipeline to search for FRBs, which will extend beyond the current pipeline's capabilities. We are also currently in the process of determining final requirements for the aforementioned new pipeline which will support the desired pluggability and customization for selection of different methods for determining single pulse candidates. We are in the process of re-implementing and improving upon modulation index calculation as one potential option, as well as exploring other methods to implement.

#1541215

Use Case 5: Water Resource Management Using OpenMORDM
Our investigation is currently focused on: (1) developing automated tools for provisioning and connecting cloud VMs, in particular Ansible, Terraform, jclouds, and bash scripts over the Jetstream OpenStack API, (2) developing productive contexts for orchestrating container applications at scale, in particular Docker swarm, Singularity, and Open MPI communication parameters. While the investigation and prototyping is occurring in the context of the Water Resource Management Use Case, these tools will be useful to all of the use cases, and the Singularity project will allow for bursting to traditional HPC resources if required.

Use Case 6: Mapping Transcriptome Data to Metabolic Models of Gut Microbiota
Previous work on tuning the SteadyCom algorithm and setting up infrastructure in Red Cloud (Cornell's Aristotle site) has allowed us to perform a first pass of Flux Variability Analysis (FVA) on the four microbes composing the Drosophila gut microbiome: an L. plantarum model with 1028 reactions and 933 metabolites, L. brevis with 946 reactions and 840 metabolites, A. tropicalis with 1228 reactions and 1108 metabolites, and A. pomorum with 1058 reactions and 938 metabolites.

During this process we discovered a discrepancy between the original FVA results for a single model, and the trivial case of constructing a multi-species model from a single model; this was due to the COBRA script not preserving extracellular flux constraints. We are in the process of writing a test for the COBRA Toolbox to reapply constraints from a single model to verify FVA results are the same in the trivial multi-species model. Once this is complete, we can then move on to applying various constraint sets to the community model (the entire gut microbiome) for SteadyCom's implementation of FVA and FBA (Flux Balance Analysis) and ordinary FBA and FVA analyses; the expectation is that the SteadyCom results should be more realistic than the ordinary counterparts due to enforcement of a single steady-state growth rate across all microbes in the system, which gives rise to numerous modeling improvements over joint FBA and FVA.

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

Frost prevention project: The science team has begun Phase 2 deployment for the citrus orchard at Visalia, California. The installation infrastructure (30-foot mounting poles for sensors and networking) has been completed and the networking is in place. The team is waiting for the local ISP to enable a new network for the sensing infrastructure which is scheduled for late June. The next steps are to deploy the sensors and to install video cameras both to act as sensors (via image processing) and to monitor the deployment. The team is planning a trip to meet with the local growers and try and complete the Phase 2 deployment in July 2018.

Differential irrigation project: The differential irrigation project is making progress. The science team has a new board design that corrects a bug in the Analog-to-Digital conversion circuitry and also allows a much larger number of analog sensors to be connected to each controller (dramatically lowering the installation cost). A test deployment of the new board is taking place this week; if this goes smoothly, the new board will be deployed at scale (approximately 30 sensors) ahead of the up-coming growing season.

Grape irrigation project: The grape irrigation project at Sedgwick is sun setting. We helped the land managers reduce the irrigation requirements substantially, but the land use is now shifting to ranching (due to local infrastructure issues). We are working with the land managers to determine how Aristotle might accommodate this new land use case.

#1541215

5.0 Community Outreach and Education

5.1 Community Outreach

- Aristotle's UB team presented "Federated Keystone Single Sign-On with FreeIPA and OpenID Connect: at the OpenStack Summit:" https://www.openstack.org/videos/vancouver-2018/federated-keystone-single-sign-on-with-freeipa-and-openid-connect
- Aristotle co-PI Wolski presented "Emerging Trends in the Economics of Cloud Computing" at the NSF Workshop on Cloud Economics, May 2018, Stanford University, Palo Alto, CA: https://federatedcloud.org/papers/nsf-econ-2018.pdf
- Wolski also presented "Tracking Casual Order in AWS Lambda Applications" at IEEE International Conference on Cloud Engineering, April 2018, Orlando, FL: http://conferences.computer.org/IC2E/2018/program.htm
- Aristotle REU student Joan Song presented "Genome-scale Metabolic Modeling of Gut Microbiota in the Fly Gut" poster at the American Society for Engineering Education Meeting: https://federatedcloud.org/papers/Poster4_19_18.pdf
- News releases and news features:
    - "Apply now for NSF-funded Research for Experience for Undergraduates summer positions at Cornell" - https://www.cac.cornell.edu/about/news/180420.aspx
    - "Cornell receives NSF award to implement federated cloud software" - https://www.cac.cornell.edu/about/news/180416.aspx
    - Use case scientist Angela Douglas featured in "How Microbes Benefit Human Health" - https://research.cornell.edu/news-features/how-microbes-benefit-human-health

5.2 Education

- Wolski taught a class at the UCSB College of Engineering on how to use OpenStack.
- Cornell hosted a half-day education workshop on AWS storage.
- Next quarter, the Cornell Aristotle team plans to build and teach a workshop for Upward Bound high school students on introductory programming concepts.