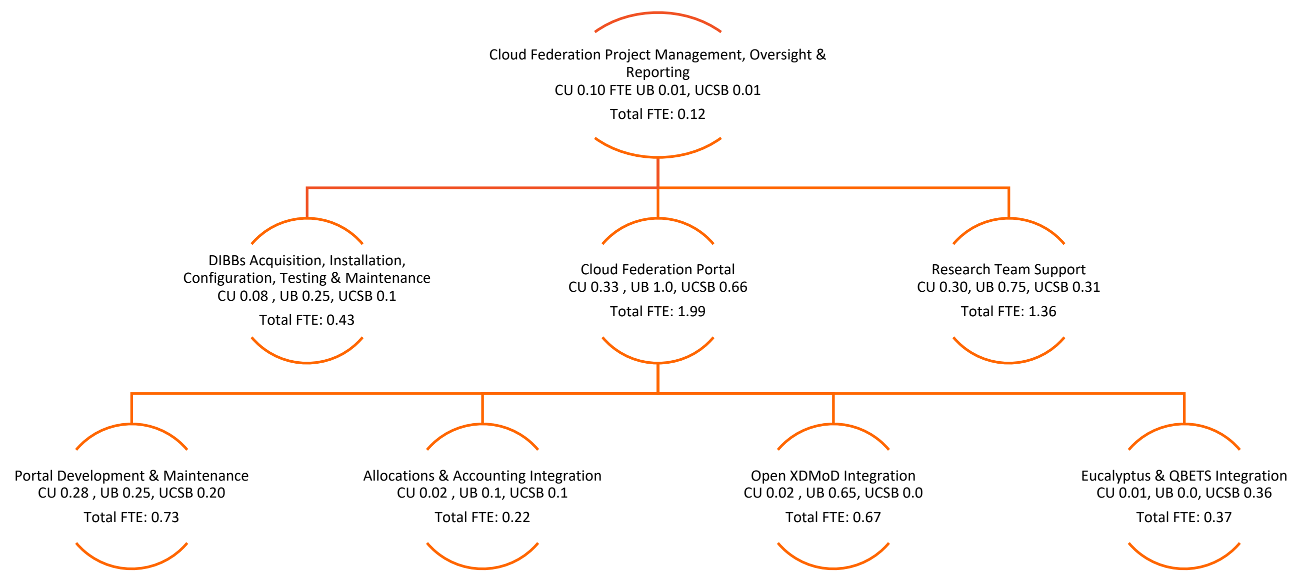
CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation

Program Year 5: Quarterly Report 4

9/14/2020

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

This is the Program Year 5: 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

1.0 Cloud Federation Project Management, Oversight & Reporting 3
1.1 Subcontracts 3
1.2 Project Change Request.. 3
1.3 Project Execution Plan.. 3
1.4 PI/Partner Virtual Meetings 3
1.5 Project Planning Virtual Meetings/Status Calls 3

2.0 DIBBs Acquisition, Installation, Configuration, Testing & Maintenance Report........... 8
2.1 Hardware Acquisition.. 8
2.2 Installation, Configuration, and Testing 9
2.3 Federated Identity Management... 9
2.4 Cloud Status by Site 9
2.5 Tools.. 10

3.0 Cloud Federation Portal Report.. 10
3.1 Software Requirements & Portal Platform 12
3.2 Integrating DrAFTS into the Portal 12
3.3 Integrating Open XDMoD into the Portal 12
3.3.1 Application Kernels (AK) Containerization in the Cloud 12
3.3.2 XDMoD Cloud Integration.. 13
3.4 Allocations & Accounting 13

4.0 Research Team Support 14
4.1 Science Use Case Team Updates 14
Use Case 1: A Cloud-Based Framework for Visualization & Analysis of Big Geospatial Data 14
Use Case 2: Global Market Efficiency Impact... 14
Use Case 4: Transient Detection in Radio Astronomy Search Data 15
Use Case 5: Water Resource Management Using OpenMORDM 16
Use Case 6: Mapping Transcriptome Data to Metabolic Models of Gut Microbiota............... 16
Use Case 7: Multi-Sourced Data Analytics to Improve Food Production & Security.............. 16

5.0 Community Outreach and Education 17
5.1 Community Outreach 17
5.2 Education 17

1.0 Cloud Federation Project Management, Oversight & Reporting

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

1.2 Project Change Request
A No Cost Extension request was submitted and approved 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.

1.4 PI/Partner Virtual Meetings
- The PIs met to discuss submitting a No Cost Extension (NCE). At Cornell, late approval of supplemental funds reduced our time to work on the proposed effort; this also delayed UCSB in hiring a postdoc to work on the supplemental. A one-year NCE was approved by NSF for 10/1/2020 to 9/30/2021 without additional funds. We will complete all remaining project and supplemental work during the NCE period. The final Financial Report, Technical Report, and Project Outcomes Report will be submitted by the specified 1/28/2022 due date.
- Aristotle's infrastructure team had additional discussions with Red Hat's technical staff on how to move to OpenStack v.16. We are optimistic Red Hat will help us design the best path forward.

1.5 Project Planning Virtual Meetings/Status Calls

6/23/2020 progress call:

- We successfully installed Nvidia CUDA Toolkit 10.0 and TensorFlow in Anaconda using the GPUs in Red Cloud. A Docker image with TensorFlow GPUs and a Jupyter Notebook is being tested. There are issues with the Jupyter Notebook inside of Docker not finding the GPUs. We plan to document how to effectively use and connect these technologies.
- The Aristotle team at Cornell and UCSB, with REU support, are building an OpenFOAM instance for researchers to use on the Citrus Under Protective Screening project. The CUPS project involves a team of researchers specializing in computer science, entomology, farm operations management, and agriculture pathology. They are attempting to grow citrus in an indoor environment to protect the trees from the bacterial disease HLB which is spread by the Asian citrus psyllid and reduced Florida's citrus production by 50 percent from 2003-2017. OpenFOAM simulations on Aristotle will be used to better understand the flow of air in order maintain proper temperatures within the indoor growing environment.
- There is an issue with entering our Nix environment inside a Singularity container which means we are not getting the exact reproducibility we desire. The NixOS is trying to change the file system while Nix is trying to enforce the Read-Only file system. We have a workaround while we work to resolve this issue.

6/30/2020 status call:

- Cornell received 4 more Nvidia T4s with SR-IOV enabled. We are trying to disable that so we can virtualize the GPUs. We are assigning a GPU per instance. Cornell is also working on upgrading their network.
- UCSB is working with Red Hat on how to unblock the CLI.
- We're excited to report that one of our former REU students, Nevena Golubovic, who went on to complete her PhD is now using Aristotle technology (our Centaurus cloud service for K-means clustering: https://sites.cs.ucsb.edu/~ckrintz/papers/centaurus-journal18.pdf) to correlate California power grid usage data with water usage data. Golubovic is with a start-up that received a grant from the California Energy Commission. Two Aristotle REUs to date have gone on to complete their PhDs and a third has been accepted in a PhD program.
- Our current REU students are learning it's not trivial doing GPUs with containers in the cloud.
- We plan to assess a wide range of AWS instance types with our DrAFTS 2.0 tool.
- Our paper—"Reproducible and Portable Workflows for Scientific Computing and HPC in the Cloud:" https://dl.acm.org/doi/abs/10.1145/3311790.3396659—was accepted by PEARC '20. We will be providing a workshop on this topic as well.
- The Aristotle container investigation team added internal "users" to the team who have no container experience. These users will keep notes on their experiences learning about containers, choosing the right container, and running science applications with containers. Our goal is to capture their experiences and share it with the CI and research communities who to a large degree have no experience with containers. We will focus on identifying best practices and document lessons learned. The users will start by containerizing HPL benchmarks with Docker and Singularity. Others science applications such as Lake Problem, WaterPaths, FRB, and/or WRF may then be used to walk them through what it takes to containerize an app. Our team will consult with the users on how to install and run sample applications, but they will do the work and take notes in order to be sure that we identify the "gotchas." We believe that fresh eyes are essential to better understand the process. We will capture the results in a white paper and share it with the community. The users will start with a single node application and then move on to multi-node MPI.

7/10/2020 status call:

- An UCSB REU student built a Visualizer that looks at science team data in real time. It will run on a MAC or hosted. We'll demo it during our next meeting to get feedback. The REU OpenFOAM CFD modeling project is going well. We believe ARM will perform better for CFD than Intel and will be having a call with the ARM CFD team in the UK with the student participating.
- Cornell is working with UCSB on a CLI issue which needs to get resolve to finish DrAFTS 2.0.
- We are pursuing the extension of Red Hat support for another year. Budget and accounting for support is being sorted out by Cornell.
- Cornell added more GPU nodes for the Red Cloud pool. They have a total of 12 to date which will soon be expanded to 20. A science use case is being ported to use the GPUs and the OpenFOAM project is also planning to test GPUs. UB is working on quotes for their final round of hardware purchases, including new control servers.
- Dartmouth is taking a lot of notes for our how-to guide. They are currently trying to figure out the right OpenStack mechanism for unsticking a bit; they need to find the right configuration file and UCSB is assisting.

- Tomorrow we are having another meeting with Red Hat technical support who is assisting us with templates for building a new cloud and helping UCSB with its install. UB is testing the new install with TripleO 16.1 which is still in beta. They're going to use the beta to test the deployment and see how far they can get so that they are more prepared when 16.1 comes out.
- The Aristotle container investigation team reached out to the Cornell X-Container team to further that collaboration.

7/21/2020 progress call:

- We made a Docker 3.8 container; however, v.4 still doesn't build. We have line-by-line build scripts and will make them public. We have a working Docker container that anyone can build and ran NCAR's commands from their GitHub.
- Cornell's REU student has benchmarks running and good numbers for our Terrraform Ansible cluster. We will deploy multi-VM MPI next and are documenting as we go. We are planning 16 node parallel executions on Red Cloud, Stampede2 bare metal and containerized runs, and possibly Google containerized runs. Our focus is on investigating the portability for non-trivial scalable real parallel MPI codes and an exhibition of what the portability of that performance looks like.

7/28/2020 status call:

- The Cornell Science Use Case Team lead is very happy with the REU students that we selected. Vaillancourt and Wineholt are marshalling their efforts, including GPU work which is difficult in the cloud. Wineholt is creating some excellent documentation to assist the REUs. The REUs participate in Aristotle meetings as well general Cornell CAC meetings.
- UCSB is making good progress with their REU as well. The visualization tool he developed is being used to visualize real-time telemetry data from ag sensors. A data acquisition system pulls in all the sensor data and stores it in a website. The visualizer tool then connects directly to the data acquisition system to produce a live graph of the data. You can download it and run it on your MAC or you can run as a software service (it uses Docker), deploy it in the cloud and run. The REU student and his mentors are thinking of writing a paper about this tool and the REU will be applying to the UCSB PhD program this fall. In the meantime, he is working with a programmer and a physics student to try to get a CFD code running to do the Citrus Under Protective Screening environmental modeling in real-time. The back-end will run on Aristotle. While the Citrus project is moving forward, other research projects are stalled due to COVID-19.
- All energy at the Dartmouth team is focused on getting their OpenStack deployment up and running.
- We discussed when users sign up for services, they should always be able to see options (all buttons) in the sign-up process whether they are enabled or not.

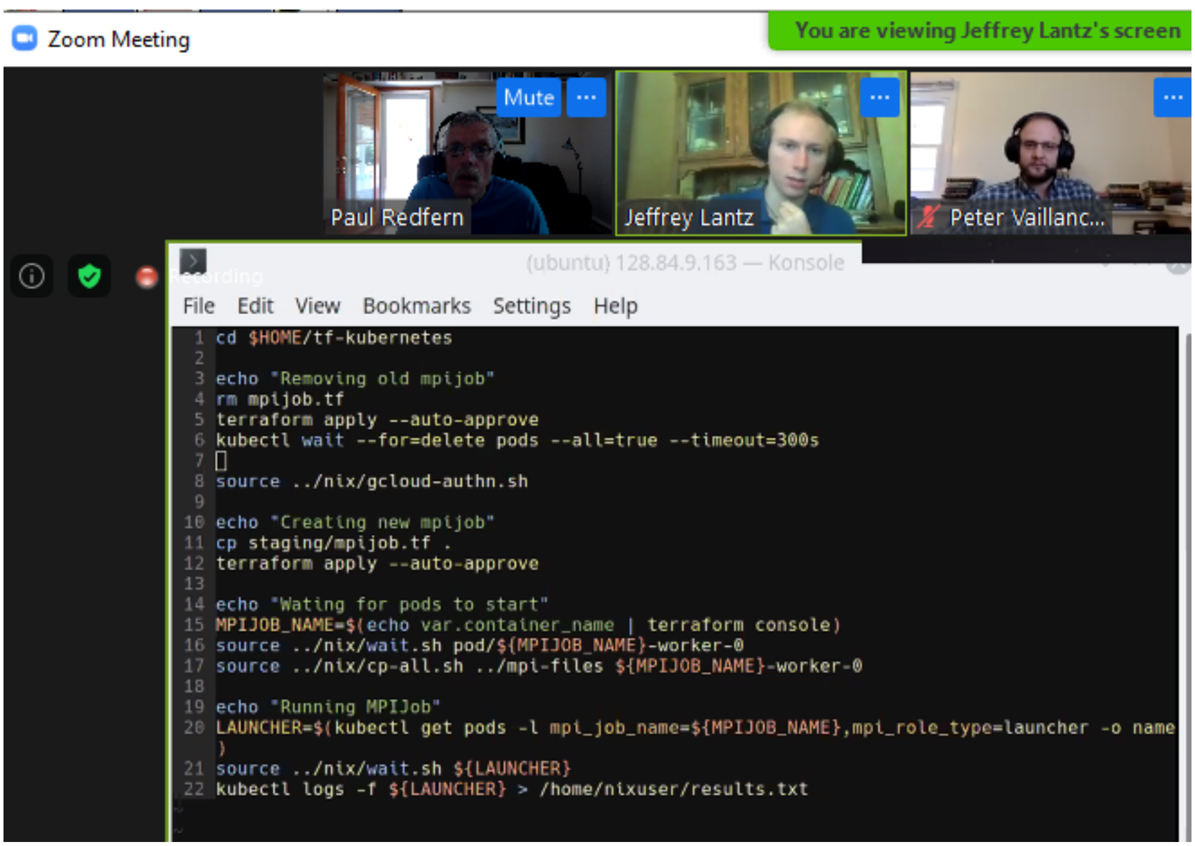
8/3/2020 progress call:

- Regarding software to be compared, we have a Docker container on Terraform/Ansible as well as Terraform/Kubernetes for HPL testing. We also have a container, but no run script, for the WRF model software. Our Terraform/Ansible deployment scripts for Docker containers will form the basis for other science software. Other multi-VM MPI comparisons are under consideration.

- Our Terraform/Ansible method deploys a multi-VM MPI capable cluster in the cloud and Terraform/Kubernetes that is faster than our Terraform/Ansible.
- We plan to pursue a Stampede2 start-up allocation to compare performance on different platforms.
- Our focus in the container investigation is on what new users need to know, what their pain points are, as well as performance. People need to know which container technologies are useful and which are not, and especially, which codes are likely to do well in containers.

8/6/2020 REU student presentations:

- This year's REU students demonstrated a high-level of skill, ingenuity, and dedication. Our mentors, Brazier, Vaillancourt, Wineholt, and Barker, provided advice on how to get started, how to resolve errors while setting up VM instances particularly when running detectable GPUs, etc.
- At the end of his project, for example, Cornell REU student Jeffrey Lantz demonstrated a multi-VM MPI-capable Terraform/Kubernetes Cluster deployment he built with an HPL benchmark.



8/11/2020 status call:

- Cornell continues to help UCSB resolve their CLI issues for OpenStack.
- Walle and Mehringer are creating templates of the portal and the Aristotle database to share with the broader community via GitHub.

#1541215

- We are testing our REU student's OpenFOAM documentation and will be sharing that as well.
- We discussed running a meaningful parallel test set of WRF with data sets on a wide variety of cloud and HPC platforms and the need for WRF expertise from the scientists to do that.
- The visualizer tool built by the Aristotle REU student at UCSB was demonstrated. The snapshot below is of live temperature data coming from the weather sensors at UCSB's Edible Campus farm site on the coast, from the Santa Barbara airport (published on the web), and from the backyard of the student's mentor located 3.5 miles from the coast. This tool can run locally on a Mac in your web browser or run as a service in a VM so it can be used locally as a visualization tool or it can be hosted in the cloud for a larger user community. There is a lot of packaging and deployment innovations in this work and we plan to use it as a live visualizer for the various sensor data we have. We are considering using this tool for a big history study on climate change as well.

8/25/2020 status call:

- Our Red Hat renewal for another year has been processed.
- At Cornell, all our GPU nodes are in place now: 20 T4s and 4 V100s. We are working on documentation for our GPU nodes so that we can promote their availability and provide the necessary guidance.
- UB received new servers with the new version of Red Hat Director. We are testing it and will then make the transition to it. We are experiencing more interest in GPUs from scientists and plan on bringing on more users with V100s.
- UCSB got their old cloud up. Our new 16.1 cloud is running. We sent out a list of topics Red Hat can hopefully help us with – migrating workloads, connecting to our various authentication methods, particularly Globus (which is officially supported by Red Hat in 16.1). Red Hat will help

UCSB with their upgrade and with templates so when UB and Cornell are ready, UCSB will be able to help us.

- We plan to continue supporting science teams with reasonable access to Aristotle cloud resources providing they are actively collaborating with us during the no cost extension period.
- The cloud marketplace investigation researchers concluded that the time resolution and data granularity of RightScale's cost tracking tool is not well-suited to analyzing the usage of particular applications or runs, making it difficult to profile individual science workflows. They decided to pursue the cost analysis of multi-VM MPI-capable clusters which appear to be of the most interest to researchers. We published work-to-date on the creation of multi-VM MPI environments in multiple clouds at *PEARC '20*. We will continue this work and create a reproducible container deployment for the WRF weather modeling Aristotle use case which will be extensible to the broader atmospheric community. We then plan to write a paper that captures costing information for comparison of different resources.

8/31/2020 progress call

- The Docker, Singularity, and X-Containers performance team summarized their results to date. They have: (1) developed application containers that are convertible from Docker to Singularity and verified functionality on a variety of cloud platforms, (2) implemented reproducibility of builds and environments within containers by leveraging the Nix package manager, (3) determined a set of recommendations for the conversion of containers from Docker to Singularity for XSEDE HPC machines, (4) created a multi-cloud automated deployment method using Terraform and Ansible for deploying multi-VM MPI-capable clusters with Docker, (5) developed an automated deployment method for Kubernetes using Terraform and created MPI-capable clusters (tested on Google Cloud), (6) developed a cluster benchmarks container in both Docker and Singularity and tested it on multiple platforms.
- Futures plans are: (1) complete benchmarking runs of an MPI sample science application (most likely WRF). This benchmarking will be done on bare metal and containerized with Singularity on an HPC machine, and containerized with Docker for benchmarking on multiple clouds, (2) explore and report on Kubernetes orchestration and considerations for creating and managing workflows with each runtime, (3) explore interoperability between Docker, Singularity, and X-Containers, (4) complete a technical report detailing the performance and considerations for running Docker, Singularity, and X-Containers, (5) share the technical report, lessons learned, and any practical outcomes with the wider scientific community.

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

2.1 Hardware Acquisition

- Cornell had no acquisitions this quarter.
- UCSB ordered 6 cloud nodes for an additional 288 cores and 4608TB RAM along with 4 storage nodes for an additional 768TB Ceph storage (raw).
- UB ordered new infrastructure (servers, switches, and PDUs) to support production and development clouds running Red Hat OpenStack 16.1. They also ordered a data transfer node (DTN) with a 100GB interface to allow for data and images to be more easily transferred.

#1541215

2.2 Installation, Configuration, and Testing

- Cornell added 16 Nvidia T4 GPUs to Red Cloud (acquired via alternative funding). They will be testing their RHOSP 16.1 installation and plan to migrate workloads as soon as possible.
- UCSB upgraded to RHOSP 16.1 and will be assisting Cornell and Buffalo with their RHOSP 16.1 installations.
- UB installed their RHOSP 16.1 development and production clouds. They plan to migrate workloads as soon as possible. They are also setting up their DTN node.
- Dartmouth has installed their RHOSP 16.1 cloud and will be connecting it to the federation this quarter.

2.3 Federated Identity Management

Researchers use single sign-on at any member site.

2.4 Cloud Status by Site

The chart below shows each site's production cloud status. Dartmouth's cloud is in test mode.

	Cornell	**Buffalo**	**UCSB**
Cloud URL	https://redcloud.cac.cornell.edu	https://lakeeffect.ccr.buffalo.edu/ (access only to federation)	https://openstack.aristotle.ucsb.edu/
Status	Production	Production	Production
Software Stack	OpenStack	OpenStack	OpenStack
Hardware Vendors	Dell	Dell, Ace	Dell, HPE, DXC
DIBBs Purchased Cores	*616	**792	***740
RAM/Core	8GB	up to 8GB	9GB Dell, 10GB HPE
Storage	Ceph (1.6PB)	Ceph (768TB)	Ceph (720TB)
10gb Interconnect	Yes	Yes	Yes
Largest Instance Type	28core/192GB RAM	24core/192GB RAM	48core/119GB RAM
Globus File Transfer	Yes	In Progress	In Progress
Globus OAuth 2.0	Yes	Yes	Yes
Total Cores (DIBBs purchased cores + existing cores) = 2776	* 616 additional cores augmenting the existing Red Cloud (1316 total cores).	** 792 total cores (UB Lake Effect Cloud and CCR cloud will be one pool after upgrade to Red Hat Director).	***740 cores in UCSB Aristotle cloud (956 total cores, Aristotle is separate from UCSB campus cloud)

NSF
#1541215

2.5 Tools

- Red Hat – Cornell and UCSB infrastructure teams met with Red Hat OpenStack technical staff again regarding UCSB's RHOSP 16.1 installation. The plan is for Red Hat to assist UCSB with the necessary templates allowing UCSB to support Cornell and UB on their installations.

3.0 Cloud Federation Portal Report

Content updates to the project portal are ongoing (https://federatedcloud.org). In tandem with the database, schema, the main portal site (skeleton, project and allocation functionality, and database integration) was shared with the greater CI and research community via GitHub.

Open XDMoD continues to monitor data ingestion from all sites, as well as provide the utilization data (https://federatedcloud.org/using/federationstatus.php).

The portal planning table was not 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.
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.

#1541215

3.1 Software Requirements & Portal Platform

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

3.2 Integrating DrAFTS into the Portal

DrAFTS was developed and successfully deployed as a tool for predicting how to use the AWS spot market and reduce costs. Coverage in the general press as well as research publications, e.g. *IEEE International Conference in Cloud Engineering* (https://ieeexplore.ieee.org/document/8790118) and conference presentations documented this achievement. While this tool provided value to research groups such as Globus, AWS made changes to their product that obscured our ability to make accurate spot market predictions.

In light of this, we decided to invest additional time and energy in developing a second tool that would run the TOP500 LINPACK Benchmark on all Aristotle instance types and all AWS instance types, and then sort the results by cost and performance. With DrAFTS 2.0, users could ask questions such as: Which AWS instance type is the most equivalent to an instance type in Aristotle? Which is less costly? Which is faster? Or, if I want to spend 20% more and go 30% faster, which instance type should I use?

Unfortunately, both the student working on DrAFTS 2.0 and the Aristotle-supported staff member are now terminating (the student to work for Google and the staff member to enter a Ph.D. program), but both have agreed to try and finalize the needed work on their own time while transitioning into their new respective positions. We believe there is a high probability this work will be completed. If DrAFTS 2.0 is not completed by the project's original end date (9/30/2020), we will complete it as soon as the OpenStack CLI functionality issues have been fully resolved, working during the no cost extension period should that be necessary.

3.3 Integrating Open XDMoD into the Portal

3.3.1 Application Kernels (AK) Containerization in the Cloud

The majority of scientific and engineering applications take advantage of the great advances in modern CPU architectures. For efficient utilization of that computing power, applications must be compiled with a specific CPU target in mind. Because such a target is unknown for a universal Docker container, the generation of compute efficient containers can be challenging. In the early version of AK containers, we manually compiled programs for four generations of vectorized instructions (SSE2, AVX, AVX-2 and AVX-512). In the final version of AK containers, we switched to Spark for CPU specific executables creation. Spark is a package manager for HPC systems that has multiple advantages over the manual installation. It builds software and its dependencies automatically using provided recipes and compiles it for specific CPU architecture, allowing automated container generation for multiple CPU targets while significantly simplifying software updates.

We have created seven AK containers: HPCC, HPCG, IOR, MDtest, NAMD, NWChem and Enzo. Each AK container consists of the application compiled for the four most common vectorized instructions sets and input for that application. The container automatically detects the most suitable executable, detects a number of cores to use, executes the application with provided input parameters, and outputs the results.

#1541215

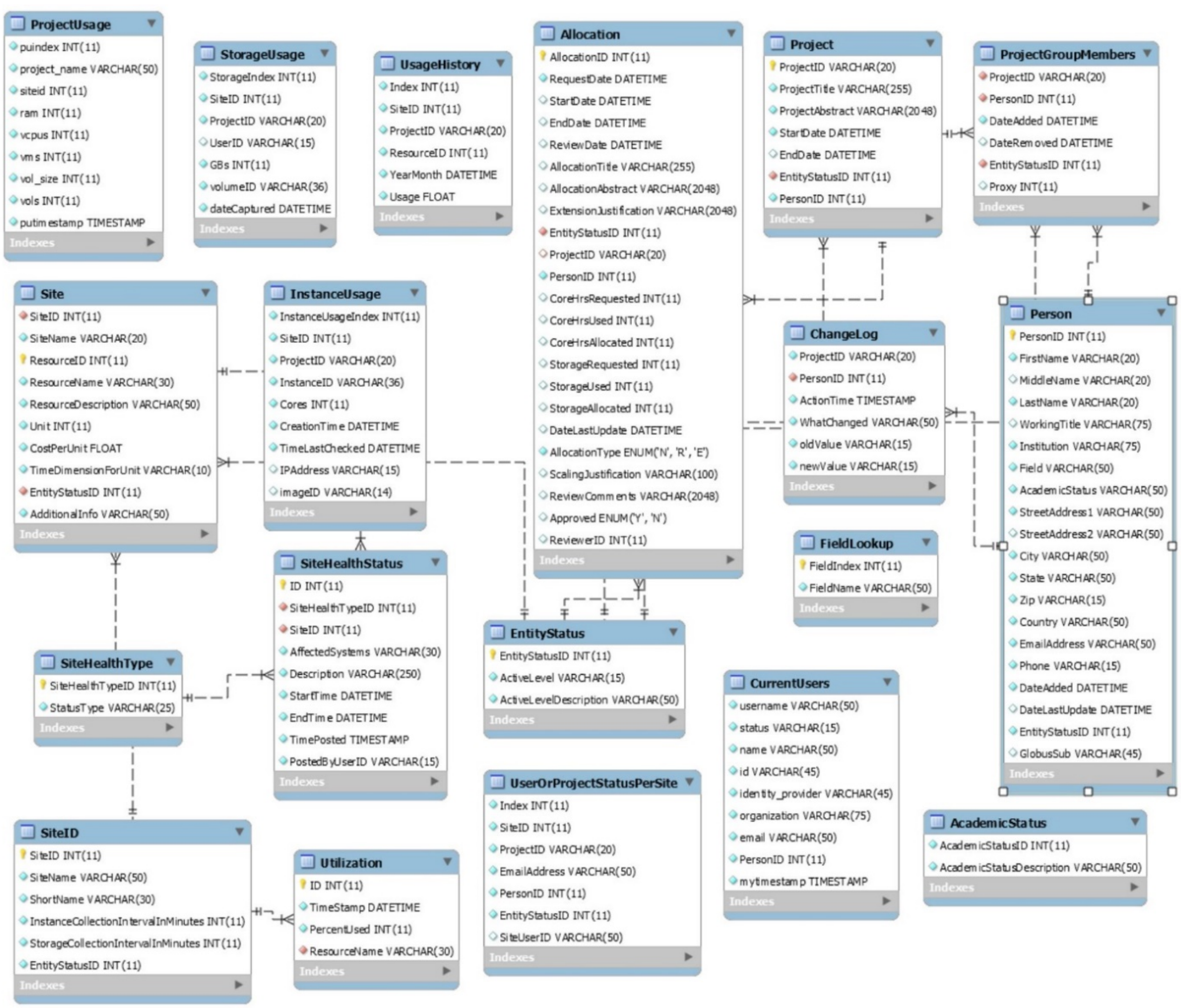
AK containers are used in all Aristotle OpenStack instances within the XDMoD performance monitoring module.

3.3.2 XDMoD Cloud Integration

Federated Open XDMoD, a tool to monitor affiliated computing resources including clouds, has added cloud metrics to its functionality. Cloud metrics are now available for average cores reserved, average memory reserved, average root volume storage reserved, average wall hours per session, total core hours, number of active sessions, number of sessions ended, and number of sessions started. These metrics can be grouped or filtered by instance type, project resource, and VM size (core/memory). For example, here are XDMoD cloud metrics for Cornell's Red Cloud (note: the sweet spot appears to be 8-core instances): https://openxdmod.cac.cornell.edu/#tg_usage?node=statistic_Cloud_none_cloud_avg_cores_reserved.

3.4 Allocations & Accounting

The database schema was shared with the broader community via GitHub.



#1541215

4.0 Research Team Support

4.1 Science Use Case Team Updates

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

Aristotle use case scientist and UB CS professor Varun Chandola created a new platform to support research activities at the University at Buffalo called the OUTSTEPS Community Platform. It facilitates interactions between more than 100 practitioners, community stakeholders, and researchers from 8 urban communities with universities located along the Lower Great Lakes coast. Universities participating include the University of Michigan, University of Toledo, Cleveland State University, Penn State at Behrend, University at Buffalo, SUNY ESF, SUNY Oswego, and Clarkson University.

The framework, hosted on the Aristotle cloud, is available at https://www.community-outsteps.org. We recently submitted a planning proposal to the NSF in response to the NSF Civic Innovation Challenge that builds upon this community in order to develop new ways to increase community resilience to natural disasters such as the COVID-19 pandemic in Erie County, NY. As a next step, we are merging this platform with the webGlobe infrastructure (developed earlier as part of the Aristotle grant) to allow practitioners to share geo data and knowledge with each other.

Use Case 2: Global Market Efficiency Impact

This Aristotle-enabled research team led by UB's Dominik Roesch received a revise and resubmit request from the *Journal of Financial Markets* for their paper entitled "Financial market frictions and learning from the stock price."

Use Case 3: Application of the Weather Research and Forecasting (WRF) Model for Climate-Relevant Simulations on the Cloud

Précis objectives of simulations performed using the Weather and Research Forecasting (WRF v.3.8.1) model by Cornell Professor Sara C. Pryor's group:

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 90^{th} 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).

Activities during this quarter:

- Finalizing analyses of previously conducted WRF simulations:
  - *Objective 1:* Analyses of our convection-permitting high-resolution simulations with the WRF model conducted at 2 and 4 km grid-spacing (resolution) over the eastern USA to examine the presence/absence of low-level jets and examine the frequency of non-ideal wind speed profiles and the impact on those phenomena from varying resolution.
  - *Objective 5:* Analyses of our WRF output to characterize wind farm wakes (i.e., disruption of downstream near-surface properties) as simulated using the Fitch and EWP wind farm parameterizations applied at 2 and 4 km resolution, and developing a rigorous framework to use those simulations to provide guidance for a planned field experiment.
- Guiding research by the NSF REU student focusing on forecasting of wind gusts based on antecedent atmospheric conditions and assessment of relative skill of generalized linear models v. machine learning.
- We were unable to undertake the planned final WRF simulations because using a single node VM would be too slow and the development of a multi-node instance is temporarily delayed due to staff availability issues at CAC. However, we were able to pivot and innovate on a different science hypothesis focused on simulation of derechos (fast-moving damaging deeply convective systems). We selected a case study (29 June 2012; https://www.weather.gov/lwx/20120629svrwx) and undertook a five member ensemble of the 5 days around the event using different microphysics schemes. Those simulations are almost completed as of 19 August 2020 (4 of 5 are completed and the simulation with the Morrison scheme running) and the output is being transferred to an XSEDE instance for analysis. Pryor wrote and was awarded an XSEDE supplement to enable the transition from conducting WRF simulations on Aristotle to XSEDE resources.

Journal manuscripts published this quarter:

- Barthelmie R.J., Shepherd T.J., Aird J.A. and Pryor S.C. (2020): Power and wind shear implications of large wind turbine scenarios in the U.S. Central Plains. *Energies* ***13***(16) 4269 https://doi.org/10.3390/en13164269
  Note: The analyses were performed on an XSEDE allocation but the original simulations were performed on Aristotle.

Use Case 4: Transient Detection in Radio Astronomy Search Data

The Jim Cordes research team and Aristotle use case scientists completed and tested the single container of radio astronomy software combining the pipeline components developed for Pulsar and other transient detections that can be deployed either on the cloud (Docker) or on an XSEDE HPC resource (Singularity), and have released the container build files and images publicly via GitHub and DockerHub. The Singularity container is planned to be used by an astronomy graduate student on Bridges large memory nodes to process a large dataset as part of an upcoming XSEDE allocation request. Additionally, our summer REU student has completed the summer project, but will be continuing to work on the project through the academic year with the goal of working towards a new FRB detection. We will be continuing to make improvements to the FOF algorithm in the FRB_pipeline and complete the comparison with Single Pulse Search method in PRESTO with the modulation index calculation. A new dataset that has not been processed yet will then be explored with the FOF algorithm in search of FRB detections.

Use Case 5: Water Resource Management Using OpenMORDM

In the last quarter, the Cornell Aristotle team modified an existing Docker container to create a VM for initial tests with WaterPaths, which automatically clones and compiles WaterPaths and Borg and downloads all necessary data files for the Pat Reed Research Group. On this test VM, WaterPaths was successfully adapted to run on Red Cloud, after some issues with binary data formats were resolved, which allowed for standalone WaterPaths simulations and small-scale optimization runs with four MPI processes (one master and three slaves). Lastly, the Aristotle staff and the Reed group defined the guidelines for the next steps, which will include timings and MPI profiling and on whose setup the Aristotle team is currently working:

1. Using the same size VM with 28 cores, set up 3 Docker containers to use 8 cores each (total of 24 cores)
2. Set up a small cluster of 3 VMs, each with 8 cores, with a single Docker container on each that uses all 8 cores
3. Set up a larger cluster of 8 VMs, each with 28 cores, with a single Docker container on each VM

The timings and MPI profiling will allow for at least two publications, one focused on WaterPaths and the other on the cloud itself.

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

The Angela Douglas Lab and Aristotle support team focused on developing new algorithms and workflows for analyzing the impact of the order of arrival of a microbe (priority effects) on community structure and function. We developed a new method, Semi Dynamic SteadyCom, to investigate priority effects among Drosophila gut microbes. Semi Dynamic SteadyCom employs SteadyCom with discrete time steps. At each step, flux uptake bounds are restricted or relaxed based on consumption or production patterns in the prior time step. We have also developed a modeling framework with strong static typing, named hCOBRA, in the Haskell programming language. hCOBRA extends the COBRA Toolbox for MATLAB with the aim of making large scale simulations both more reliable and more scalable. Our development and implementation of Semi Dynamic SteadyCom and hCOBRA has facilitated navigating the complexity of the metabolic modeling and recursive structure of simulations associated with priority effects. We are currently performing our final round of simulations and will analyze data from these simulations for the preparation of a manuscript.

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

Citrus Frost Prevention (Lindcove Research and Extension Center, Exeter, CA):
The Citrus Under Protective Screens (CUPS) project is stalled due to COVID-related construction delays and infection precautions. The science team has been waiting to be cleared to visit the site to assess how to make progress, but the town where the site is located has had a severe outbreak so no access has been possible. Similarly, the completion of the facility (which is necessary for the instrumentation installation to begin) has been paused pending a reduction in cases locally. However, under stay-at-home work rules, the team has begun to develop the hybrid machine-learning and CFD model that will be used to support CUPS spraying operations and frost prevention. The team has also been working with an Aristotle REU

NSF
#1541215

(Kerem Celik) to develop a telemetry visualizer that works either as a local tool or as a service hosted in Aristotle.

Edible Campus and Food Security (UCSB):
The Edible Campus farm (located on UCSB campus grounds) remains closed to student access. The science team obtained permission for a brief assessment visit and the instrumentation infrastructure remains functional; however, growing efforts remain stalled pending the safe return of the student growers and local farming consultants.

Sedgwick Reserve Ecology (Santa Ynez, CA)
UCSB is slowly enabling limited research activities on campus but as yet has not lifted restrictions on research at the research preserves (including Sedgwick) that it operates. Faculty (but not students) can get clearance to visit in order to assess the state of on-going experimental projects.

5.0 Community Outreach and Education

5.1 Community Outreach

- The Cornell team is preparing to participate in Indiana's Jetstream 2 project and is currently reviewing hardware specifications. The regional Jetstream 2 system deployed at Cornell will consist of 1,024 computer cores and 869TB storage. This "New York zone" will be used to explore federation of clouds and to make OpenStack enhancements that will be shared with the rest of the project team and disseminated to the broader research community. Cornell will draw on our Aristotle experiences to create campus software so that campuses can set up their own clouds

5.2 Education

- 5 undergraduate students (4 at Cornell, 1 at UCSB) participated in Aristotle cloud and accelerated GPU research projects thanks to funding from the NSF REU program. Their research experiences and lessons learned were reported in a September 9, 2020 news story featured in *HPCwire*: https://www.hpcwire.com/off-the-wire/cornell-students-immersed-in-latest-cloud-technologies-thanks-to-nsf-research-experiences-for-undergraduates-program/
- At UB, Mohammed Zia used the Aristotle cloud service to host JupyterHub with the nbgrader module enabled on an Ubuntu 18.04 image for his *Programming and Database Fundamentals for Data Scientists (EAS 503)* class. The nbgrader module allows him to create and grade programming assignments in JupyterHub, a streamlined method that enables more frequent distribution of assignments (increased from 4 to 13). Aristotle is cost-effective because Zia can choose the instance type based on the expected class size. There are usually 120 students in the Fall semester and 50 students in the Spring semester. Zia selects the instance type he needs based on the class size, and after the semester is over, deletes the instance. Ease of deployment and the ability to handle all of the student's interactive Python needs in one place is also a plus.

NSF
#1541215