ARISTOTLE
CLOUD FEDERATION

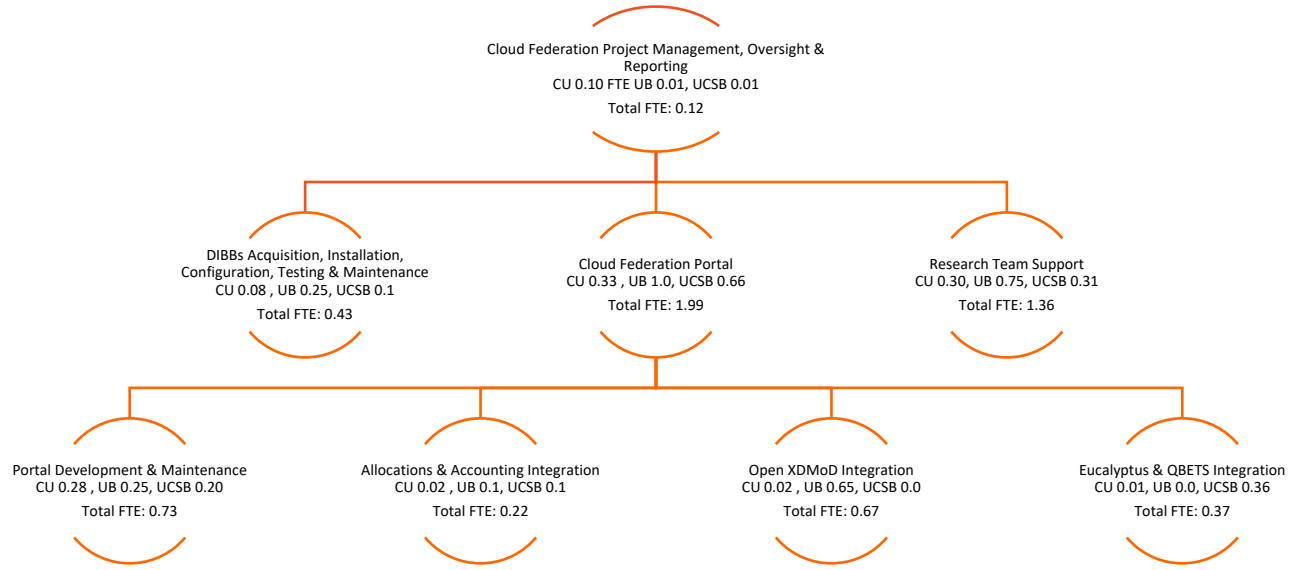
CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation

Program Year 5: Quarterly Report 3

6/29/2020

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

This is the Program Year 5: 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

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/Partner Meetings
- Cornell's Aristotle team hosted the CIO, research computing director, and 5 IT staff members from Binghamton University for briefings on our experiences moving to the cloud, the containerization of research applications, and the federated cloud model.
- Aristotle's infrastructure team had discussions with Red Hat's technical staff on how to move to OpenStack v.16. We're optimistic that Red Hat will help us design the best path forward.
- At NSF's request, Aristotle PI David Lifka participated in a CISE call with 6 other CIOs to provide opinions on restarting research activities at U.S. campuses. Among the ideas discussed were the recommendations that NSF provide incentives for faculty to use centralized services as opposed to each faculty building their own system and the expansion of oversubscribed XSEDE resources.

1.5 Project Planning Meetings/Status Calls

4/7/2020 progress call:
- Cornell secured an AWS allocation and co-PI Wolski will be using it to develop the DrAFTS 2.0 AWS cost comparison tool. UCSB's Edible Campus, Sedgwick Reserve camera trap, and Citrus Under Protective Screen (CUPS) projects are on hold due to the COVID-19. Wolski submitted a preproposal to the Citrus Research Board (CRB) to augment the CUPS project and was encouraged to submit a full proposal. If successful, the CRB award will fund additional monitoring and analytics all of which will be Aristotle-based.
- UB's Aristotle team is working with use case scientist Dominik Roesch to get his financial framework working on UB's cluster in the cloud.
- When completed, the Aristotle portal will include the DrAFTS 2.0 AWS cost comparison tool with availability to the broader research community. It will compare TOP500 benchmarks on all AWS instances and, eventually, all Aristotle instances.
- Our proposal for 6 summer 2020 Aristotle REUs was submitted and is pending NSF approval. We plan to share lessons learned on the virtual management and productivity of remote-based REUs.
- Cornell's "cloud first" strategy, previously implemented, enabled a seamless transition for researchers, students, and staff when the campus closed due to COVID-19. There was no negative impact on research computing or data access.
- Cornell took its oldest production node off line and made it the meta data server for the Ceph cluster.
- A new physics use case at UB would like to use containers on UB's cloud. UB prefers to use Magnum for container functionality rather than standing up Kubernetes or Red Hat OpenShift.

#1541215

Cornell will share the documentation they've developed on installing Magnum. The Red Hat OS doesn't include Magnum; the installation is a hack.

- UB has NVIDIA GPUs up and working well. One limitation is you can't offer multiple GPU types per hypervisor; you can only pick one on the OS version UB is currently using. UB will be able to separate GPUs with the newer OS. Users are happy with the GPUs so far.
- Cornell continues work on containerizing MPI-based applications.

4/8/2020 progress call:

- Dartmouth is scrambling to provide online classwork due to COVID-19 which is impacting their desire to deploy a cloud system with the Aristotle team and, possibly, the XSEDE CRI team.
- Cornell is working on standing up an MPI cluster at UB for potentially running WRF site-to-site.
- Angela Douglas is now using Windows containers to run simulations to improve the understanding of wind climate variability with a focus on wind energy and wind turbine applications.

4/21/2020 status call:

- UB is assessing which type of GPUs provide the biggest bang for the buck.
- David Doermann, an AI expert at UB, would like to have a submission portal with the Lake Effect cloud on the front end and HPC clusters on the back end and a SLURM interface. The goal is a pipeline to bridge from cloud to HPC. Doermann is training ML models with TensorFlow.
- UCSB's onsite research remains on hold due to COVID-19. However, we are now taking measurements from individual farms of the amount of moisture plants give off on a daily basis. We will use this data to create the optimal watering cycle (i.e., if a plant is giving off a lot of water due to evaporation transportation, we will provide more water on that plant). The correlation analysis between electrical power and evaporation transportation will take place on Aristotle.
- DrAFTS 2.0 AWS cost comparison analysis benchmarking scripting is underway. When the tool is completed, it will be available for download from GitHub with installation scripts and on the Aristotle portal as an on-demand tool for the research community. Our plan is to use the DrAFTS tool to run the TOP500 benchmark on all AWS instances in each AZ and compare performance to price. We plan to perform theses analyses on the Aristotle clouds as well.
- The Cornell Aristotle team received 40 REU applications; offers will go out this week.

5/5/2020 status call:

- Two women and two men REUs were selected by Cornell.
- At UCSB, an REU student built an edge visualizer for meteorological data that we're collecting; this tool provides real-time data graphs. We plan to demo this new capability in June. This REU student has decided to join UCSB's CS PhD program! Aristotle REUs have been a great recruiting tool for UCSB's PhD program.
- We are currently working on the packaging of DrAFTS 2.0. The research community will be able to choose any AWS or Aristotle instance type and make comparisons (via a ranking) of performance and price. An REU student will work on benchmarking DrAFTS 2.0 this summer and integration of the tool into the Aristotle portal.
- UB is working on the next version of XDMoD; they plan to release v. 9.0 at PEARC20.

5/19/2020 status call:

- The Aristotle portal team is discussing how to complete the Aristotle portal template and share it with the rest of the research community.

#1541215

- UI work continues on DrAFTS 2.0 to make it easier to use. The performance ranking is done; we will add the price ranking feature next. It will then be skinned with the Aristotle style sheet and added to the portal for public access.
- We will have to rebuild our clouds to get to OpenStack v. 9.0. Red Hat is providing some templates that should be helpful particularly since all 3 sites will be working together on the upgrade and actively sharing lessons learned as they go. We are also exploring the cost for ongoing support with Red Hat after the grant ends.
- We are investigating which future workshops we wish to participate in in order to share our experiences running MPI-based applications in containers.
- Finance researchers at UB led by Dominick Roesch and the Aristotle team are working on getting the finance workflow running on UB's HPC cluster. The financial database is running on UB's Aristotle cloud; researchers want to grab the database from anywhere and use it on their local workflows.
- UCSB plans to install video cameras at the Sedgwick Reserve and their REU student is working on a machine learning app to capture and process the data. Their Edible Campus project has the surveillance video up and running, and students in Wolski's cloud computing class are building essential real-time human detection software that will provide alerts when people are on the campus farm. Edge clouds are running the machine learning applications but the instrumentation for the farm needs to get operational.

5/27/2020 progress call:

- We discussed writing additional papers on what it really takes to use containers to get portability and our successes doing so. The focus will be on providing portability to any platform (cloud, HPC, etc. and Kubernetes, or not) for legacy MPI codes.

6/2/2020 status call:

- We conducted an orientation session for our new REUs, provided them with Slack channels, and plan to have weekly Zoom meetings to ensure that they stay engaged and feel like they are a vital member of the Aristotle team. This summer's REU projects will include cloud application enablement, accelerated computing with GPUs, building a subsystem to support the monitoring of IoT sensors, building new tools for digital agriculture such as an REU-created real-time meteorological data access tool, and enabling camera trap image analysis with video.
- Cornell and UCSB expect to get some software reuse between the Aristotle and the SCIMMA projects (e.g., a code base that wraps Kafka as a publication hub system for astronomy, a new monitoring system, etc.).
- Work continues on packaging DrAFTS 2.0 for research community use. It will enable researchers to find the sweet spot when deciding which AWS or Aristotle instance to run on for the best price/performance. It will sort all instance types by cost per hour and allow researchers to comparison shop and get the most bang for the buck. The capabilities of this pricing engine should surpass what is currently available in terms of features and accuracy.
- UB has many application kernels in the XDMoD framework that they want run with DrAFTS 2.0.
- Dartmouth is reinstalling their cloud software.
- Cornell added 6 new Ceph nodes. We plan to add 16 more T4 GPUs in the coming weeks.
- UB plans to add more compute and storage to their Ceph cluster.

NSF
#1541215

6/9/2020 progress call:

- REU student Jeff Lantz started working with Docker and Singularity and is looking at Terraform and Kubernetes. His goal is to deploy a Kubernetes multi-VM MPI cluster starting with an OSU benchmark container and then using WRF or some other MPI application.
- We had discussions with NVIDA regarding a GPU version of WRF. The Pryor group is currently using just the binary v. 3.8.1 because a custom framework on top would need source access. The other alternative is a fee-based version from TempoQuest but the Pryor group considers it too expensive. TempoQuest appears to be a leader in WRF; they build the free community binary and ship it out. NVIDIA's WRF effort is community based so you can talk to them, sign an NDA, and get source. Unfortunately, using 3.7.1 is not compatible with the modules that the Pryor group wants. They have an HPC version they have to run on Summit.
- Wolski is about to start a project using CFD to model turbulence in the Citrus Under Protective Screening (CUPS) orchard. He'll likely use OpenFOAM's CFD Toolbox which has MPI support. CA and FL are trying to figure how to use very fine mesh screens and screen houses to grow citrus. They can't control the humidity or temperature so studies looking at how to model airflow inside this space is critical. They may use an ARM NVIDIA system on a chip and Nautilus which is part of Larry Smarr's fabric Kubernetes clusters that attach to GPUs like a batch system. NAMD is another possibility for the REU student to work on.
- We plan to add a table to the Aristotle portal by Sept. 4th with some sample scientific workflows and a cost-performance analysis of multi-VM workflows, and possibly, Kubernetes.
- Longer term, we are planning multiple papers: one on how to get equivalent performance (hopefully) from a containerized MPI program in a public cloud compared to XSEDE (without waiting in a batch queue)—how hard is it and what does it cost? A second paper would focus on multiple cloud and multiple MPI-based apps—is there a performance penalty? our tools and workflows make it easy (with or without Kubernetes), is there is a cost-efficient way to deploy it? etc.

6/16/2020

- Our REU students gave summaries on what they've been working on thus far during our weekly Zoom REU call. The atmospheric sciences REU student discovered that you can spend half your time making the data analyzable due to irregular sampling and other factors. Other REUs are making progress with application containerization projects and looking at FRB algorithms and the impact of false positives and radio frequency interference. The accelerated computing project was too difficult for an undergraduate to build, so we will build an OpenFOAM installation for the student who will then focus on the Dev side of DevOps.
- UCSB kicked off the CFD project to model the turbulent airflows inside screen houses. This has never been done before and is generating excitement in the ag community. An REU student is building a large eddy simulation CFD using OpenFOAM. A mathematician collaborator in LA is excited about this project and is writing a collaborative NSF grant to secure additional funding. We would like to find out if we can use the turbulent flow model to not only model airflows, but to automatically spray citrus using robots. Currently, we don't know how much to spray. Also, since we can't use a wind machine to do frost prevention in the screened houses, we want to use CFD to see if we can model heat transfer out of the canopy. This will be done initially on an X86 architecture at UCSB and then Cornell will replicate the computation using their GPU infrastructure.
- A follow-up infrastructure team call with Red Hat will happen this week to help us with our TripleO builds and to review available templates.

#1541215

- UCSB plans to purchase more compute nodes and Cornell is looking at upgrading and adding to their networking.
- The portal team plans to improve the new account interface to make it easier to add students to a project.
- DrAFTS 2.0 TOP500 benchmarking is now underway for all AWS NE regions.
- Work on the container performance table for inclusion in the Aristotle portal continues and will be completed by the end of summer.

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

2.1 Hardware Acquisition

- Cornell purchased 6 Dell servers to add to their Ceph storage pool. UCSB and UB had no acquisitions this quarter.

2.2 Installation, Configuration, and Testing

- Cornell installed the new Dell servers. This increased their Aristotle Ceph storage capacity to 1.6PBs.
- UCSB and UB performed routine cloud maintenance this quarter.
- Dartmouth made progress rebuilding their OpenStack system; they plan to connect to the federation next 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	**256	356
RAM/Core	8GB	up to 8GB	9GB Dell, 10GB HPE
Storage	Ceph (1644TB)	Ceph (768TB)	Ceph (528TB)

#1541215

10gb Interconnect	Yes	Yes	Yes
Largest Instance Type	28core/240GB 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) = 2424	* 616 additional cores augmenting the existing Red Cloud (1252 total cores).	** 256 additional cores augmenting the existing Lake Effect Cloud (600 total cores).	***356 cores in UCSB Aristotle cloud (572 total cores, Aristotle is separate from UCSB campus cloud)

2.5 Tools

- Red Hat – Cornell, UCSB, and UB infrastructure teams meet with Red Hat's OpenStack technical team to outline an OpenStack upgrade plan.

  Cornell and UB will build new OpenStack environments and migrate existing ones (there is no upgrade path from their current installations). UCSB will be able to upgrade because they installed using TripleO. Red Hat will provide templates tailored to each site to assist in the new builds.

  Cornell's Steven Lee will take a Red Hat TripleO class to facilitate a smooth transition at Cornell.

3.0 Cloud Federation Portal Report

Content updates to the project portal are ongoing (https://federatedcloud.org).

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.

#1541215

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	Display usage and CPU hours by account or project on the portal.	Automate project (account) creation by researcher, via the portal.	Report on usage by account, if the researcher has multiple funding

#1541215

collection across the federation. Design database schema for Users, Projects and collections of CPU usage and Storage Usage of the federated cloud.	Integration hooks for user and project creation/deletion and synchronization across sites. Note: due to OpenStack move, account creation across sites is delayed.		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 DrAFTS into the Portal

DrAFTS 2.0 remains at the prototype stage although it has been tested with the UCSB campus cloud. The staff programmer on the project is developing the needed deployment and maintenance scripts. However, the OpenStack CLI is no longer functional due to a software upgrade. Aristotle staff are working to restore the CLI functionality that is necessary to transition DrAFTs 2.0 to production operation.

In addition, while waiting for OpenStack repairs, the prototype is being integrated with the portal for test purposes so that when the system is operational, functionality will be immediately available.

3.3 Integrating Open XDMoD into the Portal

3.3.1 Application Kernels Containerization in the Cloud

No work was performed on application kernel containerization this quarter.

3.3.2 XDMoD Cloud Integration

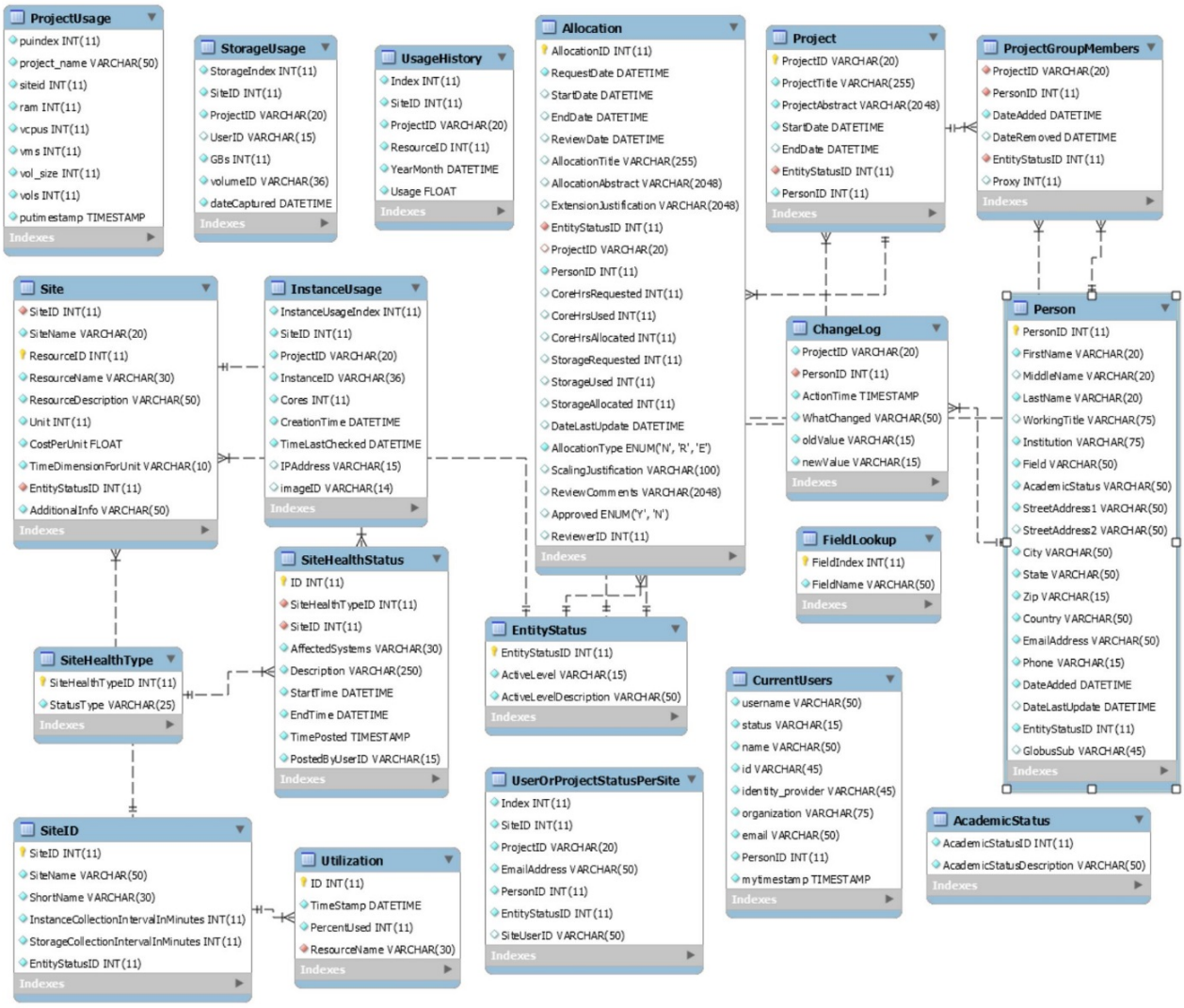
No work was performed on XDMoD integration this quarter.

3.4 Allocations & Accounting

A draft usage report that shows core usage per site/per month was reviewed by the project directors and, after minor improvements, implemented. The federation's allocations and accounting systems is completed.

No changes were made to the database schema this quarter:

#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

UB Professor Varun Chandola and collaborators created a system to support the Lower Great Lakes Resiliency Research Group. It runs on top of Aristotle (https://www.community-outstepsteps.org) and allows community participants in and around the Lower Great Lakes (universities, NGOs, and city/state governmental organizations) to interact and share information and data using a built-in, open source communication platform called RocketChat (https://rocket.chat/). The next step is to integrate the WebGlobe framework into this platform to allow users to share geospatial datasets and workflows.

#1541215

Use Case 2: Global Market Efficiency Impact

UB professor Dominik Roesch with UB professor Jonathan Brogaard and University of Utah professor Matthew Ringgenberg wrote a new paper that investigates whether the closure of the floor of the New York Stock Exchange due to COVID-19 affected global financial markets. The paper—"Does the Trading Floor Matter?"—is currently being revised and credits the Aristotle grant. In addition, a new UB CS student is using the Aristotle framework for her research in collaboration with Roesch.

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

Précis objectives of Cornell professor Sara C. Pryor's and post doctorate associate Tristan J. Shepherd's 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).

This quarter's activities have included preparation of an instance and data sets for use by the REU student and supervision of that student. Unfortunately, our WRF simulations exhibited slow compute speeds on the single node VM and due to pressing commitments on Cornell staff time no progress was made this quarter generating a multi-node instance for use with WRF. Thus, our activities have exclusively focused on analysis of our previous simulations in support of:

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.

Activities planned for next quarter:

- Complete transitioning from conducting WRF simulations on Aristotle to XSEDE resources in anticipation of the conclusion of the Aristotle project.
- Finalizing analysis of previously conducted WRF simulations.

#1541215

- Continue to support and guide research performed by the NSF REU student.

Journal manuscripts finalized this quarter:

- Pryor S.C., Shepherd T.J., Volker P., Hahmann A.N. and Barthelmie R.J. (2020). Diagnosing systematic difference in predicted wind turbine array-array interactions. *Journal of Physics Conference Series: Special issue from The Science of Making Torque from Wind* (in press).
- Aird J., Barthelmie R.J., Shepherd T.J. and Pryor S.C. (2020). WRF-simulated springtime low-level jets over Iowa: Implications for wind energy. *Journal of Physics Conference Series: Special issue from The Science of Making Torque from Wind* (in press).
- Barthelmie R.J., Shepherd T.J. and Pryor S.C. (2020). Increasing turbine dimensions: Impact on shear and power. *Journal of Physics Conference Series: Special issue from The Science of Making Torque from Wind* (in press).

Use Case 4: Transient Detection in Radio Astronomy Search Data

The Radio Transient Detection Use Case has been focused on combining the pipeline components developed for Pulsar and other transient detections (that are not FRBs) with the FRB_pipeline in a single container that can be deployed either on the cloud (Docker) or on an XSEDE HPC resource (Singularity), and on improving the performance of the components of both of these pipelines in preparation for full scale runs of processing large datasets on XSEDE. Pre-existing optimizations of the PRESTO code exist, and we are exploring leveraging these to improve the overall performance of the pipeline, as well as improvements to pipeline components through parallelization and other means. We are exploring a few options of where to run through an XSEDE Campus Champions allocation, and are intending to pursue a Startup Allocation, and then a full allocation request. Due to the large data sizes to be processed, and how current methods for processing these datasets work, we may need to pursue access to large memory nodes or queues. Additionally, our Radio Astronomy REU student this summer is working on improving and validating the FOF algorithm in the FRB_pipeline with plans to run a comparison on a full dataset of the FOF algorithm to the Single Pulse Search method in PRESTO with modulation index calculation to sort pulse candidates.

Use Case 5: Water Resource Management Using OpenMORDM

Pete Vaillencourt has installed MPI profiling tools in Docker and Bennett Wineholt has created a large Virtual Machine with 28 CPUs and 224GB memory for Bernardo Trinidade to use.

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

During this quarter we have submitted our manuscript—"The Predicted Metabolic Function of the Gut Microbiota of Drosophila melanogaster"—to *Cell Reports* (currently in review). We have extended our results from that paper to investigate priority effects, which are the order-dependent effects resulting from introduction of species within a community. Our goal for this follow-up study is to identify the nutrient conditions that make Lactobacilli, widely recognized for their potential as a probiotic, colonize established microbial communities better. An additional goal for this project is to identify specific nutrients than can be added to microbial communities to make specific Lactobacilli species dominant. While using much of the prior code and infrastructure developed during the course of the Aristotle project, the increase in the number of simulations, as well as the means of running them, has required some additional development efforts. This work is ongoing, and our initial exploration suggests several avenues of potential biological

interest as well as some technically-interesting cloud-computing results related to caching scientific workflow results for improved performance.

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

Citrus Frost Prevention (Lindcove Research and Extension Center, Exeter, CA):
CUPS construction has been slowed by COVID-19 labor safety practices but is still under way. We met virtually with the CUPS manager at Lindcove and the PI of the USDA grant from the University of Florida to discuss Aristotle analysis of under-screen meteorological data. Apparently little in the way of analytics has been applied to understanding the growing environment under CUPS. The screen attenuates 20% of the UV solar radiation, but it is unclear how much wind restriction is introduced, how much additional humidity, and what effect it has on the screen house temperature. The University of Florida CUPS is hydroponic making the California installation the first instrumented CUPS with trees growing in the soil. UCSB science researchers will be leading the analytics development and deployment using Aristotle for CUPS.

Edible Campus (UCSB):
The Edible Campus is shut down for student activity except for the emergency maintenance necessary to keep the crops alive. The UCSB science team installed surveillance cameras in December 2019 to help train students in farm operations by allowing experts to observe them without visiting. However, with COVID-19 the concern is that students may be putting themselves at risk by visiting for non-essential activities. The Aristotle PI taught a cloud computing class during which time several students developed video analysis applications to help identify when people were visiting the farm without an essential farming activity to perform. While no such incidents were recorded, the software was able to screen thousands of false positive motion detections caused by passing vehicles, spider webs waving in front of the motion detectors, and blowing drizzle.

Sedgwick Reserve (Santa Ynez, CA)
The Sedgwick camera trap operations are stalled because of COVID-19 restrictions on visitations to the site. UCSB appears to be lifting those restrictions in the coming weeks. When it is safe, the UCSB science team will visit the site and assess the maintenance that will be necessary to reinitiate the camera trap observations.

5.0 Community Outreach and Education

5.1 Community Outreach

- A new Aristotle paper, "Reproducible and Portable Workflows for Scientific Computing and HPC in the Cloud," was accepted by PEARC20: https://arxiv.org/pdf/2006.05016.pdf
- We will be producing a white paper that compares different ways of containerizing applications. We are also planning an SC workshop paper that will provide performance comparisons of WRF.
- The Cornell team will be participating in Indiana's Jetstream 2 project. The regional system deployed at Cornell will consist of 1,024 computer cores and 869TB storage. The "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 campuses can set up

NSF
#1541215

their own clouds: https://www.cac.cornell.edu/about/news/200603.aspx

5.2 Education

- Co-PI Wolski is currently teaching a cloud computing class and some of the student projects are Aristotle-based.
- The Aristotle team will be providing a PEARC20 tutorial titled "Deep Dive into Constructing Containers for Scientific Computing and Gateways."
- Use case scientists and the Aristotle team are providing 6 REUs with the information and skills they need to succeed in their remote summer REU experience. Last summer's UCSB REU has committed to their CS PhD program.
- *Cornell Cloud REU Project*: a student is focused on exploring Kubernetes clusters with MPI capabilities in public clouds. A Docker container will be deployed with benchmarks to gauge the effectiveness of Kubernetes in supporting HPC in the cloud. The use of Kubernetes will also be compared to our previously explored method of automatically deploying multi-VM MPI clusters using Terraform and Ansible. Deploying Kubernetes using cloud vendor managed services (such as GKE) as well as scripted deployment through Terraform will also be evaluated and reported on.
- *Cornell Accelerated REU Computing Project*: The goal is to facilitate GPU computation for our science partners. We have demonstrated GPU usage with a basic machine learning framework program and identified a science team interested in accelerating airflow simulation codes. Specifically, we have used Aristotle virtual machine resources to install CUDA 11 and 10.1 as well as libraries using the related NVIDIA GPU drivers. GPU-enabled TensorFlow 2.2 was installed using Anaconda and Jupyter Notebook servers were deployed using community Docker builds. Multiple interdependencies between available system packages, kernel module loading, hardware drivers, and software libraries make the installation process a nontrivial challenge for the average scientific user. We intend to demonstrate a few more common libraries and then provide a pre-built virtual machine image and accompanying user documentation to facilitate user productivity using the federation's GPU resources. We have identified a testing research group to use this new GPU-enabled server image to accelerate their intended airflow simulation codes and will continue to work with them to prototype a proof-of-concept simulation run this summer.

#1541215