My Desktop
Prepare & Submit Proposals
Prepare Proposals in FastLane
New! Prepare Proposals (Limited proposal types)
Proposal Status
Awards & Reporting
Notifications & Requests
Project Reports
Submit Images/Videos
Award Functions
Manage Financials
Program Income Reporting
Grantee Cash Management Section Contacts
Administration
Lookup NSF ID

Preview of Award 1541215 - Annual Project Report

Cover |
Accomplishments |
Products |
Participants/Organizations |
Impacts |
Changes/Problems

Cover

Federal Agency and Organization Element to Which Report is Submitted:	4900
Federal Grant or Other Identifying Number Assigned by Agency:	1541215
Project Title:	CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation
PD/PI Name:	David A Lifka, Principal Investigator Thomas R Furlani, Co-Principal Investigator Richard Wolski, Co-Principal Investigator
Recipient Organization:	Cornell University
Project/Grant Period:	10/01/2015 - 09/30/2020
Reporting Period:	10/01/2017 - 09/30/2018
Submitting Official (if other than PD\PI):	N/A
Submission Date:	N/A
Signature of Submitting Official (signature shall be submitted in accordance with agency specific instructions)	N/A

Accomplishments

*** What are the major goals of the project?**

The major goals established at the launch of the 5-year Aristotle DIBBs project are:

- Implement a scalable and sustainable multi-institutional cyberinfrastructure cloud federation model that provides data analysis building blocks in support of multiple research disciplines requiring flexible workflows and analysis tools for large-scale data sets. Federation sites are Cornell University (CU), University at Buffalo (UB), and University of California, Santa Barbara (UCSB).
- Support seven strategic science use cases from intentionally diverse disciplines (earth and atmospheric science, finance, chemistry, astronomy, civil engineering, genomics, and agriculture) to demonstrate the potential of a federated cloud as a campus bridging paradigm. Explore data analysis techniques and their applicability to different disciplines. Document tools, workflows, challenges, and best practices for each use case.
- Encourage and reward data analysis resource sharing with a new allocations and accounting model that provides a fair exchange mechanism for resource access between and across multiple institutions. Develop and build a new tool for cloud metrics into Open XDMoD that includes DrAFTS (Durability Agreement From Time Series) statistics to make online forecasts of future performance and allocations levels available to users.

During Aristotle's eighteen-month independent project review at the National Science Foundation, the review panel concluded that the "Awardee and colleagues need to be highly commended for their progress thus far" and recommended continued funding. With regards to project goals, the panel stated that Aristotle has the potential to:

- serve as a model for providing cloud resources locally while providing elasticity and access to resources such as software, datasets, or hardware not available locally,
- provide relevant metrics that facilitate the exchange of resources among a federation of institutions (via the incorporation/further development of Open XDMoD).

The panel also noted that Aristotle isn't just "another cloud" or a specific cloud solution, but rather an exchange and an enabling technology for researchers to achieve results faster.

*** What was accomplished under these goals (you must provide information for at least one of the 4 categories below)?**

Major Activities: **Architect and install PY3 DIBBs infrastructure/storage assets at 3 federation sites**

Specific Objectives:

- Installed cloud infrastructure at Cornell University, University at Buffalo, and University of California, Santa Barbara with DIBBs PY3 funding. Cornell installed infrastructure servers and a network switch for their Red Cloud OpenStack cloud; UB installed a Ceph storage cluster for their Lake Effect OpenStack cloud, and UCSB installed OpenStack node controllers and network switches.
- Installed a Foreman/Puppet system at Cornell to automate all future Aristotle OpenStack deployments.
- Onboarded friendly users to UB's OpenStack cloud, created a web service to report OpenStack usage, and ordered SSD drives to add to UB Ceph cluster (will reconfigure OSDs to maximize performance).
- Worked on improving the redundancy of UCSB Ceph cluster connectivity, configured UCSB networks, and tested OpenStack-Ansible automated installations for OpenStack deployments.
- Configured the web consoles for the new OpenStack clouds to support Globus Single Sign-On.

- Worked on integrating user identity and access management across the portal and the OpenStack clouds.

Develop portal code to allow Globus authentication by users

- Implemented Globus Auth on the portal which will be used to provide Single Sign-On to users across multiple sites and facilitate access to federated Aristotle data.

Develop Phase 3 portal content

- Updated portal content on an ongoing basis to reflect the federation's latest capabilities, publications, science news, use case accomplishments, plans, and products, and user guide enhancements.
- Implemented software to verify the data ingestion API is running for the Aristotle usage graph.
- Upgraded, refined, and added functionality to the accounting and allocations database and tables (added AcademicStatus, User History, UserOrProjectBySite, etc.).
- Created multiple new Stored Procedures so Aristotle users and projects can be added to the portal.

Implement Open XDMoD cloud data collection across the federation

- Developed beta version of federated Open XDMoD which will support the collection and aggregation of data from individually managed sites into a single federated instance of Open XDMoD in order to display federation-wide metrics.

Integrate Metrics and Predictions into Open XDMoD and associated data collection for portal

- Integrated initial cloud metrics (number of instances running/started/ended, number of cores utilized, CPU hours per instance/total, and wallhours per instance/total, etc.) into Open XDMoD and tested and refined metrics for upcoming Open XDMoD 8.0 release.
- Used DrAFTS to audit and analyze AWS changes to spot instance pricing and reliability.
- Continued to use DrAFTS as an AWS pricing tool but without the reliability guarantees.

Support 7 science use case teams with current cloud resources while implementing plans for federated cloud usage

1. **A Cloud-Based Framework for Visualization & Analysis of Big Geo Data** (Varun Chandola, UB). Refined webGlobe's integrated visualization and analysis capabilities that support a variety of data formats prevalent in the climate and earth science communities. Continued development of an Energy-Water Knowledge Discovery Framework portal using webGlobe technology currently running on the Aristotle cloud. Performed comparative evaluations of various machine learning methods on Aristotle resources to better understand the Energy-Water nexus. Shipped a version of webGlobe to collaborators at ORNL to support their research activities in the area of climate data analytics.
2. **Global Market Efficiency Impact** (Dominik Roesch, UB). Created VMs in collaboration with Varun Chandola and Jialiang Jiang that share the same image and hold all the finance data; they are executing well. Demonstrated

the framework to PhD students so they can analyze the data independently on Aristotle. Launched new investigations, i.e., how price deviations (market inefficiencies) affect liquidity (the ease at which you can buy and sell), how the recent increase in tick-size (the minimum price movement of US stocks) affects liquidity, etc.

3. **High Fidelity Modeling and Analytics for Improved Understanding of Climate** (Sara C. Pryor, Cornell). Analyzed high-resolution numerical simulations of the effects of wind turbines (WT) on regional climate. Completed simulations to test the sensitivity of the climate impacts to the precise description of the WT aerodynamics (the extraction of momentum and introduction of turbulence behind the turbine rotor). Analyzed long-term simulations with the WRF model to examine inter-annual variability of annual mean wind speeds at/near typical wind turbine hub-heights, and applied the power curve of the most commonly deployed WT to post-process the 10-minute wind speed output into estimated annual energy production (goal is to create a more robust prediction of the value of wind energy projects; analyses rendered possible by mounting a 100TB hard drive to an Aristotle instance).
4. **Transient Detection in Radio Astronomy Search Data** (James Cordes, Cornell). Developed a new containerized solution that features a full suite of all relevant pulsar-processing software, Python Conda distributions, the core PRESTO pulsar search package to perform transient search using the Spitler modulation index method, and the decimation code. Implemented an improved reproduction of the PALFA2 pipeline for detecting single pulse candidates that may be a Fast Radio Burst (FRB) source; this pipeline includes PRESTO functionality, modulation index calculation, parameter customization, and the production of graphic data output. Built a new flexible framework for running radio astronomy searches: data are read from their native format into NumPy Arrays and the pipeline's routines are selected from a configuration file and includes a friends-of-friends search, and also allows the running of the pipeline of Laura Spitler who discovered FRB 1211102.
5. **Water Resource Management Using OpenMORDM** (Patrick Reed, Cornell). Containerized and published build scripts for the Lake Problem code and successfully ran it in MPI across multiple cloud virtual machines using Docker; this will allow distributed scientific software to be executed faster at cloud scale, in existing institutional clouds, XSEDE resources, and public provider clouds.
6. **Mapping Transcriptome Data to Metabolic Models of Gut Microbiota** (Angela Douglas, Cornell). Integrated the SteadyCom computational framework to model multi-species metabolic interaction in the insect gut. Performed SteadyCom on test models in Aristotle using a newly configured Windows VM with MATLAB and COBRA Toolbox. Built a Linux VM with Docker, MATLAB, and Gurobi to containerize the simulation environment. Performed a Flux Variability Analysis on the four microbes composing the *Drosophila* gut microbiome. Constructed multi-species metabolic networks in silico of 2, 3, 4 and 5 bacterial species, quantified predicted metabolite fluxes, and wrote a MATLAB function to restore exchange constraints to a multi-species model created from a single species. Verified multi-species models behave as expected with standard flux algorithms.

7. **Multi-Sourced Data Analytics to Improve Food Production and Security** (Kate McCurdy, Sedgwick Reserve; Elizabeth Grafton-Cardwell, UC Lindcove Research & Extension Center; Chandra Krintz, UCSB SmartFarm with Cal Poly, San Louis Obispo and Fresno State). Installed hardware and software to instrument almond trees in a Fresno, CA test orchard to see how much water can be saved by irrigating the different sides of the root stock in proportion to its dryness. Develop innovative analytics with a minimally invasive instrumentation footprint for an Exeter, CA citrus test orchard to get highly accurate temperature readings at night when frost could form.

Significant Results:

Key outcomes or
Other achievements:

*** What opportunities for training and professional development has the project provided?**

Cross-Training & Knowledge Sharing

Expertise was shared between sites every two weeks on Aristotle team conference calls to ensure timely cross-training and knowledge sharing, and in-depth follow-up calls occurred to solve specific technology implementation issues and to share lessons learned.

Cross-site discussions and collaborations between use case scientists facilitated the sharing of research tools and cloud computing techniques, e.g., big geo data use case scientist Varun Chandola and PhD student Jialiang Jiang helped financial market use case scientist Dominik Roesche develop two finance VMs that share the same image).

A train-the-trainer approach has been used as a training multiplier, i.e., the training of one use case team member was documented in order to facilitate the training of the entire research group.

Knowledge gained from training events such as the 2017 AWS Global Summit were beneficial to the federation as a whole.

Undergraduate & Graduate Student Development

Five undergraduate REU students made valuable contributions to Aristotle science use cases at Cornell and UCSB, and gained valuable domain-specific knowledge and first-hand experience using clouds for data analysis.

Plato Deliyannis, a Cornell REU student, implemented the flexible radio astronomy pipeline architecture and took responsibility for coding a friend-of-friends algorithm to examine the dynamic spectrum of high time-resolution radio observations, highlighting the best candidates from the observation. All of this work was done in Python and is available on GitHub. This code was run on several test data sets, including the discovery plot for the 121101 repeating FRB discovered by Laura Spitler.

Cornell REU student Cindy Wu was looking for trends and patterns in the growth of multiple species microbial communities in relation to the number of microbes in the community, working with five computational models of microbes that could be found in the gut microbiota of fruit flies. Coding was primarily done using COBRA toolbox, a linear programming toolbox that was added to MATLAB that contained scripts that could perform linear optimization on the fluxes of the computation models. All flux analyses were done to optimize the biomass equation in the models. Cindy wrote a script to construct multiple species models of any size from given individual models and after constructing the multiple species models, wrote code to analyze the exchange reactions of each individual model and all the possible combinations of different sized community models from the five given

models.

And, Cornell REU student Peter Cook conducted several analyses of the operating conditions for wind turbines (WT) using 10-minute output from long-term numerical simulations using the Weather Research and Forecasting (WRF) model (Mar. 2001 – Dec. 2016) produced by the atmospheric sciences use case. In order to assess the degree of coherence of wind speed variability at WT locations, Peter wrote several scripts to analyze WRF WS data using an array of statistical methods and learned how to employ parallel computing for efficiency. He also optimized non-parallel code by converting double-precision data to Boolean arrays and opting to perform vectorized and logical (or, and) operations over more complex mathematical and elementwise ones. Peter visualized these results in a large number of figures.

At UCSB, REU student William Berman completed his work with DrAFTS by designing a new interface for comparing spot instance prices. He graduated, completing his undergraduate degree and now works in industry as a professional developer.

Gareth George, another UCSB REU student, joined the Aristotle team and, to date, has participated in three different projects. Initially, he worked with William Berman on the revamp of the DrAFTS data management infrastructure. When that work was completed and William graduated, Gareth transitioned to a project analyzing Amazon's new pricing mechanism for spot instances. This work has resulted in a paper that will be submitted to IEEE International Conference on Cloud Computing (IC2E) in October. Gareth then began focusing on supporting the science teams using Aristotle for IoT analytics. His current project is developing a portable version of AWS Lambda so that the science team applications which use Lambda can be executed in remote locations.

In spring 2018, Aristotle resources supported a UCSB undergraduate Computer Science class taught by Rich Wolski on Cloud Computing (CS293B - https://www.cs.ucsb.edu/~rich/class/cs293b-cloud/). The student learning objective was to develop a multi-cloud system that used various cloud infrastructures in the best way possible. Students wrote a distributed search algorithm for non-convex optimization. They also combined Aristotle federation resources with XSEDE (Jetstream), Chameleon, CloudLab, and HTCondor resources in the same application. These large workloads (lots of small VMs) may help Jetstream better understand cloud workloads vs. HPC workloads. Wolski taught a second 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. These classes required ~200 IP addresses and an increase in network space.

UB built a Docker container with RStudio which the UB Statistics Dept. deploys on the UB Aristotle cloud and the university cloud for hundreds of statistics students. This is an example of how the Aristotle project is impacting campuses at large (beyond the Aristotle project science use cases). In a similar fashion, Cornell created 4 Window images and a Linux image to support Cornell Institute for Social and Economic Research faculty and student projects. Social sciences computing can be an ideal fit for the cloud considering statistical software is often used in a pleasingly parallel mode at moderate scale. Researchers especially like the availability of large memory instances: 28 core instances have 192GB RAM.

Many of the Aristotle science use cases also impact graduate and PhD student skills development. For example, at the University of Buffalo, Aristotle use case lead Dominik Roesch taught PhD students to use the OneTick framework and the underlying data hosted on Aristotle; they subsequently launch new investigations (e.g., impact of new SEC policy that widens the minimum quoting and trading increments–or tick sizes–for stocks).

At Cornell, after receiving training from Aristotle science team staff on how to build and use container images, PhD students in the Patrick Reed Research Group, applied this knowledge and OpenMORDM code was run across many cores on multiple Jetstream VMs utilizing Docker containers and a scripted cluster setup. They are also making progress on getting multi-instance MPI in a container operational.

Finally, the Aristotle use case support team at Cornell trained Angela Douglas students in modeling software and

best practices, as well as algorithmic issues.

K-12

Cornell is building a website to teach a focused workshop for Upward Bound students who will be visiting campus this fall. The workshop will teach the high school students introductory programming concepts using graphical interfaces, the Python language, and Aristotle Cloud Federation hosting.

How-To Documentation & Training

How-to user guides continued to be developed on GitHub and, when ready, publicly released on the Aristotle portal.

We have been engaged in extensive containerization work with a focus on training scientists in each research group so they can train other members of the group. In the future, we plan to develop and train users to use automation tools to start multi-node jobs on various supported clouds.

*** How have the results been disseminated to communities of interest?**

Scientific Meetings, Publications, and Conferences

Science use case researchers and the Aristotle team presented results at scientific meetings where they referenced the Aristotle project and its contribution to their success. Use case researchers also published papers which acknowledged the project. Sample activities are detailed in the “Products” section of this report which includes 39 publications, presentations, and videos.

Aristotle PI and co-PIs are actively involved in professional societies and conferences such as the Practice & Experience in Advanced Research Computing (Aristotle co-PI Tom Furlani is PEARC ‘19 Chair), Coalition for Academic Scientific Computing (CASC), SC Conference, and NSF workshops all of which have afforded opportunities to share progress on the Aristotle project. In addition, PI David Lifka has a leadership role in the eXtreme Science and Engineering Discovery Environment (XSEDE) project and keeps XSEDE management abreast of developments in cloud computing. Lifka also serves on the NSF Directorate for Computer, Information Science & Engineering (CISE) Advisory Committee. In addition, Resa Reynolds, Aristotle infrastructure lead, is president of the Dell XL HPC Consortium and Susan Mehringer, Aristotle portal lead, is the XSEDE training lead

Cornell and the University at Buffalo featured the Aristotle project at their November 2017 SC conference exhibits. Progress in five Aristotle use cases—a webGlobe interface to visualize and analyze geospatial data, computing market efficiency for all US stocks, mapping transcriptome data to metabolic models of gut microbiota, the successful use of IoT, edge computing and Aristotle for irrigation scheduling, and evaluating wind speed variability using Docker-installed WRF physics—were highlighted in a presentation at Cornell's exhibit. Aristotle team PIs and leads meet with industry partners and supporters, universities, and OEMs/ ISVs to brief them on the federated cloud concept and the project’s current status. Aristotle co-PI Rich Wolski’s paper on spot instances was published in SC17 Proceedings.

Lifka met with Google and AWS at EDUCAUSE and has had subsequent meetings with Microsoft to discuss the Aristotle open cloud marketplace concept. Lifka also met with Dartmouth who plans to join the federation to explore the federated cloud concept; Dartmouth will bring new use cases to the project, including digital humanities, and be the first university outside the funded partners to pilot the federation.

Aristotle Portal

The Aristotle portal provides scientists and the cyberinfrastructure community extensive information on the

project, including detailed use case accomplishments, plans, and products, publications, news and events, documentation, and detailed project reports (https://federatedcloud.org/).

Communicating to a General Audience

News stories about Aristotle use case scientists broaden the public understanding of the value of information technology and engineering in solving societal problems. For example, a February 2018 UCSB news story and video highlighted how computer scientists are tackling ecology and helping ecologists by bringing machine learning into nature to identify wildlife (see http://www.news.ucsb.edu/2018/018702/wheres-bear); a March 2018 NSF CISE tweet featured REU students contributing to the Aristotle Cloud Federation project by processing large-scale weather models in a complex dataset; and a July 2018 news article described how wireless farming may keep frost away from citrus. See Aristotle news at https://www.cac.cornell.edu/about/news/180328.aspx).

*** What do you plan to do during the next reporting period to accomplish the goals?**

Infrastructure & Portal Plans

- Order/install/configure PY4 storage assets and cloud infrastructure at each site to enhance science use case research capabilities and to continue to build a sustainable federated cloud model.
- Improve the OpenStack installations at each site to enhance research progress of the 7 science use case teams and new use cases from Dartmouth (first institution to test the federation).
- Complete the transition from a local to a federated accounting and allocations system.
- Complete the portal dashboard which will display usage data on a project level to science team members and automate account creation.
- Expand User Guide with best practices documentation. Add advanced and federation-specific topics.
- Provide science use team training on how to use the federated cloud and topics such as how to build containers efficiently.

Metrics & Usage Plans

- Complete integration of cloud metrics into federated Open XDMoD and test and refine with use case researchers.
- Ingest summaries into Open XDMoD and display performance data with Job Viewer.
- Integrate DrAFTS cloud-based metrics into Open XDMoD.

Science Use Case Plans

1. **A Cloud-Based Framework for Visualization and Analysis of Big Geo Data** (Varun Chandola, University at Buffalo). Scale up climate simulation data analyses and run across federated sites. Package the webGlobe software into an open source toolkit library that will be released for public use. The toolkit will enable a broad scientific community dealing with geo problems to launch analytical workflows that involve studying massive spatial and spatio-temporal datasets through machine learning-driven analytics and advanced visualization. The toolkit will consist of two open-source and extensible core libraries that will allow scientists to create integrated or stand-alone workflows for analytics and visualization. Additionally, the toolkit will also include an integrated browser-based system that will allow scientists to drive the analytics and visualization of geo data from a single platform.
2. **Global Market Efficiency Impact** (Dominik Roesch, University at Buffalo). Submit research investigating the impact of arbitrage using Thompson Reuters Tick History data and the recent tick-size pilot study by the U.S. Securities Exchange and Commission to finance journals. Complete the first draft investigating global market efficiency using Thompson Reuters Tick History Data and U.S. market efficiency. Start a new project investigating the impact of jumps in prices on holding period returns and volatility. Continue to mentor four finance PhD students using the framework hosted on Aristotle. Work together with an UB information technology student to automate part of the framework hosted on Aristotle.
3. **High Fidelity Modeling and Analytics for Improved Understanding of Climate** (Sara C. Pryor, Cornell

University). Continue to run high-resolution numerical simulations to improve the understanding of wind climate variability and change with a specific focus on applications to the wind energy industry. Continue to conduct high-resolution numerical simulations to improve the understanding of local climate perturbations resulting from the action of wind turbines in harnessing the kinetic energy of the atmosphere and converting it into (carbon-free) electricity.

4. **Transient Detection in Radio Astronomy Search Data** (James Cordes, Cornell University). Test several detection algorithms independently, including friend-of-friends and other two-dimensional searches, and the PRESTO-based (PulsaR Exploration and Search TOolkit) detection method used for the Spitler discovery of the Fast Radio Burst (FRB) 121102, running in the new pipeline framework. Combine these methods to evaluate the improvement in signal detection sensitivity. Build a classification scheme for signals (particularly the RFI signals). Make the Docker container publicly available and document the pipeline to be installed on it. Perform a production run on 10s of TBs of raw search data.
5. **Water Resource Management Using OpenMORDM** (Patrick Reed, Cornell University). Continue development of the Aristotle MPI cluster, i.e., MPI in a container. Benchmark the scalability of the water resource management software stack and investigate whether many containers can be spun up across multiple clouds, including bursting to AWS. Achieve efficient multi-node support, and benchmark the Parallel Platypus VM for scaling. Build and test two additional software batches.
6. **Mapping Transcriptome Data to Metabolic Models of Gut Microbiota** (Angela Douglas, Cornell University). Construct and analyze multi-compartment metabolic models of host-microbiome interactions in *Drosophila*. Construct models of varying complexity, comprising up to 5 microbial taxa, for which the individual metabolic models were constructed and curated in the last year, and the host model that is under development. Perform SteadyCom flux analysis using verified SteadyCom and multi-species construction procedure on multi-compartment models in a nutrient-rich and minimal media, to identify community metabolism, i.e. metabolic products of multi-species metabolic networks that are not produced by a single species in isolation, and compare community composition in silico with empirical data on community composition. Integrate these empirical data into SteadyCom as additional constraints and analyze flux patterns, with emphasis on community metabolism. Design fluxomics experiments informed by flux analysis of the models.
7. **Multi-Sourced Data Analytics to Improve Food Production and Security** (Kate McCurdy, Sedgwick Reserve; Elizabeth Grafton-Cardwell, UC Lindcove Research & Extension Center; Chandra Krintz, UCSB SmartFarm with Cal Poly, San Louis Obispo and Fresno State). Develop sustainable land use practices at Sedgwick that employ livestock as part of the management lifecycle and install new monitoring infrastructure. Deploy phase 2 infrastructure at Lindcove in November 2018 in time for the citrus “frost season” and test new real-time analytics in a production farm context. Analyze phase 1 SmartFarm results and in Spring 2019 deploy phase 2 at scale in order to observe a full year cycle and analyze the impact of differential irrigation.

Plans to Disseminate Results

The Aristotle team will continue to engage the cyberinfrastructure community through presentations and dialogue at CASC, SC18 and PEARC '19, and future scientific meetings such as the 8th International Conference on the Internet of Things where an Aristotle use case team was invited to present “Estimating Outdoor Temperature from CPU Temperature for IoT Applications in Agriculture” and the CASC Fall 2018 Meeting which will include Cornell CAC deputy director Richard Knepper updating center and facility directors on Aristotle during a “Lessons learned in federation across institutions” panel. Aristotle PI/co-PIs will continue to respond to inquiries regarding the project and keep leadership at XSEDE and relevant DIBBs projects abreast of new developments in the federated cloud model. The Aristotle portal will highlight project results, and be updated regularly with news, events, and additions to the user guide.

Products

Books

Sara C. Pryor & Andrea N. Hahmann (2018). *Downscaling wind: Forthcoming in Oxford Research Encyclopedias: Climate Science* Hans Von Storch. Oxford University Press. Status = AWAITING_PUBLICATION; Acknowledgment of Federal Support = Yes

Book Chapters

Inventions

Journals or Juried Conference Papers

Arun Sharma, Syed Mohammed Arshad Zaid & Arun Chandola (2018). WebGlobe: a cloud based geospatial analysis framework for interacting with climate data. *Juried Conference Paper*. . Status = UNDER_REVIEW; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Hiranya Jayathilaka, Chandra Krintz & Rich Wolski (2018). Detecting performance anomalies in cloud platform applications. *Journal*. . Status = AWAITING_PUBLICATION; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1109/TCC.2018.2808289

Kee H. Chung, Albert J. Lee & Dominik Rösch (2018). Tick size, liquidity for small and large orders, and price informativeness: evidence from the tick size pilot program. *Journal*. . Status = UNDER_REVIEW; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; OTHER: https://papers.ssrn.com /sol3/papers.cfm?abstract_id=3220470

Melissa R. Allen, Syed Mohammed Arshad Zaidi, Varun Chandola, April M. Morton, Christa M. Brelsford, Ryan A. McManamay, Binita KC, Jibonananda Sanyal, Robert N. Stewart & Budhendra L. Bhaduri (2018). A survey of analytical methods for energy-water nexus discovery. *Journal*. . Status = UNDER_REVIEW; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Nevena Golubovic, Angad Gill, Chandra Krintz & Rich Wolski (2017). CENTAURUS: A cloud service for K-means clustering. *Juried Conference Paper*. 1135. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1109/DASC-PICom-DataCom-CyberSciTech.2017.183

Rich Wolski and John Brevik (2017). QPRED: Using quantile predictions to improve power usage for private clouds. *Juried Conference Paper*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1109/CLOUD.2017.31

Rich Wolski, John Brevik, Ryan Chard & Kyle Chard. (2018). Probabilistic guarantees of execution duration for Amazon spot instances. *Juried Conference Paper*. 18:1. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1145/3126908.3126953

Sara C. Pryor, Rebecca J. Barthelmie & Tristan J. Shepherd (2018). The influence of real-world wind turbine deployments on local to mesoscale climate. *Journal*. 123 (11), 5804. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1029/2017JD028114

Sara C. Pryor, Rebecca J. Barthelmie, Andrea N. Hahmann, Tristan J. Shepherd & Patrick Volker (2018). Downstream effects from contemporary wind turbine deployments. *Juried Conference Paper*. 1037 . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1088/1742-6596/1037/7/072010

Sara C. Pryor, Tristan J. Shepherd & Rebecca J. Barthelmie (2018). Inter-annual variability of wind climates and wind turbine annual energy production. *Journal*. . Status = AWAITING_PUBLICATION; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.5194/wes-2018-48

Stratos Dimopoulos, Chandra Krintz, and Rich Wolski (2017). PYTHIA: Admission control for multi-framework,

deadline driven, big data workloads. *Juried Conference Paper*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1109/CLOUD.2017.69

Syed Mohammed Arshad Zaidi, Melissa R. Allen, Varun Chandola & Budhendra L. Bhaduri (2018). Machine learning for energy-water nexus: challenges and opportunities. *Journal*. . Status = AWAITING_PUBLICATION; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Wei-Tsung Lin, Chandra Krintz & Rich Wolski (2018). Tracing function dependencies across clouds. *Juried Conference Paper*. . Status = AWAITING_PUBLICATION; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; OTHER: https://www.cs.ucsb.edu/~ckrintz/papers/lowgo18.pdf

Wolski, R., Brevik, J., Chard, R., and Chard K. (2017). Probabilistic Guarantees of Execution Duration for Amazon Spot Instances.. *IEEE International Conference on Cloud Engineering. (IC2E 2017)*. . Status = OTHER; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; OTHER:

Licenses

Other Conference Presentations / Papers
Sara C. Pryor, Rebecca J. Barthelmie & Tristan Shepherd (2018). *Assessing the fidelity of the North American wind climate and impacts on wind farms using high resolution modeling*. 98th American Meteorological Society Annual Meeting (21st Conference on Planned and Inadvertent Weather Modification). Austin, TX. Status = OTHER; Acknowledgement of Federal Support = Yes

Terrence Hendershott, Dmitry Livdan & Dominik Rösch (2018). *Asset pricing: a tale of night and day*. Research in Behavioral Finance. Amsterdam, NL. Status = OTHER; Acknowledgement of Federal Support = Yes

Thomas Furlani (2017). *Building a federated cloud model*. 1st NSF Data Infrastructure Building Blocks PI Workshop (DIBBs 17). Arlington, VA. Status = OTHER; Acknowledgement of Federal Support = Yes

David Lifka, Thomas Furlani, and Rich Wolski (2017). *CC**DNI DIBBs: Data analysis and management building blocks (DIBBs) for multi-campus cyberinfrastructure through cloud federation*. 1st NSF Data Infrastructure Building Blocks PI Workshop (DIBBs 17). Arlington, VA. Status = OTHER; Acknowledgement of Federal Support = Yes

David Lifka, Thomas Furlani & Rich Wolski (2018). *CC*DNI DIBBs: Data analysis & management building blocks for multi-campus cyberinfrastructure through cloud federation*. 2nd NSF Data Infrastructure Building Blocks PI Workshop. Arlington, VA. Status = OTHER; Acknowledgement of Federal Support = Yes

Thomas Furlani (2018). *Campus-based systems and the national cyberinfrastructure ecosystem*. Opportunities from the Integration of Simulation Science and Data Science Workshop. Washington, DC. Status = OTHER; Acknowledgement of Federal Support = Yes

David Lifka, Thomas Furlani & Rich Wolski (2018). *Challenges: DIBBs for multi-campus cyberinfrastructure through cloud federation*. 2nd NSF Data Infrastructure Building Blocks PI Workshop. Arlington, VA. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Richard Knepper (2017). *Cloud computing perspectives*. Coalition for Academic Scientific Computing Fall Workshop. Westminster, CO. Status = OTHER; Acknowledgement of Federal Support = Yes

Rich Wolski (2017). *Cloud computing status and future*. 10th IEEE International Conference on Cloud Computing. Honolulu, HI. Status = OTHER; Acknowledgement of Federal Support = Yes

Richard Knepper (2017). *Cloud day: Current use cases at academic institutions*. Coalition for Academic Scientific

Computing Fall 2017 Meeting. Westminster, CO. Status = OTHER; Acknowledgement of Federal Support = Yes

Sara C. Pryor, Rebecca J. Barthelmie, Andrea N. Hahmann, Tristan J. Shepherd & Patrick Volker (2018). *Contemporary wind turbine deployments have a minor impact on regional climate*. Science of Making Torque from Wind Conference. Milan, Italy. Status = OTHER; Acknowledgement of Federal Support = Yes

Steven Lee (2017). *Cornell Red Cloud: Campus-based hybrid cloud*. National Science Foundation Cybersecurity Conference. Arlington, VA. Status = OTHER; Acknowledgement of Federal Support = Yes

April Morton, Jesse Piburn, Robert N. Stewart & Varun Chandola (2017). *Development of a suite of analytical tools for energy and water infrastructure knowledge discovery*. American Geophysical Union Fall Meeting. New Orleans, LA. Status = OTHER; Acknowledgement of Federal Support = No

Sara C. Pryor, Rebecca J. Barthelmie & Tristan J. Shepherd (2018). *Do current and near-term future wind turbine deployments have a substantial impact on regional climate?*. European Geosciences Union General Assembly 2018. Vienna, Austria. Status = OTHER; Acknowledgement of Federal Support = Yes

Rich Wolski (2018). *Emerging trends in the economics of cloud computing*. NSF Workshop on Cloud Computing Economics. Palo Alto, CA. Status = OTHER; Acknowledgement of Federal Support = Yes

Martins Innus & Andrew Bruno (2018). *Federated keystone single sign-on with FreeIPA and OpenID Connect*. OpenStack Summit. Vancouver, BC. Status = OTHER; Acknowledgement of Federal Support = Yes

Joan Song, Brandon Barker, Nana Ankrah & Angela Douglas (2018). *Genome-scale metabolic modeling of gut microbiota in the fly gut*. American Society for Engineering Education 2018 St. Lawrence Section Meeting. Ithaca, NY. Status = OTHER; Acknowledgement of Federal Support = Yes

Sara C. Pryor (2017). *High resolution WRF simulations for resource assessment and quantification of downstream impacts of high density wind turbine deployments*. DTU Wind Energy. Roskilde, Denmark. Status = OTHER; Acknowledgement of Federal Support = Yes

Sara C. Pryor, Rebecca J. Barthelmie, and Tristan Shepherd (2017). *High-fidelity simulations of the downstream impacts of high density wind turbine deployments*. 4th Workshop on Systems Engineering for Wind Energy. Roskilde, Denmark. Status = OTHER; Acknowledgement of Federal Support = Yes

Sara C. Pryor, Rebecca J. Barthelmie & Tristan Shepherd (2018). *Improved characterization of the magnitude and causes of spatio-temporal variability in wind resources*. 98th American Meteorological Society Annual Meeting (31st Conference on Climate Variability and Change). Austin, TX. Status = OTHER; Acknowledgement of Federal Support = Yes

Dominik Roesch (2017). *Investigating the efficiency of financial stock markets with high frequency data*. 2nd Federal Reserve Bank Economic Research in High Performance Computing Environments Workshop. Kansas City, KS. Status = OTHER; Acknowledgement of Federal Support = Yes

April Morton, Jesse Piburn, Robert N. Stewart & Varun Chandola (2018). *Leveraging advances in population modeling to support energy and water nexus knowledge discovery*. American Geophysical Union Fall Meeting. New Orleans, LA. Status = OTHER; Acknowledgement of Federal Support = No

Angela E. Douglas (2017). *Metabolic conversions in insect microbiomes*. 14th Symposium on Bacterial Genetics and Ecology. Aberdeen, Scotland. Status = OTHER; Acknowledgement of Federal Support = No

Angela E. Douglas (2017). *Metabolism and microbiomes: Metabolic models meet experimental data in insect-microbial symbiosis.*. 36th Summer Symposium in Molecular Biology: Metabolism: Disease Models and Model

Organisms. State College, PA. Status = OTHER; Acknowledgement of Federal Support = No

Thomas Furlani (2018). *Metrics*. Coalition for Academic Scientific Computation. Alexandria, VA. Status = OTHER; Acknowledgement of Federal Support = Yes

Shami Chatterjee (2018). *New results in radio astronomy: Fast radio bursts and transients*. International Astronomical Union (IAU) General Assembly. Vienna. Status = OTHER; Acknowledgement of Federal Support = Yes

Nana Y.D. Ankrah (2017). *Nutritional roles of beneficial bacteria associated with insects revealed by metabolic modeling*. Copenhagen Bioscience Conference: Data-Driven Biotechnology—Bench, Bioreactor and Bedside. Copenhagen, Denmark. Status = OTHER; Acknowledgement of Federal Support = Yes

Steven Lantz (2017). *Parallel computing with MATLAB and scaling to Red Cloud*. MathWorks Technical Sessions at Cornell. Ithaca, NY. Status = OTHER; Acknowledgement of Federal Support = Yes

Rion Dooley, Andy Edmonds, David Y. Hancock, John Michael Lowe, Edwin Skidmore, Andrew K. Adams, Ryan Kiser, Mark Krenz, Von Welch & Richard Knepper (2018). *Security best practices for academic cloud service providers*. 2018 NSF Cybersecurity Summit for Large Facilities and Cyberinfrastructure. Alexandria, VA. Status = OTHER; Acknowledgement of Federal Support = Yes

Tristan J. Shepherd, Patrick Volker, Rebecca J. Barthelmie, Andrea N. Hahmann & Sara C. Pryor, Tristan J. Shepherd & Patrick Volker (2018). *Sensitivity of wind turbine array downstream effects to the parameterization used in WRF*. WRF/MPAS User's Workshop. Boulder, CO. Status = OTHER; Acknowledgement of Federal Support = Yes

Angela E. Douglas (2017). *The Drosophila model for gut microbiome research*. NIH Common Fund: The Human Microbiome: Emerging Themes at the Horizon of the 21st Century. Bethesda, MD. Status = OTHER; Acknowledgement of Federal Support = No

Shami Chatterjee (2018). *The dwarf galaxy host of a repeating fast radio burst*. International Astronomical Union (IAU) General Assembly. Vienna. Status = OTHER; Acknowledgement of Federal Support = Yes

Jibonananda Sanyal, Varun Chandola, Alexandre Sorokine, Melissa R. Allen, Anne Berres, Harrison Pang, Rajasekar Karthik, Phil Nugent, Ryan McManamay, Robert Stewart & Budhendra L. Bhaduri (2017). *Towards a web-enabled geovisualization and analytics platform for the energy and water nexus*. American Geophysical Union Fall Meeting. New Orleans, LA. Status = OTHER; Acknowledgement of Federal Support = No

Varun Chandola (2017). *WebGlobe – A cloud based geospatial analysis framework for interacting with climate data*. 6th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data. Redondo Beach, CA. Status = OTHER; Acknowledgement of Federal Support = Yes

Other Products

Audio or Video Products.

Assessing the Fidelity of the North American Wind Climate and Impacts of Wind Farms Using High-Resolution Modeling video features Cornell Aristotle use case scientist Tristan Shepherd describing how wind farms impact climate at the American Meteorological Society Meeting. See the video at https://ams.confex.com/ams/98Annual/videogateway.cgi/id/44205?recordingid=44205&uniqueid=Paper331226&entry_password=508882 .

Audio or Video Products.

Exploiting Animal-Microbe Symbiosis to Produce Pest-Resistant Crops World Economic Forum video features Cornell Aristotle use case scientist Angela Douglas describing how understanding the gut microbiome may lead to

development of microbial therapies for everything from making honey bees healthier to restoring coral reefs. See the video at https://www.youtube.com/watch?v=r-H_uBC-m_4&index=2&list=PL7m903CwFUgkQ2aiznKfb5Y-HgquhBsHk. A *Journal of Bacteriology* paper (https://doi.org/10.1128/JB.00872-16) details this use case.

Audio or Video Products.

Federated Keystone Single Sign-On with FreeIPA and OpenID Connect video features UB Aristotle infrastructure team members Martins Innus and Andrew Bruno describing integrating federated keystone with FreeIPA as the backend identity provider using OpenID Connect as the federation protocol. See the 2018 OpenStack Summit video at https://www.youtube.com/watch?v=7BSnhRZ8nhs.

Audio or Video Products.

Where's the Bear? video and UCSB Current article by Shelly Leachman describe how Aristotle use case scientists Chandra Krintz and Rich Wolski are helping ecologists by bringing machine learning into nature to identify wildlife. See the video at http://www.news.ucsb.edu/2018/018702/wheres-bear. A *Proceedings of the 2nd ACM/IEEE International Conference on Internet-of-Things Design and Implementation* paper (https://doi.org/10.1145/3054977.3054986) details this use case project.

Audio or Video Products.

Wind Speed Variability Visualization video shows Sara C. Pryor Aristotle use case team simulations used to evaluate the inter-annual variable of wind speed and other atmospheric properties over the eastern U.S. High-resolution numerical simulations were used to quantify wind climate and analyze the impact of large wind turbine developments on downstream climate. See the video at https://federatedcloud.org/science/Pryor_video_RSV.gif. A *Journal of Geophysical Research: Atmospheres* paper (https://doi.org/10.1029/2017JD028114) details this use case project.

Other Publications

Patents

Technologies or Techniques

The Aristotle project team and science use case researchers are developing:

new container and Singularity images that create reproducible research workflows which due to their portability can be shared broadly across institutions and disciplines
new techniques that facilitate edge cloud systems with support for multi-analytics (multi-tenant) interference avoidance, graceful degradation and fault tolerance, and appliance-like self-management
new programming and deployment models for cloud and IoT systems, including serverless functions-as-a-service
new K-Means a service called CENTAURUS that will enable scientists to analyze data to determine the best number of clusters for it, using the K-Means algorithm with Mahalanois distance and Bayesian Information Criterion.

Thesis/Dissertations

Websites
Aristotle Cloud Federation
https://federatedcloud.org

The Aristotle Cloud Federation portal was updated regularly to feature new web content and user guide documentation as well as Aristotle science use case accomplishments, plans, and products. Quarterly National

Science Foundation project reports are also available on the portal; they provide detailed descriptions of the project's activities, challenges, and accomplishments.

Participants/Organizations

Research Experience for Undergraduates (REU) funding

Form of REU funding support:	REU supplement
How many REU applications were received during this reporting period?	40
How many REU applicants were selected and agreed to participate during this reporting period?	5
REU Comments:	

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Lifka, David	PD/PI	1
Furlani, Thomas	Co PD/PI	1
Wolski, Richard	Co PD/PI	3

Full details of individuals who have worked on the project:

David A Lifka
Email: lifka@cac.cornell.edu
Most Senior Project Role: PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Programmatic oversight of the Aristotle Cloud Federation project ensuring deliverables outlined in the program execution plan are met on schedule.

Funding Support: No funding support from other projects used for this award.

International Collaboration: No
International Travel: No

Thomas R Furlani
Email: furlani@buffalo.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Programmatic oversight of the UB subaward, and continuous interaction with the entire Aristotle technical team.

Funding Support: No funding support from other projects used for this award.

International Collaboration: No

International Travel: No

Richard Wolski
Email: rich@cs.ucsb.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 3

Contribution to the Project: Architected the deployment at UCSB for production Aristotle services. Developed the AWS spot market prediction system. On-boarded new Sedgwick science team efforts.

Funding Support: No funding support from other projects used for this award.

International Collaboration: No
International Travel: No

What other organizations have been involved as partners?
Nothing to report.

What other collaborators or contacts have been involved?
Nothing to report

Impacts

What is the impact on the development of the principal discipline(s) of the project?

The Aristotle project is advancing the knowledge of federated and hybrid cloud computing and their potential roles as campus bridging paradigms. By building and deploying a federated cloud model with the necessary allocations, accounting, and cloud metrics, Cornell University, University at Buffalo, and UC Santa Barbara are exploring how cloud resources can be effectively shared between campuses and their impact on researchers who require flexible workflows and analysis tools for large-scale data sets. The project serves as an important model for campus cyberinfrastructure that others may follow and provides elasticity by sharing resources, data, software, and tools that may not be available locally. The project is also developing an allocation model that provides a fair exchange mechanism for resource access between institutions.

What is the impact on other disciplines?

Aristotle use case scientists are strategically exploring problems of increasing complexity and corresponding increases in data and, as a result, are advancing scientific knowledge. Data challenges from a diversity of disciplines (earth and atmospheric sciences, finance, chemistry, astronomy, civil engineering, genomics, and agriculture) are being addressed with collaborators from other academic institutions, public agencies, and research labs, as well as citizen scientists. The sharing of data infrastructure building blocks capacity and the transparent movement of instances across institutional boundaries has the potential to create wider science collaborations and increased data sharing. The ability to predict when it is appropriate to burst from local cloud resources to other campus resources, NSF clouds, or public clouds has the potential to produce a positive impact on research productivity.

What is the impact on the development of human resources?

Aristotle is pioneering the concept of federated cloud computing for research which may ultimately increase the availability of on demand resources, data, and analysis tools that engage underrepresented groups. In addition, virtual laboratories in the cloud can enhance classroom learning. For example, Aristotle cloud resources were among the multiple cloud infrastructures used simultaneously to solve large problems in a Computer Science class (CS293B) taught at UC Santa Barbara in spring 2018. The availability of campus and hybrid cloud computing may also spurn the development and dissemination of ready-to-launch VMs and containers with training software and tools preloaded. This could reduce the redundant development and preparation of educational material development and onsite computer labs administration, resulting in an increased focus on individual student learning needs. Ten REU students have participate in the project to date.

What is the impact on physical resources that form infrastructure?

The federated cloud model may impact the physical resources that form infrastructure by reducing the number of computer labs required for learning. Hybrid clouds may be installed so that campuses can cost effectively use local cloud resources and, when more capacity is needed, burst to the most suitable campus, public, or NSF cloud resource. Ultimately, federated clouds will likely become complementary resources to high-end supercomputers, e.g., performing on demand iterative tasks, streaming IoT data, etc. Integrating federated clouds (campus, labs, etc.) into the national cyberinfrastructure ecosystem may increase the sharing of resources and, subsequently, reduce physical resource expenditures at the local level.

What is the impact on institutional resources that form infrastructure?

Aristotle will maximize institutional resources through federation with other institutions by (1) offloading variable computational and data analysis demands from local infrastructure with dynamic resource allocation, (2) starting coarsely parallel computations on demand, (3) bursting to process new data, (4) providing heterogeneous instance types and sizes to allow for unpredictable computational demand.

What is the impact on information resources that form infrastructure?

Aristotle's federated cloud model will facilitate the (1) sharing of high-value processed datasets of general interest and separate data resources, (2) generation of reproducible pipelines in the form of VMs or VM configurations, and containers (3) access to multiple data sources, many of which are already in public and private clouds.

What is the impact on technology transfer?

AppScale, a UCSB startup, is now supporting Eucalyptus, an AWS-compatible cloud.

What is the impact on society beyond science and technology?

Aristotle use cases have the potential to impact urban planners interested in when local climate will change for a specific location, policymakers regulating high-frequency trading, engineers deploying large-scale wind energy resources, policymakers making water resources management decisions, manufacturers producing sustainable insect pest management products, and farmers increasing yields and protecting the environment by accessing on demand soil, water, and crop sensor data to aid decision-making.

Changes/Problems

Changes in approach and reason for change
Nothing to report.

Actual or Anticipated problems or delays and actions or plans to resolve them

Nothing to report.

Changes that have a significant impact on expenditures
Nothing to report.

Significant changes in use or care of human subjects
Nothing to report.

Significant changes in use or care of vertebrate animals
Nothing to report.

Significant changes in use or care of biohazards
Nothing to report.