*Note: this revised PY2 Annual Report is the same report submitted to Research.gov in Sept. 2017 with the exception of an addendum to Accomplishments (pgs. 4 & 5). These were inadvertently truncated when the original report was submitted.*

## Project Report Printer Friendly Version

Click here to print

# 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**<br>**Thomas R Furlani, Co-Principal Investigator**<br>**Richard Wolski, Co-Principal Investigator** |
| Recipient Organization: | **Cornell University** |
| Project/Grant Period: | **10/01/2015 - 09/30/2020** |
| Reporting Period: | **10/01/2016 - 09/30/2017** |
| Submitting Official (if other than PD\PI): | **David A Lifka**<br>**Principal Investigator** |
| Submission Date: | **10/03/2017** |
| Signature of Submitting Official (signature shall be submitted in accordance with agency specific instructions) | **David A Lifka** |

Back to the top

## Accomplishments

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

- 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 food science) 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 features QBETS statistics to make online forecasts of future performance and allocations levels available to users.

Aristotle PIs and team leaders participated in an 18-month independent project review at the National Science Foundation in April 2017. 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 XDMoD), and
- predict the durability of AWS spot pricing which may help researchers lower their costs of computing (via DrAFTS).

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:

Specific Objectives:

**Architect and install PY2 DIBBs infrastructure/storage assets at 3 federation sites.**

- Installed DIBBs storage assets and cloud infrastructure at Cornell University, the University at Buffalo, and UC Santa Barbara based on science use case requirements. Using PY2 DIBBs funding, Cornell added 768TB Ceph storage (plus with internal Cornell funds an additional 384TB Ceph storage and 112 Red Cloud cores due to ~80% core usage); Buffalo added 144TB Ceph storage and 112 cores to their Lake Effect cloud; and, UCSB added 240TB Ceph storage and 216 cores to their Aristotle cloud.
- Increased total storage assets and cores in production (including existing clouds at Cornell and Buffalo) to 1824TB Ceph storage, 562TB SAN storage, and 1,268 cores.
- Upgraded UB cloud connection to the public network to 10Gb. Installed network switches at UCSB and redesigned Ceph topology for added fault tolerance and expansion capabilities. Increased bandwidth between CU and UCSB by ~1.5 – 2 Gbits/sec by increasing transmit/receive TCP window on Red Cloud nodes to 2GB and setting the Linux power management profile from “default” to “virtual host” so CPUs are not clocked down.
- Tested Eucalyptus 4.4 which supports OAuth 2.0 login via Globus and deployed at CU and UB. Plan to investigate/integrate OAuth 2.0 with OpenStack next.

**Develop portal code to allow Globus authentication by users.**

- Implemented Globus OAuth 2.0 login for portal authentication.

**Develop Phase 2 portal content.**

- Updated portal content on an ongoing basis to reflect the federation’s latest capabilities, publications, science news, etc., (e.g., added webGlobe analysis and visualization tool developed by the Aristotle big geo data science use case team with secure portal access).
- Deployed a usage graph to show early usage data from all 3 sites: https://federatedcloud.org/using/federationstatus.php
- Provided on-demand access to all Aristotle reports: https://federatedcloud.org/reports/.
- Completed the accounting and allocations database and tables, and populated with real data.  Completed PHP Scripts to import usage data from all federation sites using the REST API.

**Maintain production capabilities at the 3 federation sites so science use case research is not impeded while preparing to transition from Eucalyptus to OpenStack.**

- Stress tested new hardware at each site, performed necessary software upgrades, performed troubleshooting on the network, and addressed other infrastructure issues.

*Addendum to Accomplishments (begins top of page 4 and ends bottom of page 5)*

- Planned and currently performing initial tests of OpenStack which Aristotle will transition to due to DXC Technology's decision to no longer support Eucalyptus.
- Negotiated with Red Hat for multi-year OpenStack and Ceph support; a "one-site" contract for all federation sites is complete.

**Implement Open XDMoD cloud data collection across the federation.**

- Completed modifications to the XDMoD schema to support cloud data.
- Enhanced the log scraper to improve coverage of desired events, i.e., detaching a volume from an instance.
- Modified the XDMoD ETL v2 process to support ingesting cloud data, e.g., scanning a directory to identify files recently modified, populating multiple database tables from a single JSON record, etc.
- Ingested data periodically (and working on a daily ingestion process) with the development of initial metrics, e.g., Resources (tracking state and settings of whole resources); Requests (tracking successful and failed request for start, stop, suspend, resume, create, attach, etc.); Events (tracking activity of instances in clouds, e.g., volumes attached); and States (tracking time instances spent in various states, e.g., running, migrating, etc.).

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

- Improved the quality of DrAFTS predictions by reducing a ~4% error rate to < 1%.
- Started a new set of experiments with Globus Genomics who are developing a new scheduler that will utilize the DrAFTS.

**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** (UB Chandola). **Requirements:** Completed first runs of UB's Gaussian Process-based change detection algorithm on 200 years of climate simulation data. Developed and released to Aristotle researchers webGlobe, a browser-based user interface to the "Machine Learning for Sustainability" framework that allows scientists to upload, visualize, and analyze Network Common Data Form (NetCDF) data sets and is, at present, the only browser-based system available with this functionality (see demo at https://www.youtube.com/watch?v=-jhCwOda9fU).

2. **Global Market Efficiency Impact** (UB Roesch). Setup the framework on a VM and began to import and successfully analyze 2TBs of international tick-by-tick financial data from the Thomson Reuters Tick History (TRTH) database. Imported 25TBs Trade and Quote (TAQ) data up until 2017 for intraday trades and quote transactions for all US exchange-listed stocks, and computed 9 measures of efficiency.

3. **High Fidelity Modeling and Analytics for Improved Understanding of Climate** (CU Pryor). Installed the physics-only version of the Weather Research and Forecasting (WRF) model using Docker and compiled it with parallel NetCDF to evaluate cloud-based performance. Ran high-resolution simulations to quantify wind climate and analyze the impact of large wind turbine developments on downstream climate (local to mesoscale). Evaluated 10-minute wind speeds from simulations relative to in-situ measurements from the National Weather Service Automated Surface Observing System (ASOS) on Jetstream to allow Aristotle to continue to focus on the numerical simulations.

4. **Transient Detection in Radio Astronomy Search Data** (CU Cordes). Designed the pipeline architecture with an emphasis on pluggability and the selection of different smoothing codes under development at Cornell. Building the code to down-resolve the data, produce graphic output, perform production runs, and extend to include new algorithms.

5. **Water Resource Management Using OpenMORDM** (CU Reed). Built and tested Parallel Platypus software (the Python version of Open MORDM) with a VM, container, and container source (Docker file) developed by the Aristotle team for use as a virtual MPI cluster on Aristotle.

6. **Mapping Transcriptome Data to Metabolic Models of Gut Microbiota** (CU Douglas). Developed and published a computational model for the whiteflies system. Analyzed and compared three independently-evolved communities in xylem-feeding insects. Generated advanced draft metabolic reconstructions for 5 bacteria needed to simulate *Drosophila*-microbial community metabolic interactions. The draft models of the 2 Lactobacilli have an average of 586 genes, 1193 metabolites and 750 reactions. The draft models of the 3 Acetobacter have an average of 686 genes, 1362 metabolites and 1129 reactions. Initiated a computational strategy on Aristotle, based on a new optimization framework SteadyCom, to obtain the metabolic fluxes for optimized relative abundances of the partners under equilibrium conditions of constant growth rates.

7. **Multi-Sourced Data Analytics to Improve Food Production and Security** (Sedgwick Reserve McCurdy/UCSB). Completed large-scale Google TensorFlow training and image classification runs on Aristotle for the *"Where's the Bear?"* camera trap application analysis project (20 training runs using 2000 cores hours per run and a classification run using 1800 core hours to classify a test sample of 10,000 images prior to classifying 240,000 images from a single camera). Used soil moisture sensing devices, edge computing, and Aristotle to schedule vineyard irrigation saving 66% of the water used previously.

**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.

Aristotle Science Team Advisory Committee (STAC) meetings facilitated knowledge sharing between diverse science use cases.

Cross-site discussions facilitated the sharing of research tools and cloud computing techniques, e.g., Cornell researcher Sara Pryor is exploring the possibility of using a webGlobe visualization and analysis tool developed by UB researcher Varun Chandola for the output of her Aristotle runs.

A train-the-trainer approach was used to 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, onsite AWS Foundation Services for Research Computing and High Performance and Research Computing presentations, and onsite Red Hat Ceph training, were beneficial to the federation as a whole.

**Undergraduate & Graduate Student Development**

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

For example, Cornell REU student Thomas Biondi (1) processed over 10TB of complex datasets produced by weather models on multiple computing platforms, (2) used Aristotle to store the large amounts of data while simultaneously running the data through code on a Linux virtual machine, (3) applied statistical metrics to compute weather model accuracy, (4) used machine learning algorithms to predict days with high wind speeds, (5) created visualizations to present the data in a way that a lay audience could understand. As a result of this work, Biondi is an author on a submission to the 98th American Meteorological Society Annual Meeting (31st Conference on Climate Variability and Change), January 2018.

Also at Cornell, working with recent PhD recipient Robert Wharton and Aristotle Science Team Lead Adam Brazier, REU students Elizabeth Holzknecht and Shiva Laskhamann built code to down-resolve radio astronomy search data, produce graphical output, and rebuilt code for production runs and extension to new algorithms. Their code will be run on datasets know to contain transient sources and on pulsar (PALFA) blind search datasets.

Angela Douglas's Cornell REU student, Joan Song, enabled initial research on one three-compartment systems to complex microbial communities, focusing on the bacterial community in the *Drosophila*.

At UCSB, REU student William Berman has been researching the ability to predict market prices in the Amazon spot market for cloud resources. Working with DrAFTS, he has developed a new web-service venue that is easier to deploy, more scalable, and has a more user-friendly interface than the original prototype. He is also beginning to look at using DrAFTS to implement a long-lived VM service with the same longevity characteristics as the on-demand service, but at a significantly lower price.

In spring 2017, Aristotle resources supported a UCSB undergraduate Computer Science class 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, Chameleon, CloudLab, and HTCondor resources in the same application.

Many of the Aristotle science use cases also impact graduate and PhD student development. For example, at the University of Buffalo, PhD students were introduced to the Aristotle finance framework and two students are now completing their first queries that will allow the finance research group to start analyzing data related to the recent tick-size pilot study by the Securities Exchange Commission (SEC).

At Cornell, Aristotle science team staff successfully trained graduate student Bernardo Trindade and other Patrick Reed use case lab members on how to use Docker (how to build container images, stop and start images, and other container management tasks) with the understanding that they would train future group members. They are currently testing Multiobjective Robust Decision Making (MORDM)-related software in a single node container environment and have expressed a strong interest in moving forward to a multi-node environment.

Finally, Aristotle, use case scientist Angela Douglas's NIH grant proposal was funded. As a result, the Aristotle project will benefit from 40 hours of effort which will be used to train students in modeling software and best practices, as well as algorithmic issues.

**How-To Documentation & Training**

Containerization work has commenced with a focus on training a scientist in each research group so they can train other members in their groups. Cross cloud-stack testing of containerized science use case solutions will initially take place on Jetstream (which is based on OpenStack) while Aristotle transitions to OpenStack.

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

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

**DIBBs17 Workshop**

Aristotle PI David Lifka Chaired and Aristotle project staff organized the 1st NSF Data Infrastructure Building Blocks PI Workshop (DIBBs17) which discussed and disseminated the results of DIBBs projects. 65 PIs and co-PIs, and 7 NSF directors participated. The January 2017 Arlington, VA workshop was a supplemental award to the Aristotle project. 37 posters on significant DIBBs success and innovations were featured at the workshop and 37 white papers were submitted identifying DIBBs project challenges and solutions. NSF program director Amy Walton considered DIBBs17 an "outstanding" PI Workshop for the DIBBs community. "Your efforts created an energetic—highly productive—environment for the workshop. These activities encouraged involvement, collaboration, and contribution. Participants were excited about what they were doing." The 67-page "Final Report: 1st NSF Data Infrastructure Building Blocks Workshop" was published in April 2017 and is available at: https://dibbs17.org/report/DIBBs17FinalReport.pdf.

**Scientific Meetings, Publications, and Conferences**

Aristotle science use case faculty presented results at scientific meetings where they referenced the Aristotle project and its contribution to their success. Use case faculty also published papers which acknowledged the project. Examples of these activities are detailed in the "Products" section of this report.

Aristotle PI and co-PIs are actively involved in professional societies and conferences such as the Coalition for Academic Scientific Computing (CASC), Practice & Experience in Advanced Research Computing (PEARC), and SC Conference which have afforded the opportunity to share progress on the Aristotle project. In addition, PI David Lifka also has a leadership role in the XSEDE 2.0 eXtreme Science and Engineering Discovery Environment project and keeps XSEDE management abreast of developments in cloud computing.

Cornell featured the Aristotle project at their SC16 conference exhibit in November 2016. Three use cases, one from each site (UB Big Geospatial Data, CU Gut Microbiota, and UCSB Multi-Sourced Data Analytics to Improve Food Production and Security) were highlighted in a presentation at Cornell's booth. Cornell briefed AWS, NSF staff, universities, and OEMs and ISVs on the federated cloud concept and the project's current status. The presentation is posted on the Aristotle portal at https://federatedcloud.org/papers /SC16AristotleCloudFederationOverviewAndUseCases.pdf.

**Aristotle Portal**

The Aristotle portal provides the greater scientific and cyberinfrastructure communities extensive information on the project, including news, a user guide, and all NSF reports which have up-to-date details on the project's successes and challenges (see 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 UCSB news story highlighted the use of SmartFarm tools and hybrid clouds by co-PI Rich Wolski and his collaborators to create agriculture analytics to that enable sustainable farming practices (see http://www.news.ucsb.edu/2017/018008/dreaming-big).

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

**Infrastructure & Portal Plans**

- Order/install/configure PY3 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.
- Transition from the Eucalyptus cloud platform to Red Hat OpenStack without negatively impacting the research progress of the 7 science use case teams.
- Implement OAuth 2.0 with OpenStack so that Aristotle users can use Globus authentication to login at any site.
- Continue to share DIBBs cloud resources and transition from a local to a federated accounting and allocations system.
- Develop the portal dashboard which will display usage data on a project level to science team members and automate account creation.
- Expand User Guides and Technical Guides 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 cloud beta implementation of Open XDMoD. Ship data from XDMoD instances at the individual sites to a master XDMoD instance at UB where overall cloud data will be displayed.

- Complete development of the next generation DrAFTS website and release publicly.
- Beta test DrAFTS with Globus Genomics users.
- Integrate DrAFTS cloud-based metrics into Open XDMoD.

**Science Use Case Plans**

1. **A Cloud-Based Framework for Visualization & Analysis of Big Geo Data** (UB Chandola). Validate the results of climate change detection model simulation outputs with expert knowledge. Scale up simulation data analysis. Develop an approximation of the change detection to reduce interactive run time. Release the webGlobe browser-based user interface for uploading, visualizing, and analyzing geospatial data sets to the general scientific community with Globus authentication. Develop a version of webGlobe that integrates climate simulations with energy and water usage GIS data sets and develop analytic tools for the combined analysis.
2. **Global Market Efficiency Impact** (UB Roesch). Clean up/extend the sample and conduct a full-scale analysis of global market efficiency using Thompson Reuters Tick History data. Compute additional measures of efficiency using TAQ data. Mentor new UB Finance PhD research projects that will be querying these databases, and include a project related to the recent tick-size pilot study by the SEC. Present the Aristotle cloud environment and the investigation of the efficiency of financial stock markets with high-frequency data to the Federal Reserve Bank. Release the framework to the greater scientific community (long-term goal).
3. **High Fidelity Modeling and Analytics for Improved Understanding of Climate** (CU Pryor). Continue to run simulations for 2001-2016 to evaluate the inter-annual variability of wind speeds and other atmospheric properties over the eastern U.S. Run simulations of future (likely mid-century) climate conditions to examine climate change. Examine contemporary climate variability and possible changes, i.e., will the degree of variability change? Continue to also focus on wind speeds and application to the wind energy industry.
4. **Transient Detection in Radio Astronomy Search Data** (CU Cordes). Continue to improve the search data software and pipeline for both Aristotle and Jetstream. Reduce the PALFA dataset by 4 in time resolution and 2 in frequency resolution. Run the code on Aristotle on a mixture of datasets known to contain transient resources and on PALFA blind search datasets for diagnostic purposes. Share lessons learned that should be generalizable to PSRFITS (a standard format for pulsar data files).
5. **Water Resource Management Using OpenMORDM** (CU Reed). Continue development of the Aristotle MPI cluster, i.e., MPI in a container. Run 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** (CU Douglas). Conduct a detailed manual curation of the 5 bacterial models, prior to the analysis of among-partner transport fluxes to generate a single multi-compartment model. Generate the 6th metabolic sub-model using the previously analyzed gut transcriptome data and format (Bost et al.). Verify that SteadyCom can run on three-compartment models already reconstructed in the laboratory. Use these models to (1) identify community metabolism, i.e., outputs of the co-metabolism of two (or more) partners in the association that cannot be generated by any individual organism, and the metabolic conditions that favor/disfavor community metabolism, (2) identify condition-dependent forbidden spaces (infeasible reactions) associated with each sub-model and the integrated multi-compartment model.
7. **Multi-Sourced Data Analytics to Improve Food Production and Security** (Sedgwick Reserve McCurdy/UCSB). Improve the accuracy of image classification and explore a collaboration with

Zooniverse or AWS to engage citizen scientists in quarterly deer surveys and validation for the automatic identification of species using the *"Where's the Bear?"* software infrastructure. Compare Google, AWS, and Aristotle performance in animal trap image identification using the *"Where's the Bear?"* application. Use SmartFarm IoT technology and the Aristotle cloud to implement more efficient, automatic irrigation scheduling based on real-time moisture sensing. Analyze the results of citrus experiments (e.g., pesticides, water usage, genetic grafting, etc.) at the Lindcove Research Extension Center packline with IoT and Aristotle and deliver the data results (i.e., the positive or negative impact on the fruit) to test orchards in real time to speed up decisions made by growers in the field. Get the Centaurus clustering system up and running on Aristotle and Jetstream so that growers and farm consultant can use it for soil analysis.

**Emerging Technology Plans**

**Containers:** An NSF supplemental award was granted to transition to OpenStack and to (1) explore the containerization of scientifically important benchmarks, (2) explore the use of containers by the Aristotle science teams to achieve cross-cloud deployment portability. The Aristotle team will identify application kernels from XDMoD for containerization and test them on the federated cloud resources in order to demonstrate functionality and performance on multiple systems. The team will also explore containerization, container orchestration, and the use of Docker for Aristotle science projects (some of which are already in production on Aristotle and have been helpful in, for example, reproducing complex builds). In addition, investigation of MPI performance within a container (i.e., network, hardware, distributed storage access issues, etc.) will continue for embarrassingly parallel and tightly coupled use cases.

**DrAFTS:** The Aristotle team plans to continue the testing and optimization of DrAFTS. The system was developed under Aristotle funding to predict the "bid price" that an AWS user should bid in the spot market to ensure a minimum duration of execution before AWS terminates the instance. DrAFTS will be the basis for cloud metrics in the federation and will be integrated with Open XDMoD.

**Plans to Disseminate Results**

The Aristotle team will continue to engage the cyberinfrastructure community through presentations and dialogue at CASC, SC17 and PEARC '18, and scientific meetings. The 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 guides.

Back to the top

## Products

### Books

Paul Redfern, David Lifka, Duncan Brown, Stephen Ficklin, Ken Koedinger, Kristin Persson, Linda Schadler, and Carol Song (2017). *Final Report: 1st NSF Data Infrastructure Building Blocks PI Workshop* Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

### Book Chapters

### Inventions

### Journals or Juried Conference Papers

Andy Rosales Elias, Nevena Golubovic, Chandra Krintz, and Rich Wolski (2017). Where's the Bear? – Automating wildlife image processing using IoT and edge cloud systems.. *Juried Conference Paper*. 247. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; ISBN: 978-1-4503-4966-6

Hiranya Jayathilaka, Chandra Krintz, and Rich Wolski (2017). Performance monitoring and root cause analysis for cloud-hosted web applications. *Juried Conference Paper*. 469. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1145/3038912.3052649

Nana Y.D. Ankrah, Junbo Luan, and Angela E. Douglas (2017). Cooperative metabolism in a three-partner insect-bacterial symbiosis revealed by metabolic modeling. *Journal*. 199 (15), e00872-16. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1128/JB.00872-16

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

Stratos Dimopoulos, Chandra Krintz, and Rich Wolski (2017). PYTHIA: Admission control for multi-framework, deadline driven, big data workloads. *Juried Conference Paper*. . Status = AWAITING_PUBLICATION; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

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, and Tristan Shepherd (2018). *Assessing the fidelity of the North American wind climate and impacts of wind farms using high resolution modeling*. 98th American Meteorological Society Annual Meeting (21st Conference on Planned and Inadvertent Weather Modification). Austin, TX. Status = SUBMITTED; 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, and Rich Wolski (2017). *Challenges: DIBBs for multi-campus cyberinfrastructure through cloud federation*. 1st NSF Data Infrastructure Building Blocks PI Workshop (DIBBs 17). 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

Steven Lee (2017). *Cornell Red Cloud: Campus-based hybrid cloud computing*. 2017 NSF Cybersecurity Summit for Large Facilities and Cyberinfrastructure. Arlington, VA. 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 = SUBMITTED; Acknowledgement of Federal Support = Yes

Sara C. Pryor, Rebecca J. Barthelmie, and 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 = SUBMITTED; 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 Workshops. Kansas City, KS. Status = OTHER; Acknowledgement of Federal Support = Yes

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

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

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

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 = UNDER_REVIEW; Acknowledgement of Federal Support = Yes

### Other Products

*Audio or Video Products.*
Aristotle geo data use case scientist Varun Chandola (University at Buffalo) produced a video that demonstrates how webGlobe allows access and visualization of remotely available NetCDF data. The key capability of webGlobe is under the hood. To allow the visualization and analysis of a wide variety of geospatial data sets, webGlobe has decoupled client-server architecture. Watch the demonstration at https://www.youtube.com/watch?v=-jhCwOda9fU.

*Audio or Video Products.*
Professor Thomas Bosch, Cell and Developmental Biology, Kiel University led a 2017 interview with Aristotle use case scientist Angela Douglas (Cornell), on how her studies of insects help to understand how host-microbe interactions affect metabolism and nutrition. Her aim is to achieve biomedical models for human health and a novel target for insect pest control. See the interview at https://www.youtube.com/watch?v=jup7FXzw7Tw.

*Poster.*
Nana Ankrah, a postdoc in Douglas lab, won the poster prize at the Copenhagen Bioscience Conference in 2017 and acknowledged Aristotle science team staff member Dr. Brandon Barker of the Cornell Center for Advanced Computing and the Aristotle Cloud Federation. See the prize winning poster at http://angeladouglaslab.com/nana-wins-conference-prize/.

### Other Publications

### Patents

### Technologiesor Techniques

DrAFTS (Durability Agreements From Time Series for AWS Spot Instances) was developed under Aristotle funding. Using historical pricing data from AWS and Quantile Bounds Estimation from Time Series (QBETS)—a non-parametric time series forecasting methodology for predicting bounds on future value—DrAFTS will provide predicted bid pricing (in US dollars) and time durations (in hours) with probabilistic guarantees. After beta testing, DrAFTS will be publicly released via the DrAFTS website and Aristotle portal. The plan is for DrAFTS to be integrated with Open XDMoD in order to provide cloud metrics to Aristotle users so that they know when to run where (which academic site) as well as when to optimally run at AWS or NSF clouds.

### Thesis/Dissertations

### Websites

*Aristotle Cloud Federation*
https://federatedcloud.org
The Aristotle Cloud Federation portal was updated regularly to feature new web content such as science and technology news, events, a user guide, use case information, and the latest project reports.

Back to the top

## 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? | 6 |
| How many REU applicants were selected and agreed to participate during this reporting period? | 6 |
| 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@ccr.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

Back to the top

## 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, the University at Buffalo, and the UC Santa Barbara are exploring how cloud resources can be effectively shared between campuses and their impact on researchers requiring flexible workflows and analysis tools for large-scale data sets. The project serves as an important model for campus cyberinfrastructure that others may follow providing elasticity by sharing resources and an allocation model that provides a fair exchange mechanism for resources access between and across multiple 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 transparently moving 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 AWS has the potential to impact research productivity.

**What is the impact on the development of human resources?**

Aristotle is pioneering the concept of federated cloud computing 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 2017. The availability of campus and hybrid cloud computing may also spurn the development and dissemination of ready-to-launch VMs 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.

**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.

**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, (3) access to multiple data sources, many of which are already in public and private clouds.

**What is the impact on technology transfer?**

DrAFTS cloud metrics are likely to make an impact on Globus Genomics’ ability to cost effectively deliver software-as-a-service subscriptions for Next Generation Sequencing (NGS) data analysis.

**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 with on demand soil, water, and crop sensor data.

Back to the top

## 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.

Back to the top