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

Award No. 1541215

PI: David Lifka, Cornell University; co-PIs: Thomas Furlani, University at Buffalo; Rich Wolski, UC Santa Barbara (2015-2020)

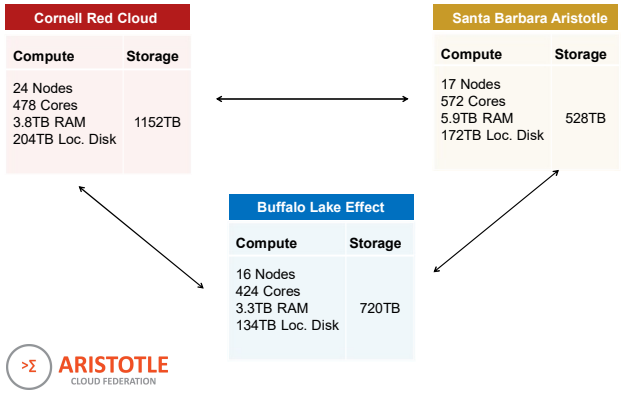
federatedcloud.org

Goals

- Implement a scalable and sustainable multi-institutional cloud federation model that provides DIBBs in support of research disciplines requiring flexible workflows and data analysis tools.
- Support seven intentionally-diverse strategic science use cases.
- Containerize use cases and move containerized applications between institutions transparently.
- Encourage and reward data analysis resource sharing with a new allocations and accounting model that provides a fair exchange mechanism for resource access between multiple institutions.
- Develop Open XDMoD cloud accounting and metrics to make online forecasts of future performance and allocation levels.
- Burst to NSF cloud and public cloud during peak usage.

New Capabilities

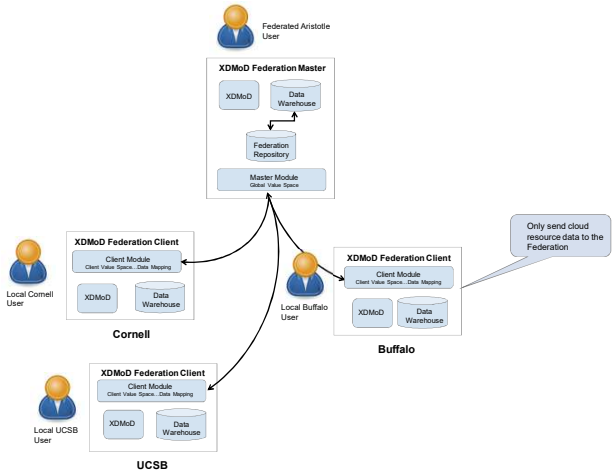
- Deployed PY3 hardware at the 3 federation sites: Cornell (CU), University at Buffalo (UB), and UC Santa Barbara (UCSB). Hardware will be augmented each year of the grant, including Ceph storage.



- Cornell and UB installed more servers to add capacity because their clouds were running at 90% utilization. UCSB added capacity as well.
- All sites transitioned their infrastructure from Eucalyptus to the OpenStack platform and are optimizing their production environments. The ability of federation sites to quickly learn from one another rather than going it alone is a major benefit of participating in a federation.
- Red Hat is the latest company to join the Aristotle project. Current partners include AWS, Globus, Dell, and HPE. Red Hat's OpenStack and Ceph expertise will enhance knowledge across the federation.

Innovative Technologies

- UB developed a federated version of Open XDMoD and is integrating initial cloud metrics into it, to be followed by more complex metrics such as UCSB's DrAFTS (Durability Agreements from Time Series – a new prediction technology). Data from all 3 federation sites—Cornell, Buffalo, and UCSB—will be available at the XDMoD Federation Master instance.



- Cornell built a new federated cloud accounting database encompassing resources, research teams, allocations, usage by site, and, in the future, exchange rates across all sites in the federation. A dashboard in the user portal will enable signed-in project members to manage their team members, view their allocation balance, view usage across sites, access all federation sites with a single sign-on, see Open XDMoD metrics, and see availability across all federated cloud sites.
- Other innovations are documented in more than 35 publications and conference presentations available at the Aristotle portal (federatedcloud.org). In addition, 11 NSF Research Experiences for Undergraduates (REU) students contributed to Aristotle technology developments and science use case progress while gaining valuable experience processing large-scale data, using machine learning algorithms, applying statistical metrics, and learning Docker.
- Future aspirations include the possible development of an open marketplace where cloud resources (campus clouds, NSF cloud, and public cloud) could be readily accessed with tools to compare, select, and request the most appropriate resource. An open cloud marketplace and ecosystem could accelerate academic research adoption of private and public cloud, and reduce time to science.

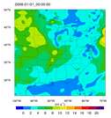
Advances in Science

A Cloud-Based Framework for Visualization and Analysis of Big Geo Data (V. Chandola et al.) – Completed 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.

This framework allows scientists to load, visualize, and analyze NetCDF data sets and is, at present, the only browser-based system available with this functionality. A demo is available at the Aristotle portal.

Mapping Transcriptome Data to Metabolic Models of Gut Microbiota (A. Douglas, N. Ankrah, J. Luan, B. Barker et al.) – CU researchers developed a computational model for the whiteflies system, and analyzed and compared three independently-evolved communities in xylem-feeding insects (*Journal of Bacteriology* & *Copenhagen Bioscience Conference* poster winner). Also, generated advanced draft metabolic reconstructions for 5 bacteria needed to simulate *Drosophila*-microbial community metabolic interactions.

High Fidelity Modeling and Analytics for Improved Understanding of Climate (S. Pryor et al.) – CU completed simulations to test the sensitivity of the climate impacts to the precise description of the wind turbine aerodynamics. A better understanding of the impact of wind turbine deployments on climate was achieved.



Multi-Sourced Data Analytics to Improve Food Production and Security (C. Krintz, K. McCurdy, R. Wolski et al.) – Soil moisture monitoring of California almond trees features a multi-scale IoT infrastructure that uses solar panel powered sensors to measure soil moisture and temperature at different depths. The analytics will result in spatial cluster in soil and differential meteorology for frost prevention. A second UCSB project is designed to measure and analyze nighttime air temperatures at a citrus orchard. Previously, researchers used a new, non-parametric time series analysis technique to analyze soil moisture data on Aristotle and discovered that grape crops were being overwatered by 66%.

Where's the Bear? uses IoT sensors (cameras), an edge cloud, and the Aristotle back-end cloud or public cloud to automatically and accurately classify over 200,000 animal images per month. See the *IoTDI Journal* publication and video at federatedcloud.org.

2nd NSF Data Infrastructure Building Blocks Workshop (DIBBs18), July 9-10, 2018, Arlington, VA