Building a Federated Cloud Model: Aristotle Cloud Federation

David Lifka, PI
Director, Cornell University Center for Advanced Computing (CAC)
CIO & Vice President, Cornell Information Technologies
lifka@cornell.edu

Thomas Furlani, Co-PI
Director, Center for Computational Research (CCR), University at Buffalo
furlani@buffalo.edu

Rich Wolski, Co-PI
Professor, Computer Science, University of California, Santa Barbara
Co-Founder, Eucalyptus Systems
rich@cs.ucsb.edu
Aristotle Overview & Goals

• $5 million, 5-year NSF award to deploy a federated DIBBs cloud at Cornell, U. Buffalo, UC Santa Barbara
  – Award made October 2015
  – 7 science teams, over 40 global collaborators
    • Requiring flexible workflows/analysis tools for large-scale data
    • Representing diversity of analysis requirements and cloud usage modalities
      – Earth and atmospheric sciences, finance, chemistry, astronomy, civil engineering, genomics, food science

• Project goals
  – Optimize “time to science”
  – Demonstrate the value of sharing resources and data across institutional boundaries
Project Structure

Cloud Federation Project Management, Oversight & Reporting

- DIBBs Acquisition, Installation, Configuration, Testing & Maintenance
- Cloud Federation Portal
- Research Team Support

- Portal Development & Maintenance
- Allocations & Accounting Integration
- Open XDMoD Integration
- Eucalyptus & QBETS Integration
Aristotle “Components”

• Federated cloud
• Ability to burst to public clouds
  – Determine optimal spot pricing with high guarantee of service
• Aristotle Portal
• XDMoD integration
• QBETS for optimizing “Time to Science”
Federated Cloud

- Multi-Disciplinary Collaborations in a Multi-Campus Cloud Federation
Aristotle Portal

• Single source for webinars, documentation, training, research project tools, help desk, system and status information, and metrics from all federation partners.
• Do it yourself cloud instructions
• Coming soon! https://federatedcloud.org/
Open XDMoD

- Comprehensive management of HPC/Cloud resources
  - Cloud usage and VM performance metrics
    - Monitor how well VM is utilized
- Federate Open XDMoD
  - Local and **global** view of federated cloud
QBETS - Quantile Bounds Estimation from Time Series

- Originally developed to predict bounds on the delay in NSF supercomputer batch queues.
- Also used to predict future load levels and the AWS spot market.
- Cloud federation user will be able to determine optimal time (in terms of wait time) to burst to the public cloud and the best price-performance when bursting is warranted.
- QBETS service will be available via the federation portal.
- Integrate its predictive capabilities with XDMoD metrics to generate statistical bounds on guaranteed delivered performance levels.
Current Cloud Infrastructure

• Cornell Red Cloud in production for 5 years
  – Red Cloud Eucalyptus upgraded to 4.2
  – Federation capabilities tested between Ithaca and WCM NYC
  – Current Aristotle portal site running in Red Cloud

• UB Lake Effect cloud installed/operational 10/16/15
  – Lake Effect Eucalyptus upgraded to 4.2
  – Integrating into existing UB core services
  – Leveraging Euca Availability Zones
  – May 2016 Aristotle DIBBs Cloud nearly in production

• UCSB Engineering Computing Infrastructure cloud in production for 2 years
  – UCSB Eucalyptus cloud upgraded to 4.2
  – Working on InCommon setup
  – First installment of DIBBs infrastructure being installed
Cornell Red Cloud

• One Campus/Aristotle Cloud

Diagram:
- Public Internet
- Eucalyptus Cloud Controller and Walrus Server (CLC/walrus)
- Eucalyptus Cluster and Storage Controller (CC/SC)
- Equallogic SAN
- Node Controller
- Node Controller
- Node Controller
- Node Controller
Buffalo Cloud

- Integration of Lake Effect and Aristotle Clouds
Global Picture of Aristotle Usage & Performance

- Federating XDMoD to collect usage and performance data for each cloud instance
- Application beyond Aristotle
Research Teams – Optimizing Time to Science

• 7 initial science teams
  – Use Case 1: A Cloud-Based Framework for Visualization & Analysis of Big Geospatial Data (UB)
  – Use Case 2: Global Market Efficiency Impact (UB)
  – Use Case 3: High Fidelity Modeling and Analytics for Improved Understanding of Climate-Relevant Aerosol Properties (CU)
  – Use Case 4: Transient Detection in Radio Astronomy Search Data (CU)
  – Use Case 5: Water Resource Management Using OpenMORDM (CU)
  – Use Case 6: Mapping Transcriptome Data to Metabolic Models of Gut Microbiota (CU)
  – Use Case 7: Multi-Sourced Data Analytics to Improve Food Production (UCSB)

• Science team meetings started from day 1
  – Understand current workflows and processes
  – Document software and hardware requirements
  – Develop cloud adoption strategy
  – Start creating VMs (e.g., geospatial analytics, finance analytics, Windows instance for MATLAB VMs already completed)
  – Demo proof of concepts to science teams
Use Case 1: A Cloud-Based Framework for Visualization and Analysis of Big Geospatial Data (UB - Chandola)

• **What?**
  – Integrate, visualize and analyze heterogeneous geospatial data from around the world

• **Why Aristotle?**
  – Offload variable computational demands from local infrastructure
  – Create an expert service hub for widely distributed data
  – Scalable and elastic computing
webGlobe: Big Geospatial Analytics on the Cloud*

* Dinh Tran, Er Ran Khoo and Varun Chandola
Computer Science and Engineering
University at Buffalo

webGlobe

• Enables access to a wealth of scientific data
  – Climate/weather simulations
  – Remote sensing and other observed data
  – Other geospatial data products

• Integrated and interactive analysis and visualization
  – Enabled by Aristotle Cloud
webGlobe Analysis Stack on Aristotle Cloud

- Allows dynamical provisioning of a virtual Spark cluster for analytics and visualization
- Enables parallel access to underlying massive simulation outputs
  - A novel spatial partitioning strategy

Try here: http://bit.ly/1XpFFCy
Use Case 2: Global Market Efficiency Impact (UB- Tin, Roesch, Wolfe)

• **What?**
  – Measure market efficiencies with high volumes of historical and current intraday market data
  – Difficult to judge whether shares (e.g., Facebook) are fairly priced
  – Difficult because no one knows the true value of company (FB) assets
  – Measure share prices in different markets across the world
  – Deviations in price indicate inefficiency in pricing share
  – Want to explain/understand this

• **Why Aristotle?**
  – Share expensive pricing database with researchers in Aristotle Consortium
  – Researchers at UB (own database) get “free” cycles on other members’ clouds
  – Burst to integrate new data
  – Share data and pipelines around the world for collaboration and reproducibility
Use Case 3: High-Fidelity Modeling and Analytics for Improved Understanding of Climate-Relevant Aerosol Properties (CU - Pryor)

• **What?**
  – Model the behavior of atmospheric aerosol particles to predict and ameliorate climate and human health impacts

• **Why Aristotle?**
  – Burst to process new data
  – Share a high-value processed dataset of general interest
  – Build an extensible toolkit for researchers, transferrable to other groups
Use Case 4: Transient Detection in Radio Astronomy Search Data (CU - Cordes)

• **What?**
  – Low-latency/on-demand search for time-varying, including one-time, signals in radio astronomy data
  – Build on existing Cornell/CAC collaboration*

• **Why Aristotle?**
  – Burst on-demand to analyze new data
  – High-speed transport of very large data sets into and around the federation

Use Case 5: Water Resource Management Using OpenMORDM (CU - Reed)

• What?
  – Simulate complex water resource management outcomes on-demand to achieve substantial beneficial environmental impact on municipalities located throughout the US

• Why Aristotle?
  – Start coarsely parallel computations on-demand
  – Very large computations, ideal for transfer to AWS or other clouds
  – Make available to many municipalities
Use Case 6: Mapping Transcriptome Data to Metabolic Models of Gut Microbiota (CU - Douglas)

• **What?**
  - Use host and gut-microbial expression data to infer which metabolic reactions occur in each organism as a result of dietary inputs

• **Why Aristotle?**
  - Heterogeneous instance types and sizes with high availability to allow for unpredictable computational demands
  - Allows generation of *reproducible computational biology pipelines* in the form of VMs or VM configurations
Use Case 7: Multi-Sourced Data Analytics to Improve Food Production (UCSB/Sedgwick Reserve - McCurdy)

• **What?**
  – Interdisciplinary analysis of data from a wide variety of sources (including the general public) to study and improve ecological outcomes

• **Why Aristotle?**
  – Scalable and portable infrastructure
  – Access to multiple data sources, many of which are already in public and private clouds
Sharing and Working with the Community

• We’re developing and identifying cloud-related tools
  – QBETS
  – Open XDMoD with cloud support
    http://xdmod.sourceforge.net/
  – CloudLaunch (potential work with Cycle Computing and AWS)
    https://www.cac.cornell.edu/technologies/CloudLaunch.pdf
  – Supercloud (testing on Jetstream)

• We want to create a "Cloudy" community focus group
  – Identify and share cloud tools, services, and solutions
  – Leverage new computing and scaling paradigms (e.g., containerization)
  – Help each other and the greater research community achieve faster time to science