Florida International University

OPEN SCIENCE DATA CLOUD

PARTNERSHIP FOR **INTERNATIONAL RESEARCH** AND EDUCATION

UCLA, 13th February, 2014
4 pm

Heidi Alvarez, PhD
Center for Internet Augmented Research and Assessment (CIARA)

Award #1129076

Open Science Data Cloud

- The Open Science Data Cloud (OSDC) is an open-source, cloud-based infrastructure that allows scientists to manage, analyze, integrate and share medium to large size scientific datasets.

- Scientists from all fields can use the OSDC resources for managing large sets of data.

NSF Partnerships for International Research and Education

Award #1129076

THE UNIVERSITY OF CHICAGO

UIC

FIU FLORIDA INTERNATIONAL UNIVERSITY

CIARA

OPEN CLOUD CONSORTIUM

Open Science Data Cloud overview

- OSDC is a Science Cloud Service Provider (CSP), operated by not-for-profit Open Cloud Consortium

- OSDC is a 6 PB / 9,000 core science cloud with 5 PB of usable storage (1 PB science data for the research community, 1 PB of biomedical data for medical research)

- We have been doubling in size each year

- We run production services for NASA and NIH researchers

- Typical job 1000s of core hours over 10-100's TB
- An open-source cloud infrastructure for supporting scientific research having moderate to large computation and storage requirements

- Split into public and protected infrastructures

- Connected to high performance research networks, open to anyone doing scientific research

Award #1129076

OSDC Data Centers and Networks

- We have three data centers
    - Chicago with 100G to StarLight
    - FIU with 10G to StarLight
    - Livermore Valley Open Campus 10G to StarLight

- We're planning one more data center with 100G connection to StarLight

- We are looking to interoperate the OSDC with international partners over 10G and 100G networks

NSF PIRE Partnerships for International Research and Education

THE UNIVERSITY OF CHICAGO

UIC

FIU FLORIDA INTERNATIONAL UNIVERSITY

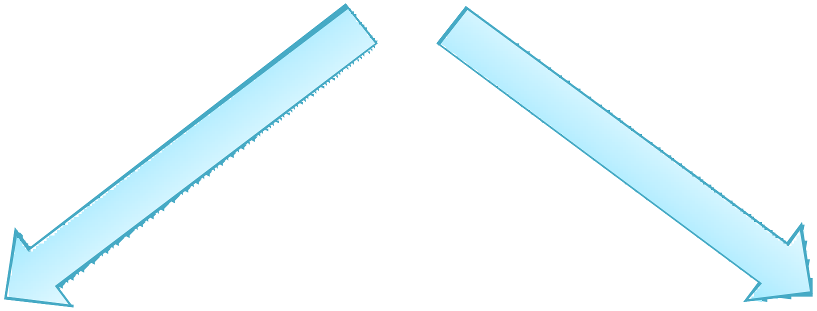
CIARA

OPEN CLOUD CONSORTIUM

Award #1129076

OSDC OPEN SCIENCE DATA CLOUD

Cloud Services for the Scientific Community



Science Cloud

Open Science Data Cloud

- Earth sciences
- Biological sciences
- Social sciences
- Digital humanities
- ACL, groups, etc.

Biomedical Cloud

PROTECTED DATA CLOUD

Designed to hold Protected Health Information (PHI) e.g. genomic data, electronic medical records, etc. (HIPAA, FISMA)

NSF PIRE Partnerships for International Research and Education

Award #1129076

THE UNIVERSITY OF CHICAGO

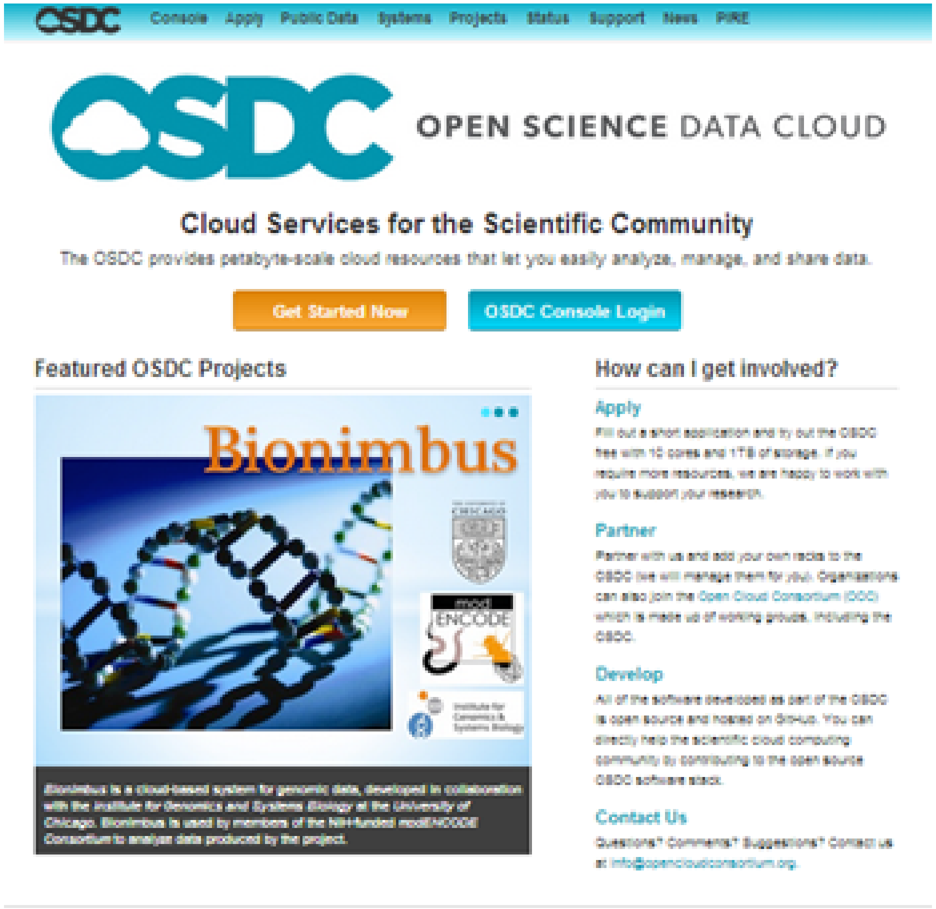
UIC

FIU FLORIDA INTERNATIONAL UNIVERSITY

CIVARA

OCC OPEN CLOUD CONSORTIUM

Current Apps on the OSDC



Cloud Services for the Scientific Community

The OSDC provides petabyte-scale cloud resources that let you easily analyze, manage, and share data.

Featured OSDC Projects

Bionimbus is a cloud-based system for genomic data, developed in collaboration with the Institute for Genomics and Systems Biology at the University of Chicago. Bionimbus is used by members of the NIH-funded modENCODE Consortium to analyze data produced by the project.

How can I get involved?

Apply

Fill out a short application and try out the OSDC free with 16 cores and 1TB of storage. If you require more resources, we are happy to work with you to support your research.

Partner

Partner with us and add your own racks to the OSDC (we will manage them for you). Organizations can also join the Open Cloud Consortium (OCC) which is made up of working groups, including the OSDC.

Develop

All of the software developed as part of the OSDC is open source and hosted on GitHub. You can directly help the scientific cloud computing community by contributing to the open source OSDC software stack.

Contact Us

Questions? Comments? Suggestions? Contact us at info@opencloudconsortium.org.

Biological Data

- **Bionimbus** is a cloud-based system for managing, analyzing and sharing genomic data, including data from next-generation sequencing devices(last publication)

- **modENCODE** The National Human Genome Research Institute (NHGRI) model organism ENCyclopedia Of DNA Elements (modENCODE).

NSF Partnerships for International Research and Education

Award #1129076

THE UNIVERSITY OF CHICAGO UIC FIU FLORIDA INTERNATIONAL UNIVERSITY CIVARA OPEN CLOUD CONSORTIUM

Current Apps on the OSDC

Earth Science Data

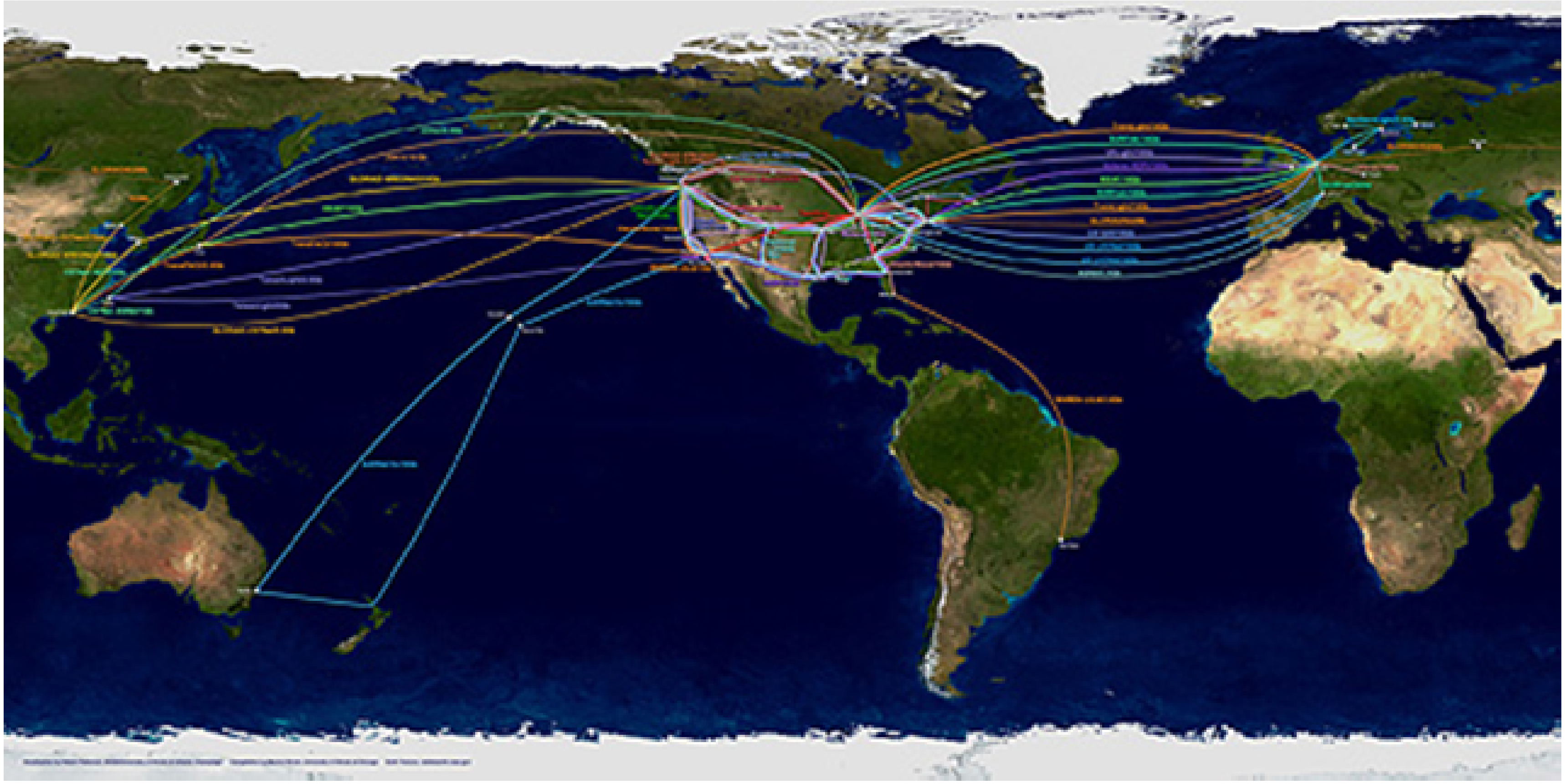
- **Project Matsu** is a collaboration with NASA. The OSDC is used to process Earth Observing 1 (EO-1) ***satellite imagery*** from the Advanced Land Imager and the Hyperion instruments and to make this data available to interested users.

Digital Humanities

- OSDC supports **Bookworm** from Harvard's Cultural Observatory and offers a way to interact with digitized book content and full text search. Bookworm uses ngrams extracted from books in the public domain and integrates library metadata, including genre, author information, publication place and date.
- ***Other Projects for public data***

Award #1129076

OSDC Network Connections



Partnerships for International Research and Education

Award #1129076

THE UNIVERSITY OF CHICAGO

UIC

FIU FLORIDA INTERNATIONAL UNIVERSITY

CIARA

OPEN CLOUD CONSORTIUM

Current State of OSDC

- An installation of petabyte scale provides storage for medium to large datasets in Hadoop or Sector large data clouds. (live status)

- Supports elastic, on demand virtual machines, similar to Amazon's EC2 service.

- Designed for long term persistent storage for scientific datasets, Nominate your favorite data set

- Utilizes high performance research networks, so datasets can be easily ingested, accessed, and shared over wide areas.

- Balanced architecture that uses data locality to support the efficient execution of queries and analysis over the data managed by the cloud.

Award #1129076

Operating System & Architecture

- Compute infrastructure based on Open Stack
- Production currently running Essex, roadmap for upgrading to Havana
- Storage infrastructure based on GlusterFS
- Currently do not use object (Swift) or block (Cinder) storage
- Hadoop clusters (Hadoop/MapReduce), R, Samba , Nagio ,OpsCode /Chef
- Databases -MySQL, PostgreSQL and SQL Server
- Manages VM images and instances, distributed file system mounted directly in VMs

View OSDC poster | Tukay poster | UDR poster | Matsu poster

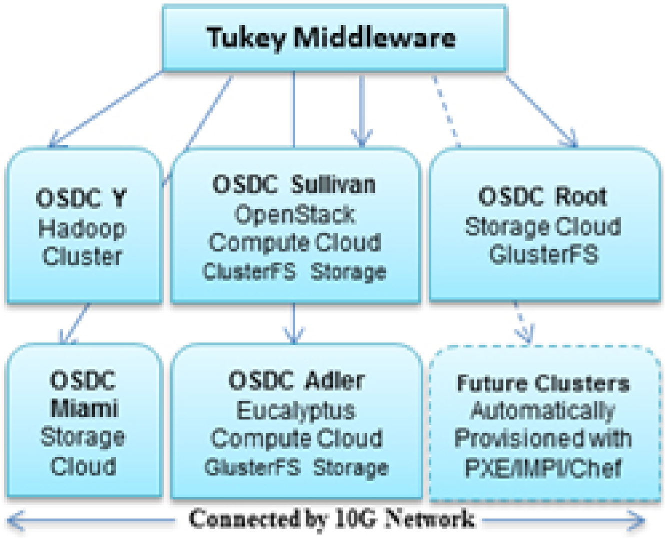
Award #1129076

OSDC Clusters

The OSDC is a distributed facility connected by 10G or greater networks so high speed transport protocols are important for enabling users to import / export data and to move data around flexibly in their analysis processes.

- UDT protocol-reliable UDP based application level data transport *protocol* for distributed data intensive applications over wide area high-speed networks

- UDR OSDC Backup Protocol – Experiment to test between Chicago, Miami, Sao Paulo underway


Tukey Middleware

OSDC Y
Hadoop
Cluster

OSDC Sullivan
OpenStack
Compute Cloud
ClusterFS Storage

OSDC Root
Storage Cloud
GlusterFS

OSDC
Miami
Storage
Cloud

OSDC Adler
Eucalyptus
Compute Cloud
GlusterFS Storage

Future Clusters
Automatically
Provisioned with
PXE/IMPI/Chef

Connected by 10G Network

NSF
Partnerships for International Research and Education

Award #1129076

THE UNIVERSITY OF CHICAGO

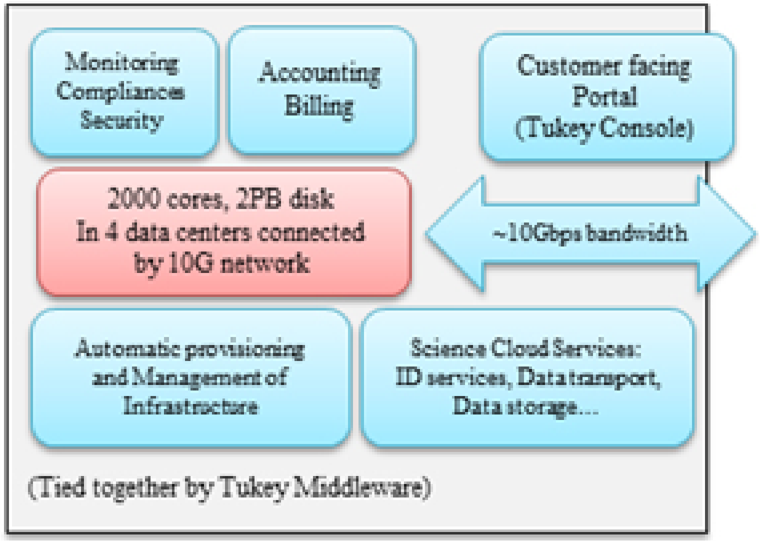
UIC

FIU
FLORIDA INTERNATIONAL UNIVERSITY

CIVARA

OPEN CLOUD CONSORTIUM

OSDC User Services



Monitoring
Compliances
Security

Accounting
Billing

Customer facing
Portal
(Tukey Console)

2000 cores, 2PB disk
In 4 data centers connected
by 10G network

~10Gbps bandwidth

Automatic provisioning
and Management of
Infrastructure

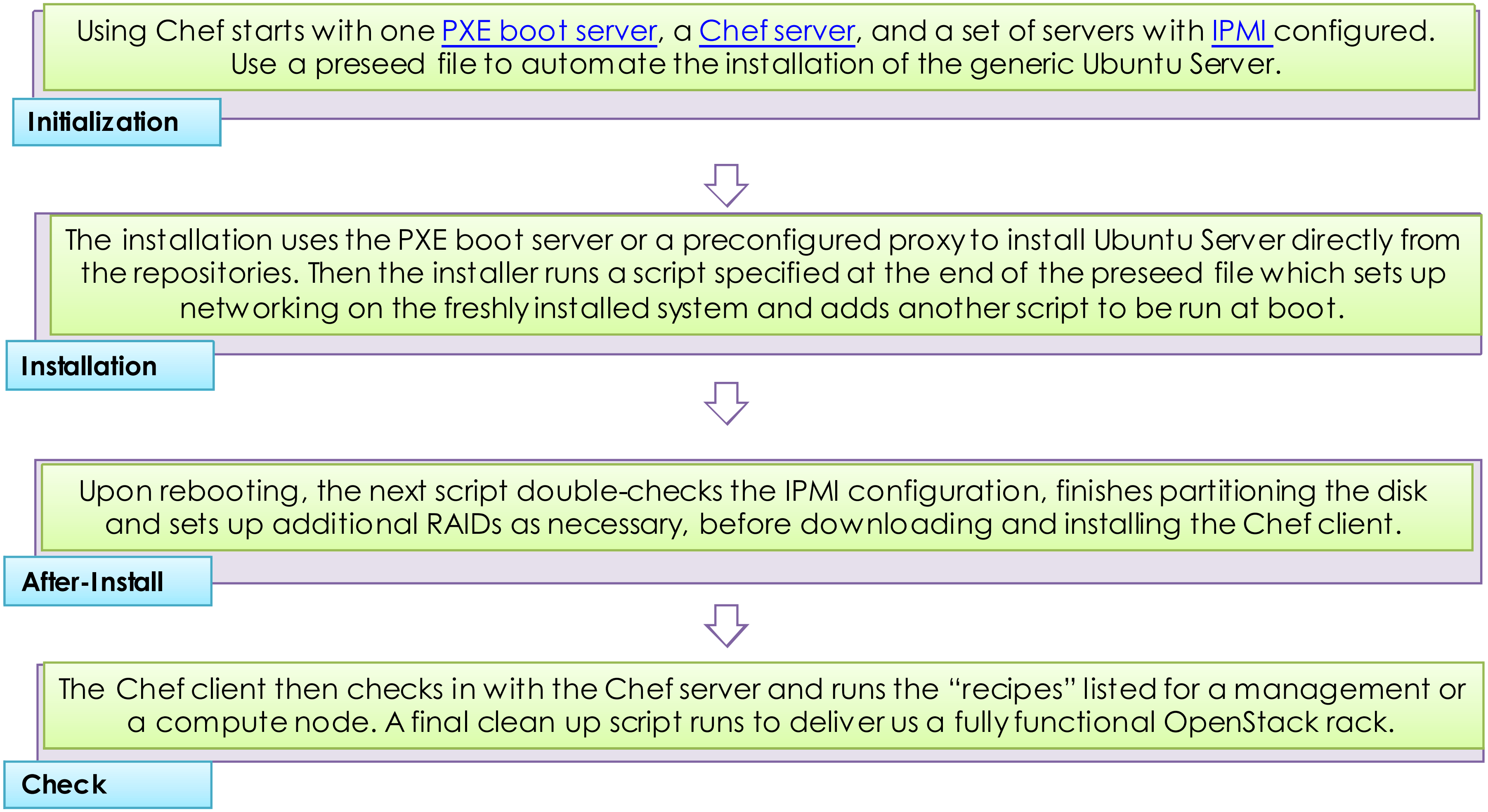
Science Cloud Services:
ID services, Data transport,
Data storage…

(Tied together by Tukey Middleware)

The OSDC user services include the ability to provision virtual machines, access usage and billing information, share files, and access to a key service and public datasets. All of the OSDC user services are tied together through a customizable web application "Tukey" and middleware that enables uniform access to the cloud services by the "Tukey Console" web application.

Award #1129076

From bare metal to a compute or storage cloud…



Using Chef starts with one PXE boot server, a Chef server, and a set of servers with IPMI configured. Use a preseed file to automate the installation of the generic Ubuntu Server.

Initialization

The installation uses the PXE boot server or a preconfigured proxy to install Ubuntu Server directly from the repositories. Then the installer runs a script specified at the end of the preseed file which sets up networking on the freshly installed system and adds another script to be run at boot.

Installation

Upon rebooting, the next script double-checks the IPMI configuration, finishes partitioning the disk and sets up additional RAIDs as necessary, before downloading and installing the Chef client.

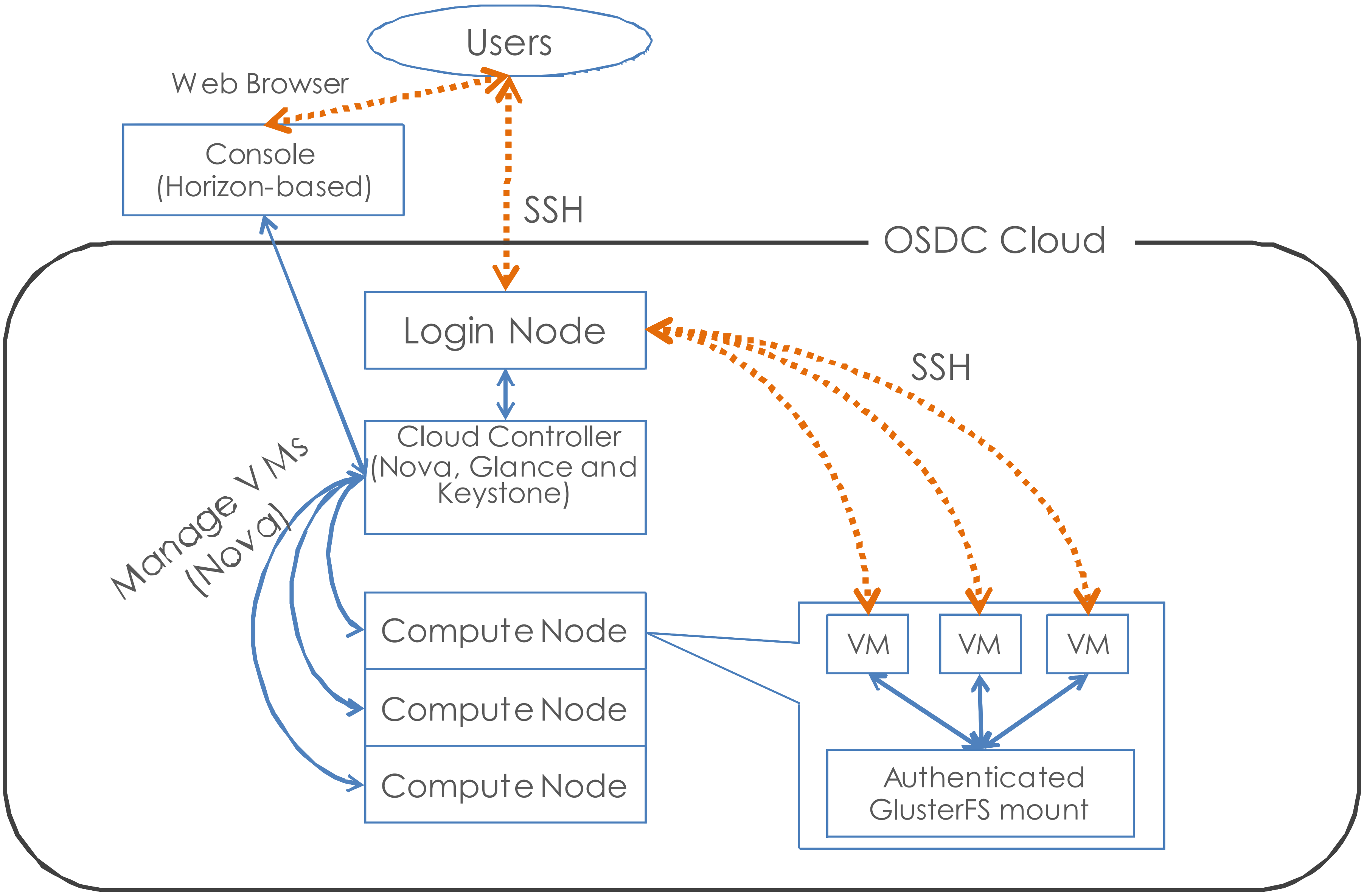
After-Install

The Chef client then checks in with the Chef server and runs the “recipes” listed for a management or a compute node. A final clean up script runs to deliver us a fully functional OpenStack rack.

Check

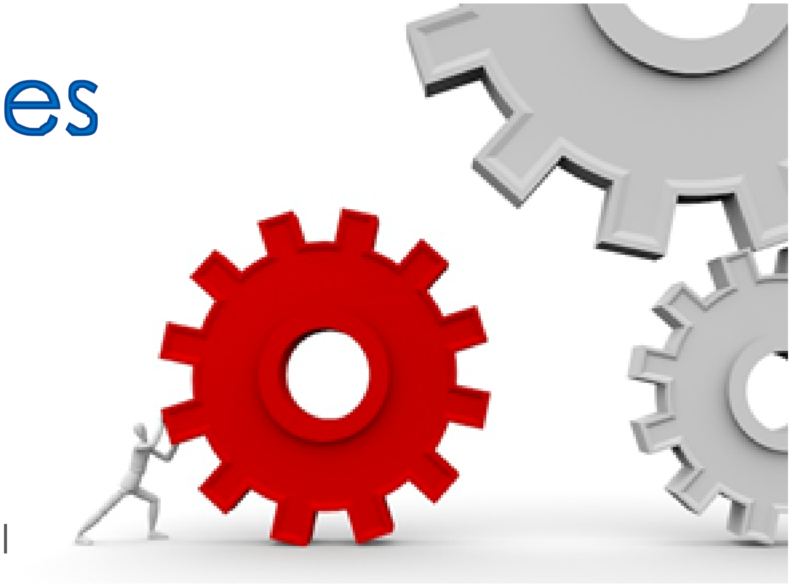
Award #1129076

OSDC A Few More Details



Award #1129076

Challenges



We are focusing on the following:

- How do we authenticate, authorize and provide access controls to researchers at our international partners to data and to cloud based services(storage and compute)

- We need open source implementations of these services

- We need trust relationships with our peers

- We are running a series of interoperability workshops to try to get this right.

Award #1129076

What You Get with the OSDC

- Login with your university credentials via InCommon
- Launch virtual machines, virtual clusters, access to large Hadoop clusters, etc.
- Access PB+ of open and protected data
- Manage files, collections of files, collections of collections
- Manage users, groups of users
- Manage accounts, sub-accounts
- Efficient transfer of large data (UDT, UDR)
- www.opensciencedatacloud.org/apply

Award #1129076

PIRE OSDC OPEN SCIENCE DATA CLOUD

- National Science Foundation Partnership for International Research and Education 5 year program 2010 – 2014 at $3.5M.
- Prepares students to compete in the global cyberinfrastructure community
- Provides **international research and education experiences** around the world!
- The student/faculty/scientist research teams help develop large-scale distributed computing capabilities data and, State-of-the-art services for integrating, analyzing, sharing and archiving scientific data.

NSF Partnerships for International Research and Education

Award #1129076

THE UNIVERSITY OF CHICAGO

UIC

FIU FLORIDA INTERNATIONAL UNIVERSITY

CIVARA

OCC OPEN CLOUD CONSORTIUM

PIRE OSDC

OPEN SCIENCE DATA CLOUD

- Students join a **prestigious international research network** focusing on the use of cyberinfrastructure.
- Training for a generation of globally-oriented IT professionals
- Become a leader in industrial and academic workforce.
- Collaborate to solve critical and nationally-important complex scientific problems with faculty scientists
- Provide a major impact on American competitiveness.

NSF Partnerships for International Research and Education

Award #1129076

THE UNIVERSITY OF CHICAGO

UIC

FIU FLORIDA INTERNATIONAL UNIVERSITY

CIVARA

OCC OPEN CLOUD CONSORTIUM

Investigators

PI: Robert Grossman
Institute for Genomics & Systems Biology, University of Chicago

Co-PI: Heidi Alvarez
CIARA, Florida International University, Outreach Lead

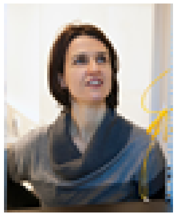
Co-PI: Philip Yu
National Center for Data Mining, University of Illinois at Chicago,
Research Lead

OSDC-PIRE US Collaborators:

Joe Mambretti - StarLight Co-Director, Open Cloud Consortium
Kevin White - Institute of Genomic & Systems Biology, UIC, liaison to Chicago Field Museum

Award #1129076

International Collaborators



Malcolm Atkinson – School of Informatics, Edinburgh University, UK host

Paola Grosso & **Cees de Laat** – Faculty of Science, Informatics Institute, University of Amsterdam

Karen Langona and **Tereza Cristina Carvalho** - LARC – Laboratory of Computer Networks and Architecture at the University of Sao Paulo, Brazil

Satoshi Sekiguchi – National Institute of Advanced Industrial Science and Technology (AIST), Japan

NSF PIRE Partnerships for International Research and Education

Award #1129076

THE UNIVERSITY OF CHICAGO UIC FIU FLORIDA INTERNATIONAL UNIVERSITY CIVARA OCC OPEN CLOUD CONSORTIUM

Objectives

- Study and strengthen storage systems
  - Integrate protocols and support data transport over wide-area, high-performance networks.
- Develop new cloud-based parallel programming frameworks
  - Apply them so that this technology is more broadly available to scientists.
- Increase involvement through workshops for a large variety of scientists & students.

NSF PIRE Partnerships for International Research and Education

Award #1129076

THE UNIVERSITY OF CHICAGO UIC FIU FLORIDA INTERNATIONAL UNIVERSITY CIVARA OPEN CLOUD CONSORTIUM

Objectives

- Train in the basics of cloud computing

- Work to ensure that cloud computing research advances to maximize the manageability and analytical power of the complex datasets unique to each scientific discipline.

- Catalyze a higher level of international engagement in the U.S. science and engineering community through international research and education collaborations.

NSF Partnerships for International Research and Education

Award #1129076

THE UNIVERSITY OF CHICAGO UIC FIU FLORIDA INTERNATIONAL UNIVERSITY CIVARA OPEN CLOUD CONSORTIUM

Research Opportunities

- **What?**
  - Fully Funded Internship, which gives you the chance to participate in sophisticated international research collaborations.
- **When?**
  - Summer of 2014
- **How long?**
  - 6 weeks
- **Where?**
  - At any of our international partners.

Award #1129076

Brazil

LARC USP – Laboratory of Computer Networks and Architecture at the **University of Sao Paulo**, Brazil

Research area

- Using large scale clustering to detect potential fire regions

- PlanetLab is a worldwide project involving over 900 nodes in over 400 different research entities. PlanetLab offers a distributed testbed for development and test of new applications and new Internet protocols.

Projects: RealTime BData

NSF PIRE Partnerships for International Research and Education THE UNIVERSITY OF CHICAGO UIC FIU FLORIDA INTERNATIONAL UNIVERSITY CIARA OCC OPEN CLOUD CONSORTIUM

Award #1129076

Scotland

University of Edinburgh ,UK

Research area

- Tune installation of Sector and Sphere on EDIM1 architecture to join OSDC
- Set up ADMIRE gateways to run DISPEL data intensive workflows and investigate dynamic mapping of DISPEL to the VMs
- Integrate scientific DBMS (RASDAMAN, MonetDB, SciDB) to combine OSDC architectures with advanced scientific DB architectures - potentially several projects
- Investigate mappings between map-reduce formulations of data-intensive tasks with DISPEL formulations of data-intensive tasks
- RAPID portals for OSDC

Projects: ***iRODS***, ***DISPEL***, ***mrBox***

Award #1129076

Japan

National Institute of Advanced Industrial Science and Technology (AIST), Japan

Research area

- FULL RESEARCH: from basic to application, focusing on innovative middleware technologies through interactive pattern acknowledgement, media interactive database search, multi-language adaptation, Geographic Information System usage.

Project: ***GEO Grid***, ***Linked Data***, ***Big Data***, ***Entertainment Computing***, ***Lavatube***, ***Data Mining***

Award #1129076

Netherlands

University of Amsterdam, Faculty of Science, Informatics Institute, Science Park, The Netherlands

UNIVERSITEIT VAN AMSTERDAM

Research area

- Phonebook for data – application (i.e. the phonebook) that allows to locate and retrieve OSDC datasets dynamically by using the most advanced high-speed and optical networks available in the research and academic community.

Project: *Phonebook*

Emerging

Singapore, Taiwan, Zambia ,Trinidad & Tobago

Participant Requirements

- Science and engineering researchers interested in cloud computing are encouraged to apply to be an OSDC PIRE participant.
- Grad Students, post-docs, early career faculty
- Graduating seniors entering a graduate education program
- **Must be either a U.S. citizens or resident**
- Strong academic qualifications & active participation in ongoing research projects have a higher chance of acceptance into the OSDC PIRE program.

Award #1129076

Participant Requirements

- Must be computer savvy
  - Computer Science majors
  - Engineering majors
  - Domain Science (e.g., Physics, Biology, Chemistry, etc.) majors

- Enrollment in at least 1 University credit (Independent study credits available) during summer travel

How to apply

- All applications must be submitted online at http://pire.opensciencedatacloud.org/pire-fellowship/pire-application/
- Application deadlines are flexible until March 31st, 2014 for travel in Summer 2014
- Applicants will be notified beginning April 1st, 2014
- Research and travel arrangements will be made shortly after notifications
- OSDC Workshop in Amsterdam, for all fellows will be June 16-20, 2014 **TBD**
- Contact **pire@opensciencedatacloud.org** or **305-348-4105** for questions**.**

NSF PIRE Partnerships for International Research and Education

Award #1129076

THE UNIVERSITY OF CHICAGO

UIC

FIU FLORIDA INTERNATIONAL UNIVERSITY

CIVARA

OPEN CLOUD CONSORTIUM

Letter of Research Interest

- One page
- Explain what you have been studying / researching in your academic program, what interests you in the the OSDC-PIRE research suite, what makes you an outstanding candidate (special scientific, technological, language skills, etc.)
- Describe your experience with the following topics:
  - Research interests and how computation, data storage, mining and retrieval is important to that research.
  - Experience using applications such as Python, Java, SQL, XML, C++, Perl, PHP and/or JavaScript. (required)
  - Experience developing Web based and client/server applications. (required)
  - Experience developing, implementing, debugging and maintaining applications.
  - Experience with complex problem solving and high technical development and activities.

Award #1129076

OSDC Community of Scholars

- Websites: pire.opensciencedatacloud.org, opensciencedatacloud.org
- Mailing list, Summer Workshops

Award #1129076

Questions?

- Visit the Open Science Data Cloud website for more information on the application, and for a list of important dates http://pire.opensciencedatacloud.org/
- Contact Heidi Alvarez at heidi@fiu.edu with any questions or comments you may have.
- Contact OSDC-PIRE Staff at pire@opensciencedatacloud.org if you have questions regarding the application process, important dates, or any other project specific inquiries.

Award #1129076

THANK YOU!

Robert Grossman, PhD.
robert.grossman@uchicago.edu

Heidi L. Alvarez, PhD.
heidi@fiu.edu

Award #1129076