Who’s who?

- **Investigators**
  - **PI:** Robert Grossman
    Institute for Genomics and 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
Collaborators

**United States:**
- Joe Mambretti – StarLight Co-Director, Open Cloud Consortium
- Kevin White – Institute for Genomics & Systems Biology, UIC, liaison to Chicago Field Museum

**Foreign:**
- Malcolm Atkinson – e-Science Institute, Edinburgh University, UK host
- Esteban Walter Gonzalez Clua – Computer Science Institute, Universidade Federal Fluminense, Brazil
- Cees de Laat – Faculty of Science, Informatics Institute, University of Amsterdam
- Minsun Lee – Supercomputing Center, Korea Institute of Science and Technology Information (KISTI)
- Sergio Novaes – Dept. of Physics, Sao Paulo State University, Brazil
- Satoshi Sekiguchi – National Institute of Advanced Industrial Science and Technology (AIST), Japan
- Chung-I Wu – Beijing Institute of Genomics (BIG), Chinese Academy of Sciences
What is cloud computing?

- Cloud computing involves clusters (the “clouds”) of distributed computers that provide potentially less expensive, more flexible, and more powerful on-demand resources and services over a network, usually the Internet, while providing the scale and the reliability of a data center.
Open Science Data Cloud (OSDC)

- 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.

- The OSDC PIRE project aims to narrow the growing gap between the capability of modern scientific instruments to produce data and the ability of researchers to control and examine the data in a reliable and timely manner.

- The emerging technology of cloud computing is a step forward from the current cyberinfrastructure.
OSDC continued...

- Through the Partnership for International Research & Education program (PIRE), the National Science Foundation (NSF) has funded a 5-year project devoted to the development of the Open Science Data Cloud.

- Scientists at the University of Chicago (UC), Florida International University (FIU) and the University of Illinois at Chicago (UIC) have collaborated with six international research groups.

- The OSDC project is also partnered with the Open Cloud Consortium (OCC) - a member driven organization that develops reference implementations, benchmarks and standards for cloud computing.
Training and Education Using the Open Science Data Cloud

Partnerships for International Research and Education

OSDC PIRE Project Overview

- Research
  - Cloud middleware for data intensive computing
  - Wide area clouds
- Training and education workshops
  - Data intensive computing using the OSDC
  - Cloud computing for scientific computing
- Outreach
  - OSDC Data Challenge
  - Awareness Workshops to help recruit a diverse work force

Foreign Partners

- Edinburgh University
- Universidade Federal Fluminense, Brasil
- University of Amsterdam
- Korea Institute of Science & Technology
- San Paulo State University
- National Institute of Advanced Industrial Science and Technology (AIST), Japan
- Beijing Institute of Genomics (BIG)

Research Focus

- Cloud architectures for data intensive computing
- Wide area clouds
- Continuous learning
- Scanning queries

OSDC Data Challenge

- Annual contest to select 1 to 2 datasets to add to the OSDC
- Will search for the most interesting datasets to add (and hopefully accompanying applications)

Ways to Participate

- Nominate one of your graduate students to spend a summer working with one of the OSDC PIRE Partners
- Send one of your graduate students to hands-on Workshops, such as Introduction to Data Intensive Computing
- Submit your most impressive dataset to the OSDC Data Challenge
- Buy a container of computers and join the OSDC

www.opensciencedatacloud.org
PIRE & OSDC

- This PIRE team intends to help develop large-scale distributed computing capabilities – the Open Science Data Cloud (OSDC) – to provide long term persistent storage for scientific data and state-of-the-art services for integrating, analyzing, sharing and archiving scientific data.
Objectives

- To study and strengthen storage systems that integrate specialized network protocols and support data transport over wide-area, high-performance networks.

- To develop new classes of cloud-based parallel programming frameworks and to integrate them into the cloud infrastructure so that this technology is more broadly available to scientists.

- To increase the involvement, in workshops and in subsequent use of the cloud cyberinfrastructure, of a large variety of scientists and their students.

- To train these groups in the basics of cloud computing and work to ensure that the cloud computing research advances to maximize the manageability and analytical power of the complex datasets unique to each scientific discipline.

- To catalyze a higher level of international engagement in the U.S. science and engineering community through international research and education collaborations.
Bionimbus, an instance on the OSDC, is a community cloud for storing, analyzing & sharing genomics & related data.

Bionimbus Virtual Machine released on Amazon EC2

Published on 2010/09/10 in Announcements. 0 Comments

A virtual machine image with common peak calling pipelines was made available on Amazon Web Services Elastic Cloud. Upon boot, it fetches pipeline library data, providing everything needed for processing user’s data.

Amazon EC2 ID: ami-aead58c7

Startup command: ec2-run-instances -n 1 -t m1.large ami-aead58c7

Upon connecting to your instance, wait for /READY-PIPELINE-DATA file to appear before commencing pipelines. This file signifies that pipeline data libraries installed successfully on your instance.

For more information see Bionimbus Machine Images.
Complete Genomics Chooses the Bionimbus as Mirror Site for CGI 60 Genomes Release

Published on 2011/02/03 in Uncategorized Closed

Complete Genomics Inc. has chosen the Bionimbus Community Cloud as a mirror site for their 60 Genomes dataset.

The 60 Genomes dataset can be found here, as part of the public data that Bionimbus makes available to researchers. With the Bionimbus Community Cloud, the data is available via both the commodity Internet, as well as via high performance research networks, such as the National LambdaRail and Internet2.

The genomes in the dataset have on average more than 55x mapped read coverage, and the sequencing of these 60 genomes generated more than 12.2 terabases (Tb) of total mapped reads. This dataset will complement other publicly available whole genome data sets, such as the 1000 Genomes Project's recent publication of six high-coverage and 179 low-coverage human genomes. Forty of the sixty genomes are available now and the remainder will be available at the end of March.
Research involvement opportunity

- We are pleased to announce an exciting international research and career development opportunity for students.

- We are recruiting for a **6-week funded** internship during summer months at foreign collaborator sites. This opportunity is available to graduate students to participate in sophisticated international research collaborations.

**SUMMER MONTH OPPORTUNITIES ONLY AT THIS TIME**

- **Summer 2011 research sites:**
  - Rio de Janeiro, Brazil
    - Applications with data sets from Navy training simulations
    - Applications with Petrobras
  - San Paulo, Brazil
    - Research at the State University and data center
    - Creation of a cloud cluster using OSDC
Participant Requirements

- Science and engineering students from all ranks (BS, MS, and PhD) and backgrounds (race, ethnicity, citizenship, and gender) who are interested in cloud computing are encouraged to apply to be an OSDC PIRE participant.

- Students with strong academic qualifications, strong research track record, and active participation in ongoing research projects have a higher chance of acceptance into the OSDC PIRE program.
Participant Requirements

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

- Assistant professors and post docs are also eligible

- Must be U.S. citizens or residents of the U.S.

- Students must be taking 3 University credits (Independent study credits available)

- Attendance to OSDC workshop in early summer in Chicago

  - **PLEASE CHECK WEBSITE FOR DETAILS ON CHICAGO WORKSHOP DATES**
Online cultural training

- Defense Language Institute, Foreign Language Center supports the US Department of Defense with foreign language needs. This is an interesting collection of information, some of which may be beneficial to our CI-PIRE students. Specifically, you may wish to explore the Downloads link or "Choose a Language" and explore pronunciations as well as cultural background. Because this site caters to support for the defense industry, not all of our destination countries are represented. This site will be helpful for Hindi, Cantonese, French, and Mandarin.

- MIT Open Course Ware (OCW) is a program supported by the Massachusetts Institute of Technology. Selected MIT course materials are publish on the web for world to reference. We have included several MIT OCW courses in this materials list. There are no costs for these courses, though students may choose to purchase text books.
  - MIT OCW: Communicating Across Cultures
  - MIT OCW: East Asian Cultures, from Zen to Pop

- My Language Exchange: Your language exchange online community
  This social network for language learners is a good place to find pen-pals or language practice partners.

- Omniglot: Writing systems and Languages of the World
  This is a wonderful linguistic reference site which contains links to language course sites.

- Open Culture: "The best free cultural & educational media on the web"
  This site offers a variety of cultural and language references.

- Other online cultural training information can be found at http://pire.fiu.edu/training/language.php
Student Application
1 page Proposal Required:

● Summary: 1 page proposal describing 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.
How to Apply:

- Please visit the Open Science Data Cloud website for more application information and for a list of important dates:

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

The OSDC PIRE project aims to narrow the growing gap between the capability of modern scientific instruments to produce data and the ability of researchers to control and examine the data in a reliable and timely manner.

The emerging technology of cloud computing is a step forward from the current cyberinfrastructure.
Questions?

- Contact Heidi Alvarez heidi@fiu.edu with your questions or comments.

- Contact Cindy Rogowski cindy.rogowski@uchicago.edu with questions regarding the application process, important dates or any other project specific inquiries.