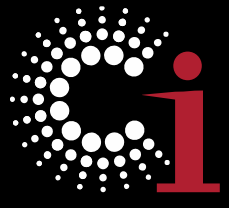


Optimizations in running large-scale Genomics workloads in Globus Genomics using HTCondor

Ravi Madduri

madduri@anl.gov

Joint work with Paul Davé, Lukasz Lacinski, Alex Rodriguez, Dinanath Sulakhe, Ryan Chard and Ian Foster

THE UNIVERSITY OF CHICAGO

Argonne NATIONAL LABORATORY

Galaxy

globus.org/genomics

Who We Are

- Globus Genomics is developed, operated, and supported by researchers, developers, and bioinformaticians at the Computation Institute – University of Chicago/Argonne National Lab

- We are a non-profit organization building solutions for non-profit researchers

- Our goal is to support the advancement of science by bringing together our strengths and capabilities to help meet the unique needs of researchers and research institutions

90% of cancer patients carry a mutation that may be responsive to a known drug

Mark Rubin, Weill Cornell Medical College and NewYork-Presbyterian Hospital in New York in *Nature, April, 2015*

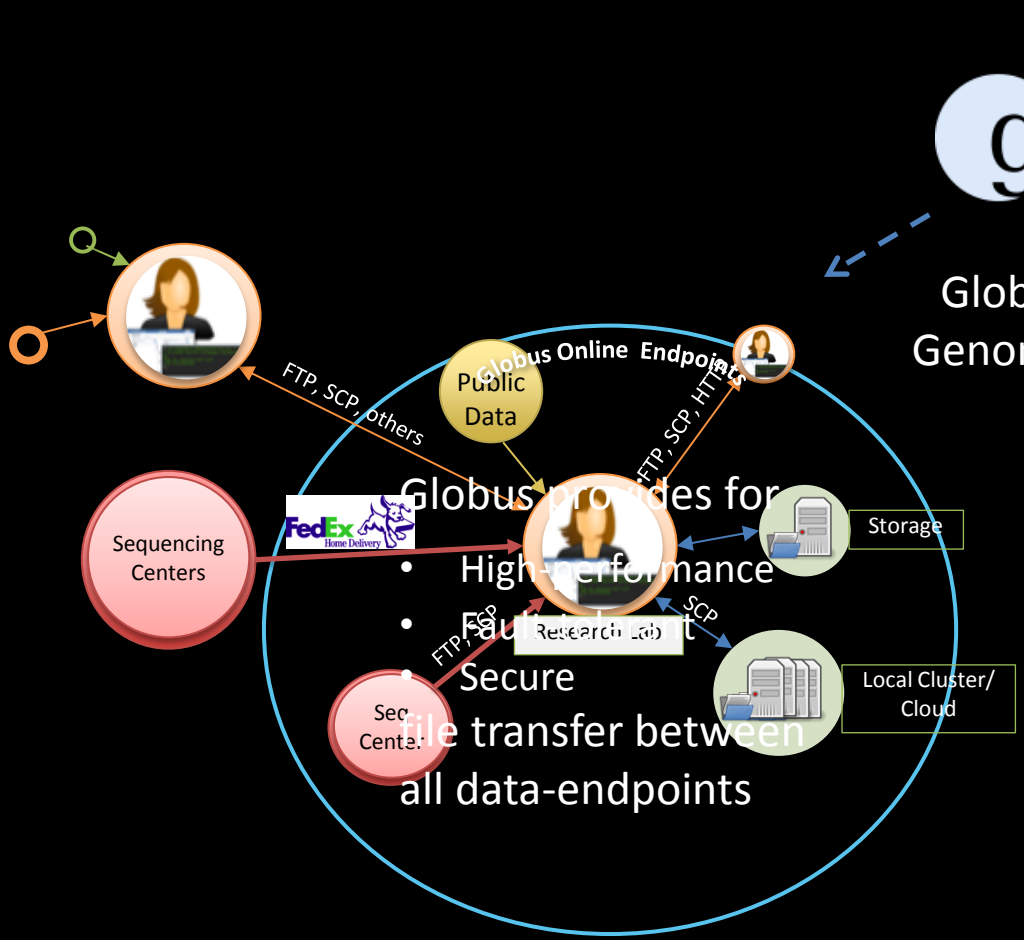
Trying to find a single causative gene for diseases with a complex genetic background is like looking for the proverbial needle in a haystack

– Nancy Cox
(Vanderbilt)

How do we accelerate discovery without requiring that every lab acquire a haystack-sorting machine?

Clayton & Shuttleworth thresher, 1910. Museum Victoria, Australia

Our answer: Globus Genomics

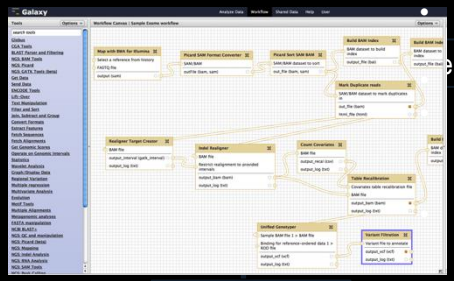


Data management

Globus Genomics

Galaxy Data Libraries

Galaxy-based workflow management



Globus integrated within Galaxy
Web-based UI
Drag-Drop workflow creations
Easily modify workflows with new tools

Analytical tools are automatically run on the scalable compute resources when possible

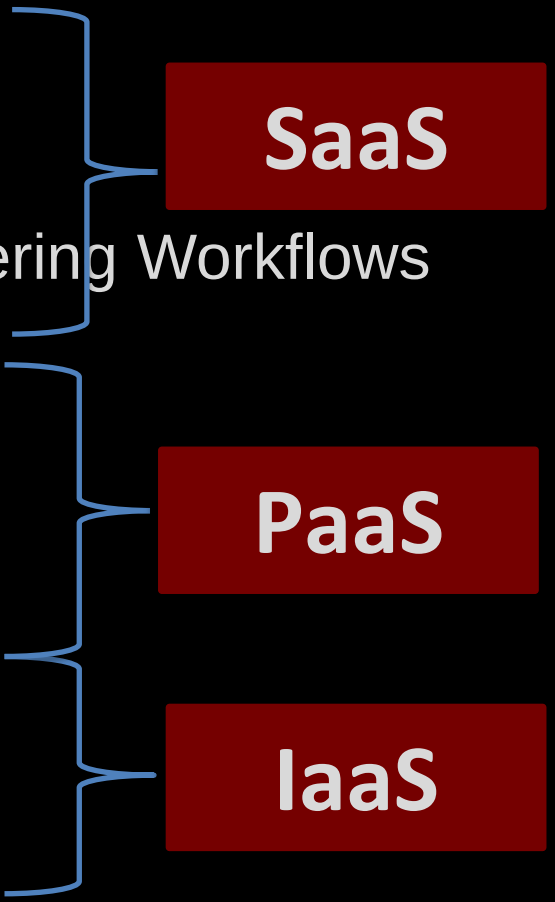
Globus Genomics on Amazon EC2

Data analysis

globus.org/genomics

Our Science Stack

- Galaxy
  - Interactive execution
  - Creation, Execution, Sharing, Discovering Workflows
- Globus
  - Data management
  - Identity Management
- AWS
  - HTCondor, Chef, EC2, EBS, S3, SNS
  - Spot, Route 53, Cloud Formation



Key Technical Bits

- HTCondor
- Computational Profiles for various analysis tools
- Elastic Spot instance provisioner
- Chef
- Nagios + Munin
- Support

Challenges/Opportunities

- Optimize for cost, performance and both
- Deliver on SLAs
- AWS has an awesome spot market with low prices for compute
- AWS billing is in hourly increments
- It takes roughly 8mins on an average to ask for a spot instance and have it join
- Global view of workloads across multiple customers

Globus Galaxies

- Each GG gateway operates a persistent head node, exposing the Galaxy service
- The head node operates an HTCondor pool and periodically executes a resource provisioning application
- The provisioner monitors the HTCondor queue for idle jobs
- The provisioner processes idle job ClassAds and monitors current EC2 spot market prices to cost-effectively acquiring resources

Globus Galaxies and HTCondor

- We have created execution profiles for frequently run genomics applications, mapping CPU, memory, and IO requirements to EC2 resource types
- Resources are dynamically contextualized to meet the requirements of GG workloads
- Each worker resource is configured as a single HTCondor slot, utilizing all of the resources advertised cores
- Once configured, the worker resource joins the pool operated by the gateway's head node

HTCondor Resource Migration

- To improve utilization and reduce cost we want to apply idle resources to another GG gateway's jobs, moving the resources between HTCondor pools
- Centralize provisioning and resource monitoring
- Employ a hierarchy of HTCondor collectors, having each gateway's HTCondor service report to additional collectors
  - Enables logical separation between Globus Galaxies services (Genomics, Materials science, etc.)
  - Global queue data can be processed
- Opt to move an idle worker resource between HTCondor masters

HTCondor Resource Migration

- Migrating resources:
  - Advertise and set attributes stating a resource is leaving a pool
  - Execute condor_off –peaceful
  - Reconfigure the resource
    - Modify slot definition to meet requirements
    - Modify host address to new gateway
  - Restart the condor service

Considerations

- Data transfer rates and limits across regions, and even between zones, are significant
- Facilitates the inclusion of gateway priority
- Enables a new billing methodology, distinct of AWS

Cox lab, UChicago

Consensus Genotyper for Exome Sequencing: Improving the Quality of Exome Variant Genotypes

Vassily Trubetskoy[1], Ravi Madduri[2], Alex Rodriguez[2], Jeremiah Scharf[3], Paul Dave[2], Ian Foster[2] , Nancy Cox[1] , Lea Davis[1]

1) Section Genetic Medicine, University of Chicago, Chicago, IL; 2) Computation Institute, University of Chicago, Chicago, IL; 3) Department of Neurology, Massachusetts General Hospital, Boston, MA

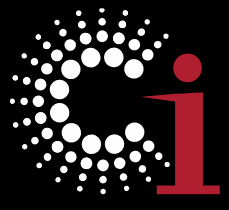
- 134 samples and 4 workflows
- 4 TB data
- 2200 core hours in 6 days

Olopade lab, UChicago

A profile of inherited predisposition to breast cancer among Nigerian women

Y. Zheng, T. Walsh, F. Yoshimatsu, M. Lee, S. Gulsuner, S. Casadei, A. Rodriguez, T. Ogundiran, C. Babalola, O. Ojengbede, D. Sighoko, R. Madduri, M.-C. King, O. Olopade

- 200 targeted exomes
- 200 GB data
- 76,920 core hours in 1.25 days



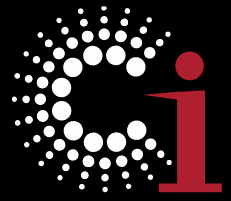
Innovation Center for Biomedical Informatics - Georgetown

A case study for high throughput analysis of NGS data for translational research using Globus Genomics

D. Sulakhe, A. Rodriguez, K. Bhuvaneshwar, Y. Gusev, R. Madduri, L. Lacinski, U. Dave, I. Foster, S. Madhavan

- 78 exomes from lung cancer study
- 2 TB data
- 125,936 core hours in 1.7 days

Other Globus Genomics users



Dobyns Lab

Seattle Children's
HOSPITAL • RESEARCH • FOUNDATION

KU MEDICAL CENTER
The University of Kansas

Berkeley
UNIVERSITY OF CALIFORNIA

Nagarajan Lab

Washington University in St. Louis

UPMC
University of Pittsburgh Medical Center

Wexner Medical Center

THE UNIVERSITY OF CHICAGO

Cox Lab
Volchenboum Lab
Olopade Lab

BD

Join the future of health.

Avera

GEORGETOWN UNIVERSITY

LA BioMed
Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center

UCSF

Genome Science Institute

Boston University Medical Center

PerkinElmer
For the Better

GEORGETOWN UNIVERSITY

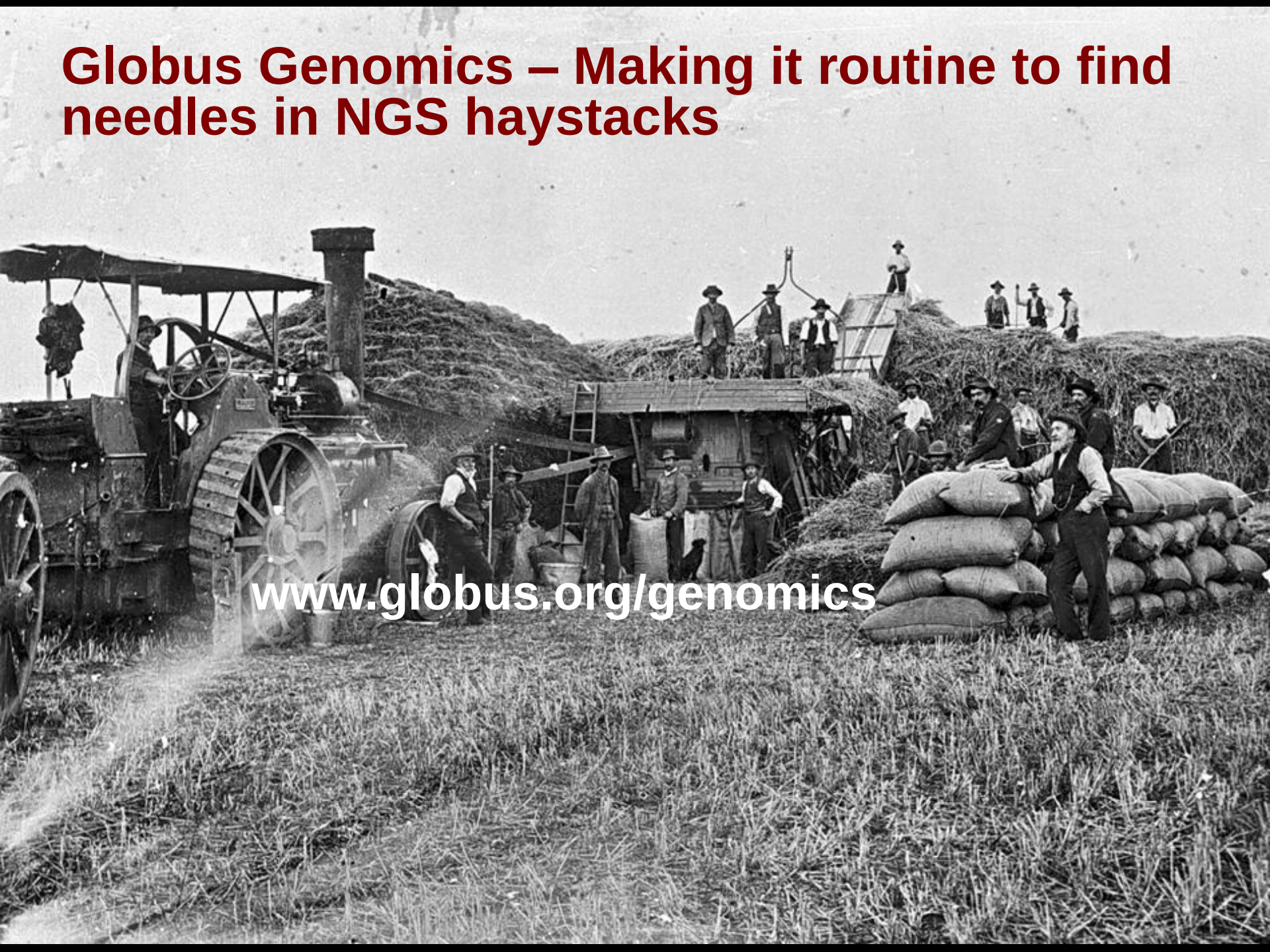
Costs are remarkably low

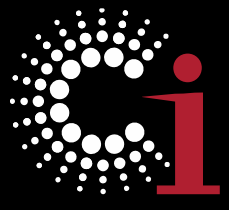
Exome	Whole Genome	RNA-Seq.
$5 - $30	$20 - $100	$5 - $10
Pricing based on example of paired-end fastq files with 5 Gbases.	Pricing based on example of paired-end fastq files with 80 Gbases.	Pricing based on example of paired-end fastq files with 5 Gbases.
Pipeline includes quality control, alignment, variant calling, and annotation using the GATK best-practices pipeline.	Pipeline includes quality control, alignment, variant calling, and annotation.	Pipeline includes quality control, alignment, exon count using cufflinks, and HT-Seq count.

Pricing includes

- Estimated compute
- Storage (one month)
- Globus Genomics platform usage
- Support

Globus Genomics – Making it routine to find needles in NGS haystacks



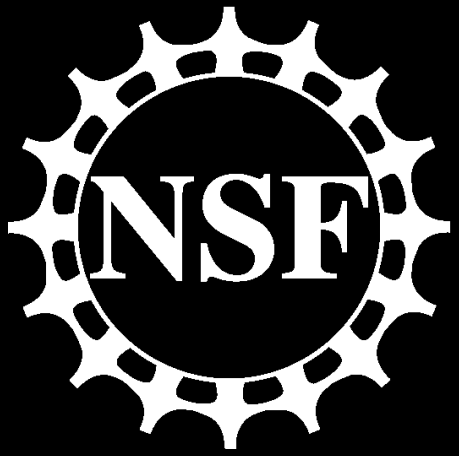


- More information on Globus Genomics and to sign up for a free trial : www.globus.org/genomics

- More information on Globus: www.globus.org

Our work is supported by:

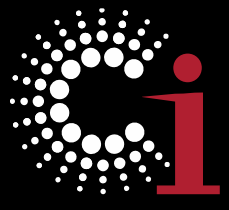
U.S. DEPARTMENT OF
ENERGY



THE UNIVERSITY OF
CHICAGO

Argonne
NATIONAL LABORATORY

globus.org/genomics



Thank you!

@madduri