Optimizing the National Cyberinfrastructure for Lower Bioinformatic Costs: Making the Most of Resources for Publicly Funded Research

William K. Barnett, Ph.D. (Director)

Richard LeDuc, Ph.D. (Manager)

National Center for Genome Analysis Support

RNA-Seq 2013, Boston MA, 6/20/2013

INDIANA UNIVERSITY

Summary

- NCGAS and its mission and cyberinfrastructure.
- Overview NCGAS server-on-demand resources on a low cost fee-for-cycles basis.
- Overview XSEDE: When you need truly large-scale resources.

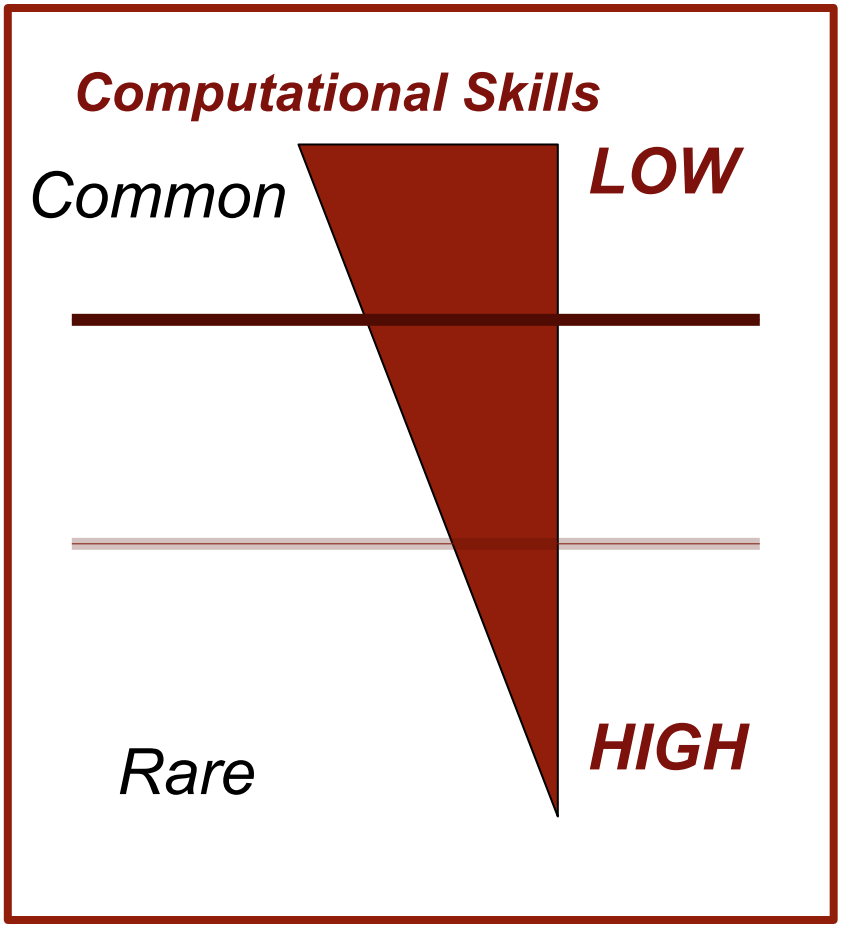
NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

INDIANA UNIVERSITY

- Funded by National Science Foundation
  1. Large memory clusters for assembly
  2. Bioinformatics consulting for biologists
  3. Optimized software for better efficiency

- Collaboration across IU, TACC, SDSC, and PSC.
- Open for business at: http://ncgas.org

Making it easier for Biologists



Galaxy

- Web interface to NCGAS resources
- Supports many bioinformatics tools
- Available for both research and instruction.

NCGAS Cyberinfrastructure at IU

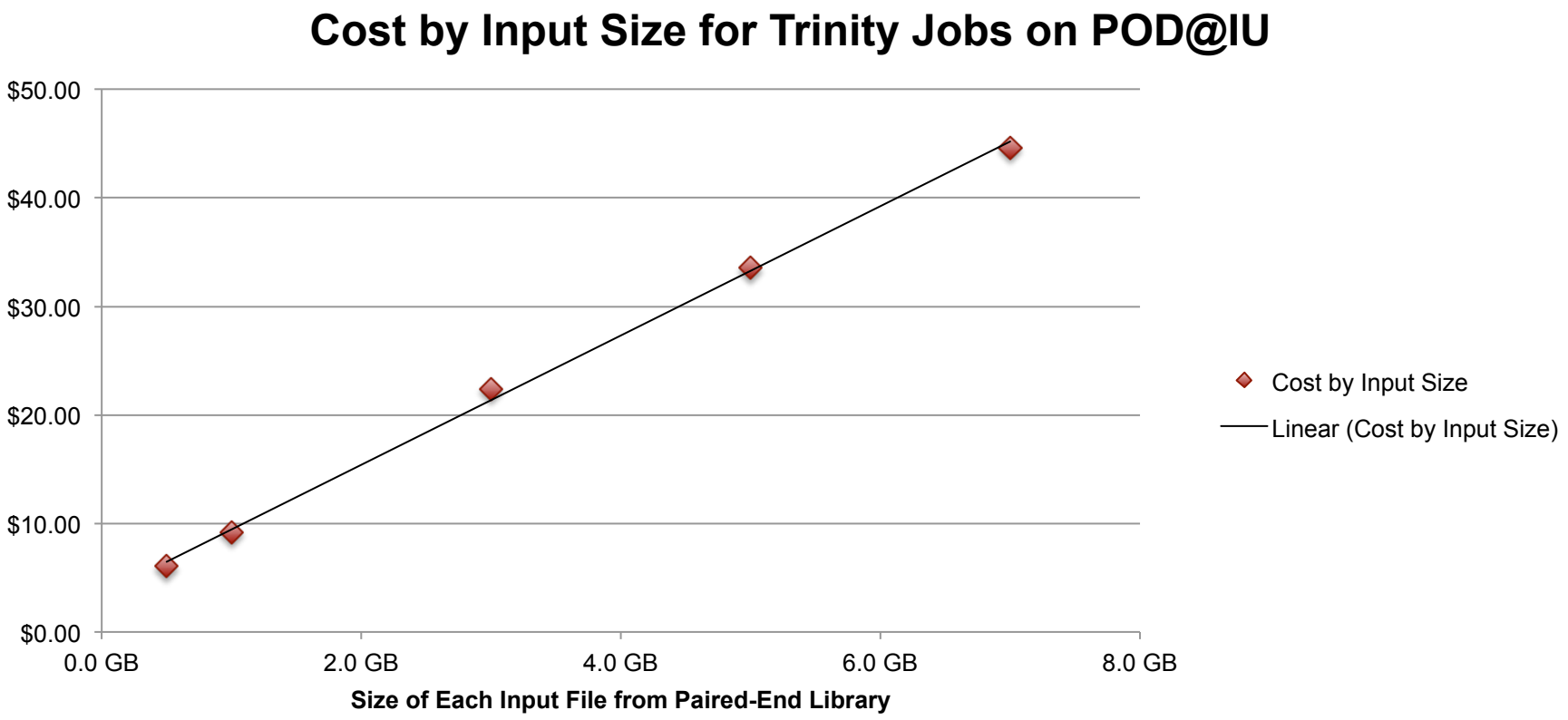
- Rockhopper: 11 servers with 48 cores and 128 GB RAM.
- Mason large memory cluster: 16 nodes with 32 cores each and 512 GB RAM per node.
- Data Capacitor: 1 PB at 20 Gbps throughput.
- Research Database Cluster for managing data sets.
- All interconnected with high speed internal network (40 Gbps)
- 100 Gbps Internet2 Backbone

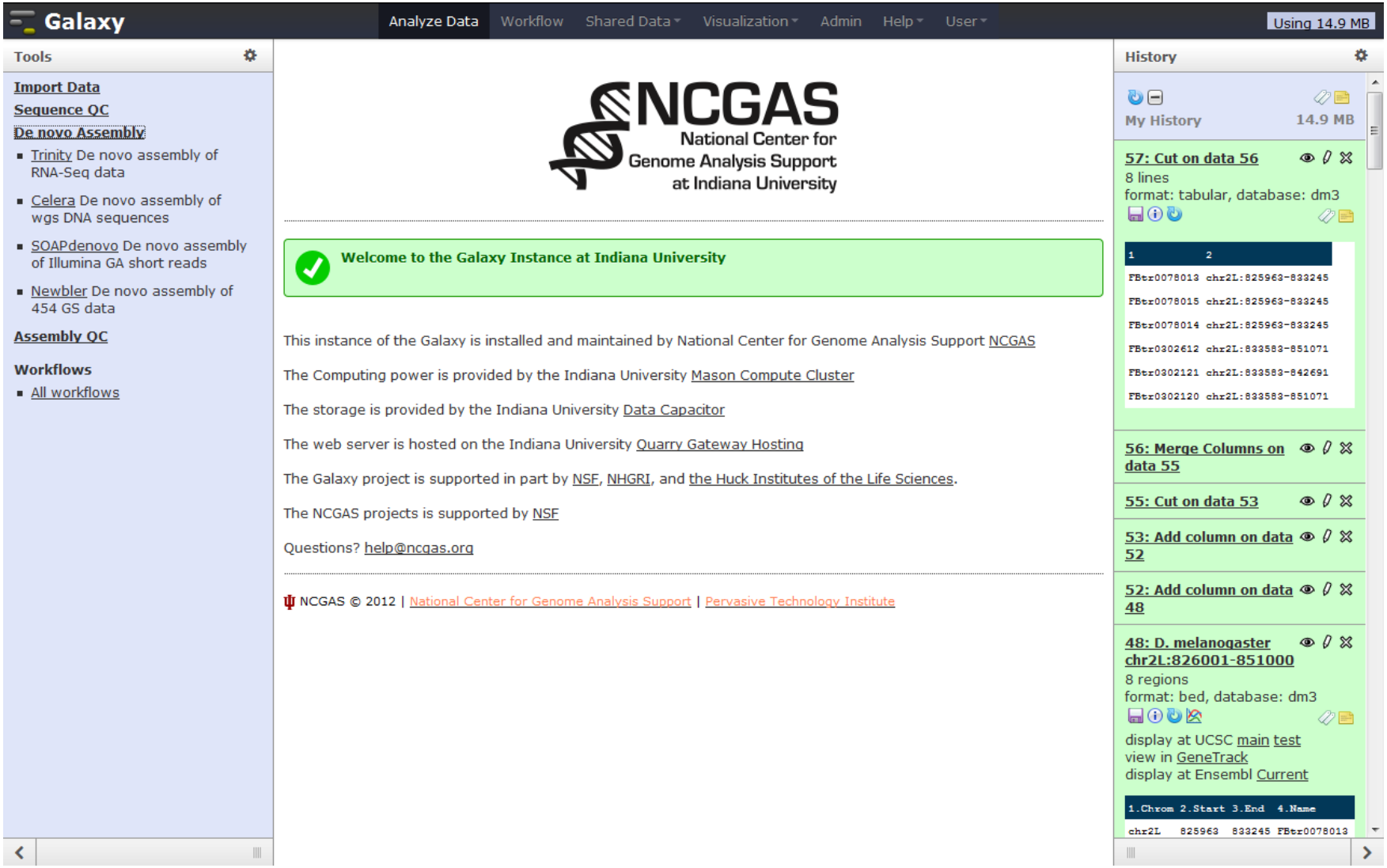
Rockhopper

- Penguin Computing's Penguin-On-Demand (POD) supercomputing cloud appliance hosted by Indiana University.
- A collaborative effort between Penguin Computing, IU, the University of Virginia, the University of California Berkeley, and the University of Michigan.
- Provides supercomputing cloud services in a secure US facility.

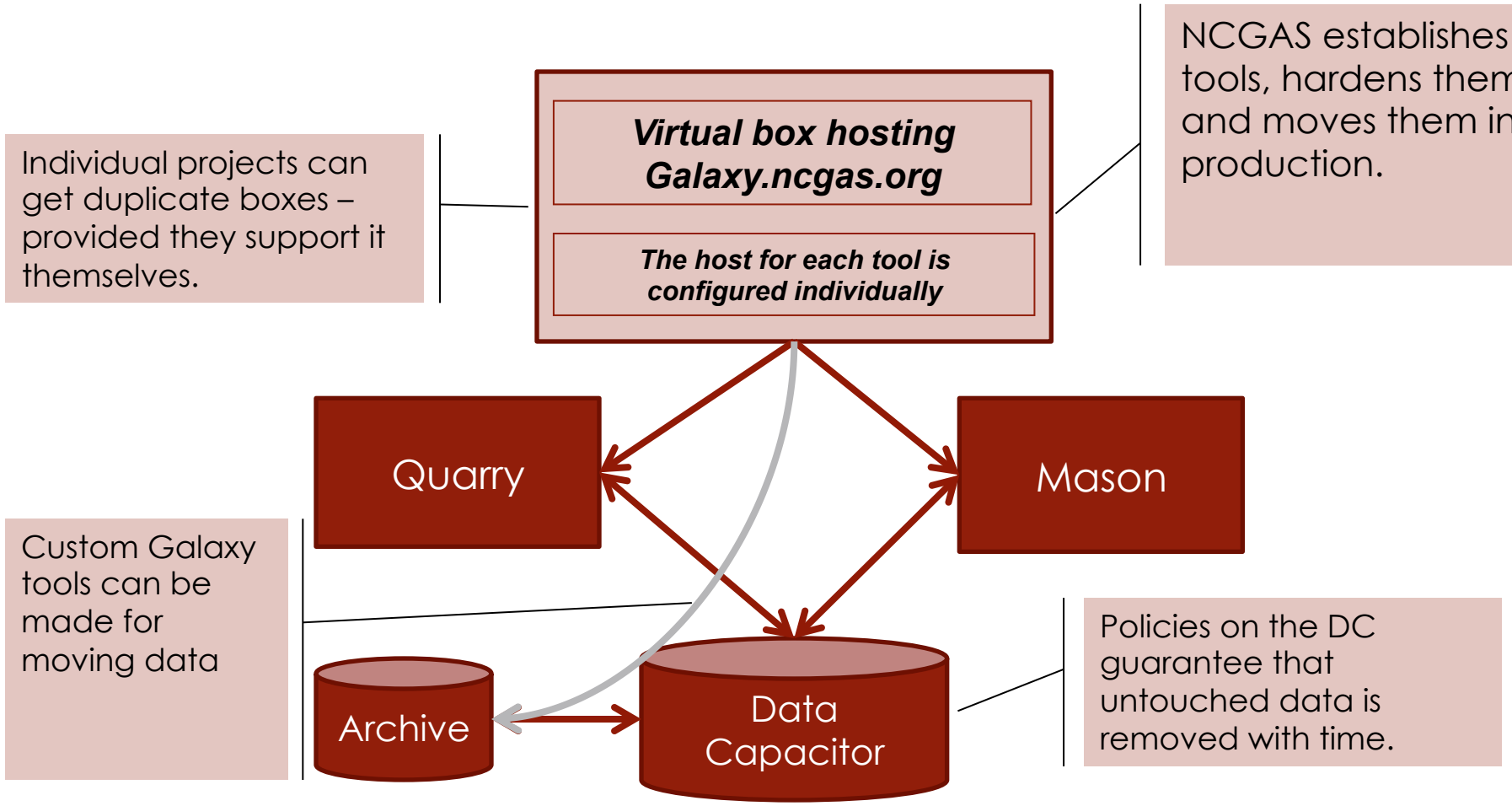
- Researchers at US institutions of higher education and Federally Funded Research and Development Centers (FFRDCs) can purchase computing time from Penguin Computing, and receive access via high-speed national research networks operated by IU.

Standardized Trinity Analyses

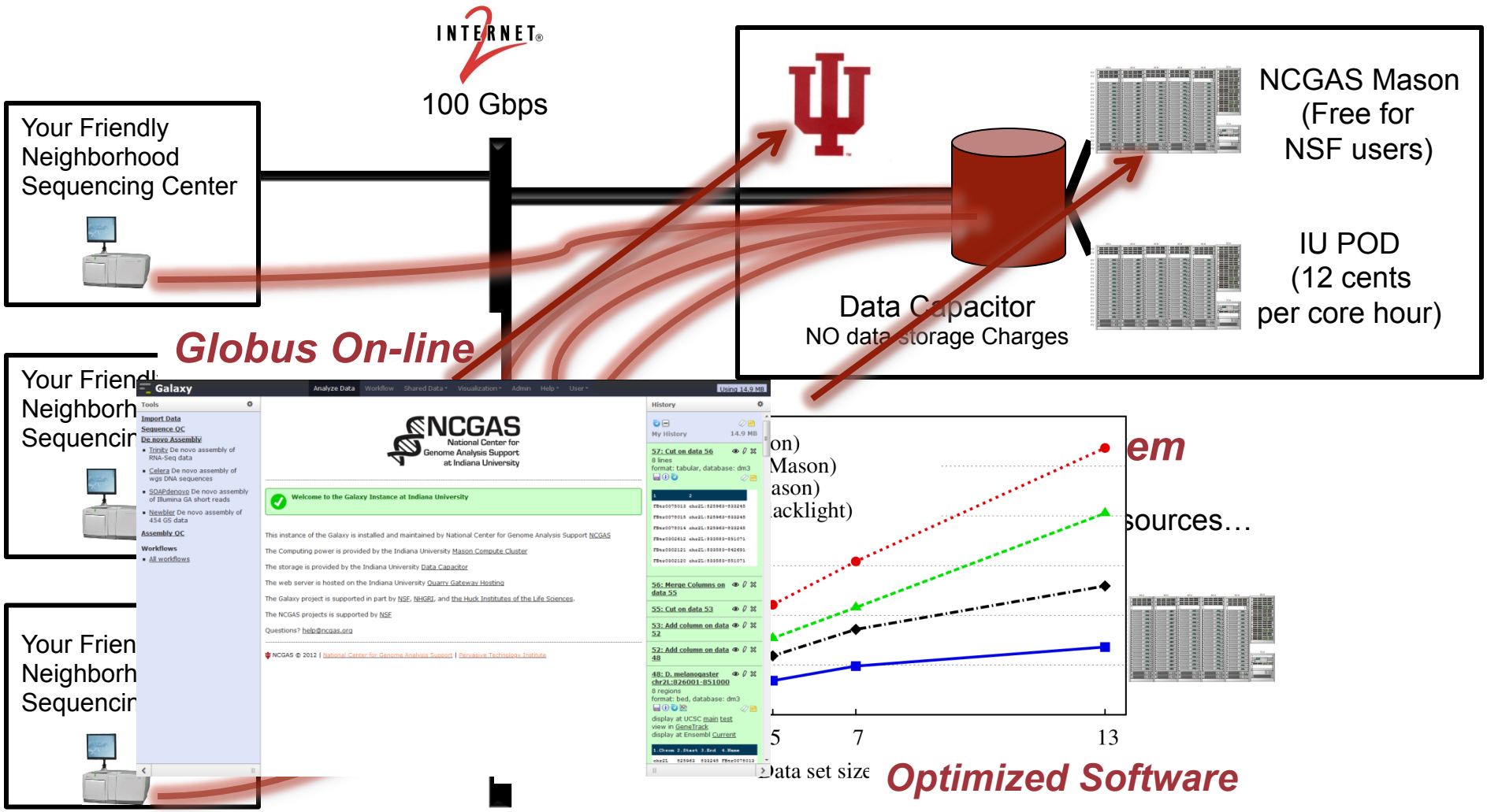


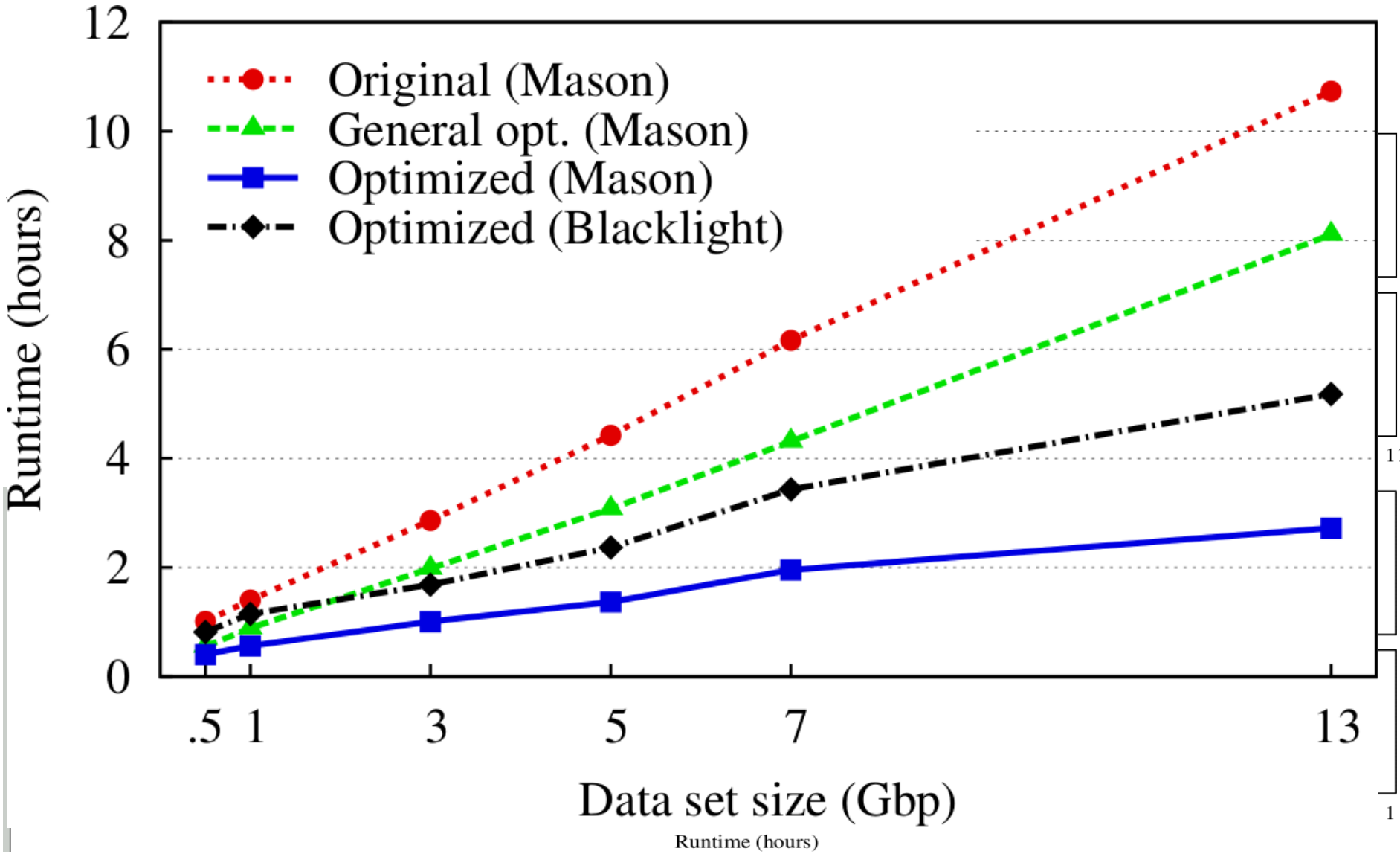

Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 14.9 MB
Tools
Import Data
Sequence QC
De novo Assembly
Trinity De novo assembly of RNA-Seq data
Celera De novo assembly of wgs DNA sequences
SOAPdenovo De novo assembly of Illumina GA short reads
Newbler De novo assembly of 454 GS data
Assembly QC
Workflows
All workflows
NCGAS
National Center for Genome Analysis Support at Indiana University
Welcome to the Galaxy Instance at Indiana University
This instance of the Galaxy is installed and maintained by National Center for Genome Analysis Support NCGAS
The Computing power is provided by the Indiana University Mason Compute Cluster
The storage is provided by the Indiana University Data Capacitor
The web server is hosted on the Indiana University Quarry Gateway Hosting
The Galaxy project is supported in part by NSF, NHGRI, and the Huck Institutes of the Life Sciences.
The NCGAS projects is supported by NSF
Questions? help@ncgas.org
NCGAS © 2012 | National Center for Genome Analysis Support | Pervasive Technology Institute
History
My History
14.9 MB
57: Cut on data 56
8 lines
format: tabular, database: dm3
1 2
FBtr0078013 chr2L:825963-833245
FBtr0078015 chr2L:825963-833245
FBtr0078014 chr2L:825963-833245
FBtr0302612 chr2L:833583-851071
FBtr0302121 chr2L:833583-842691
FBtr0302120 chr2L:833583-851071
56: Merge Columns on data 55
55: Cut on data 53
53: Add column on data 52
52: Add column on data 48
48: D. melanogaster chr2L:826001-851000
8 regions
format: bed, database: dm3
display at UCSC main test
view in GeneTrack
display at Ensembl Current
1.Chrom 2.Start 3.End 4.Name
chr2L 825963 833245 FBtr0078013

GALAXY.NCGAS.ORG Model

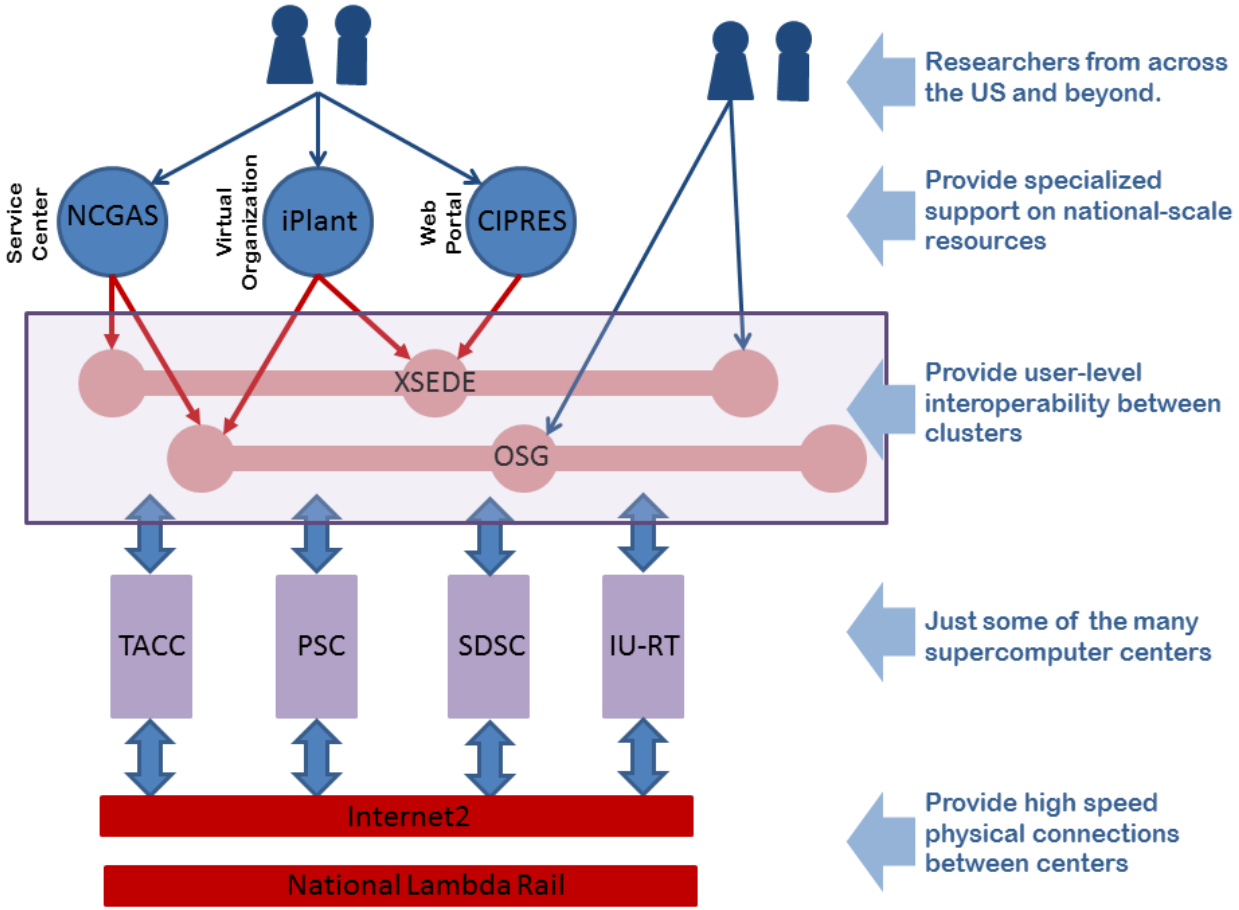


Moving Forward



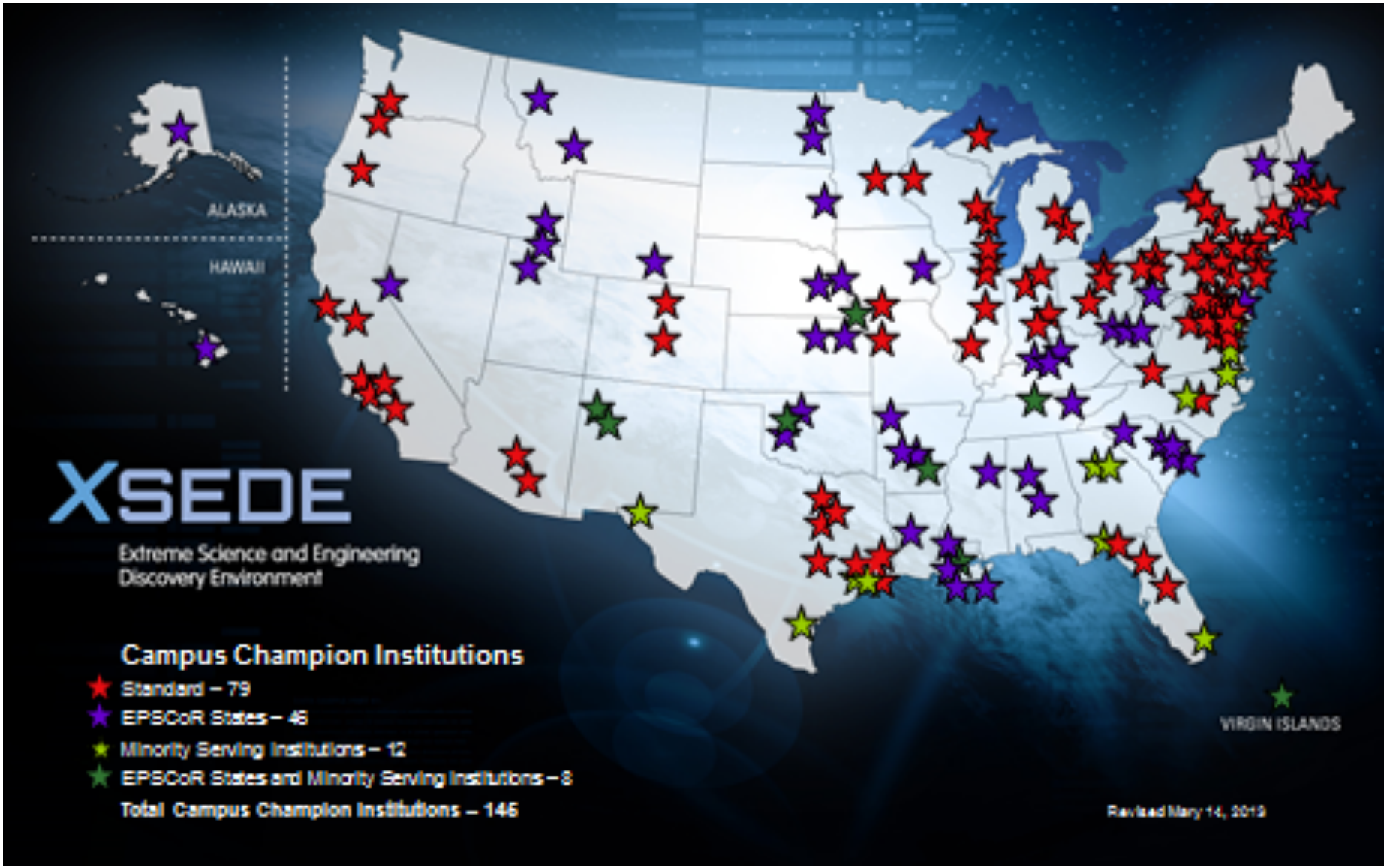

12
10
8
6
4
2
0
Runtime (hours)
Original (Mason)
General opt. (Mason)
Optimized (Mason)
Optimized (Blacklight)
.5
1
3
5
7
13
Data set size (Gbp)
Runtime (hours)

The National Cyberinfrastructure



XSEDE: Extreme Science and Engineering Discovery Environment

- About 100 requests every quarter.
- About 50% of need is met.
- 75% from two systems.
- Allocations in the millions of SU.


ALASKA
HAWAII
XSEDE
Extreme Science and Engineering Discovery Environment
Campus Champion Institutions
Standard – 79
EPSCoR States – 46
Minority Serving Institutions – 12
EPSCoR States and Minority Serving Institutions – 8
Total Campus Champion Institutions – 145
VIRGIN ISLANDS
Revised May 14, 2013

In Sum…

- NG Sequencing is creating a analytical problem that cannot be solved at sequencing centers
- NCGAS can provide a global scale infrastructure to better serve the needs of biologists who cannot become bioinformaticians to accomplish their research.
- XSEDE allows scaling to larger projects.

Thank You

Questions?

Bill Barnett (barnettw@iu.edu)

Rich LeDuc (rleduc@iu.edu)

Le-Shin Wu (lewu@iu.edu)

Carrie Ganote (cganote@iu.edu)

NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

INDIANA UNIVERSITY

Acknowledgements & disclaimer

- This material is based upon work supported by the National Science Foundation under Grants No. ABI-1062432
- This work was supported in part by the Lilly Endowment, Inc. and the Indiana University Pervasive Technology Institute

- Any opinions presented here are those of the presenter(s) and do not necessarily represent the opinions of the National Science Foundation or any other funding agencies

License terms

- Please cite as: LeDuc, R.D., Optimizing the National Cyberinfrastructure for Lower Bioinformatic Costs: Making the Most of Resources for Publicly Funded Research, presented at RNA-Seq Summit 2013, Boston MA, 6/20/2013.
- Items indicated with a © are under copyright and used here with permission. Such items may not be reused without permission from the holder of copyright except where license terms noted on a slide permit reuse.
- Except where otherwise noted, contents of this presentation are copyright 2011 by the Trustees of Indiana University.
- This document is released under the Creative Commons Attribution 3.0 Unported license (http://creativecommons.org/licenses/by/3.0/). This license includes the following terms: You are free to share – to copy, distribute and transmit the work and to remix – to adapt the work under the following conditions: attribution – you must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work). For any reuse or distribution, you must make clear to others the license terms of this work.