

PENNSTATE®



Dynamically Scalable, Accessible Analysis for High- Throughput Sequence Data

Enis Afgan, Anton Nekrutenko, James Taylor, and
the Galaxy Team

Bio-IT World, April 13, 2011 - Boston, MA



EMORY

Principles for Computational Science

Accessibility

- ▶ getting data, methods
- ▶ running tools
- ▶ creating workflows

Reproducibility

original data +
methods +
execution +
context =
meaningful results

Transparency

- ▶ communication
- ▶ repeat, reuse, extend

Galaxy: accessible analysis system

The screenshot displays the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. On the left, a 'Tools' sidebar lists various categories like 'Recently Used', 'Get Data', 'Text Manipulation', and 'NGS TOOLBOX BETA'. The main content area features a 'Galaxy 101' tutorial with the text 'Start small The very first tutorial you need' and 'Live Quickies' for 'Advanced fastQ manipulation' and '454 Mapping: Single End'. Below this is a paragraph about the Galaxy team and a tweet from the galaxyproject. On the right, a 'History' panel shows a workflow named 'Galaxy 101' with steps: '1: Exons', '2: SNPs', '3: Join on data 2 and data 1', '4: Group on data 3', '5: Sort on data 4', and '6: Select first on data 5'. A table of genomic coordinates is visible under step 1.

1:Chrom	2:Start	3:End	4:Name
chr22	18824444	18825832	uc002eac.2_1
chr22	20476281	20463200	uc002eac.2_2
chr22	21738147	21743267	uc002eac.2_3
chr22	46652457	46659219	uc002eac.2_4
chr22	21489526	21489925	uc002eac.2_5

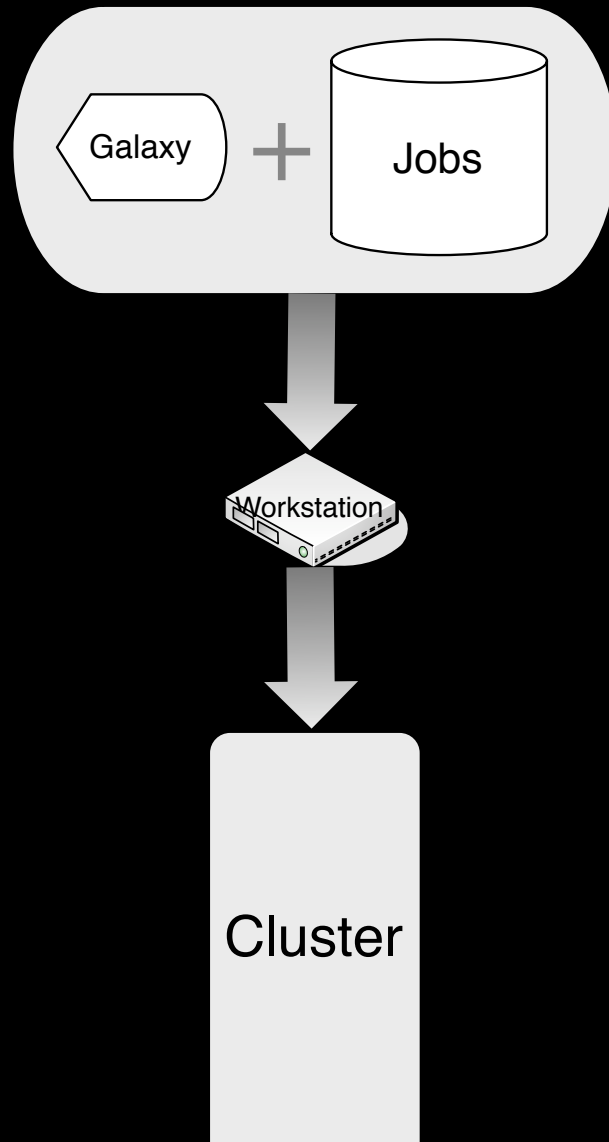
- Easily integrate new tools
- Consistent tool user interfaces automatically generated
- History system facilitates and tracks multistep analyses
- Exact parameters of a step can always be inspected, and easily rerun
- Sharing: analyses, processes, tools, results
- Workflow system
- Data visualization with Trackster

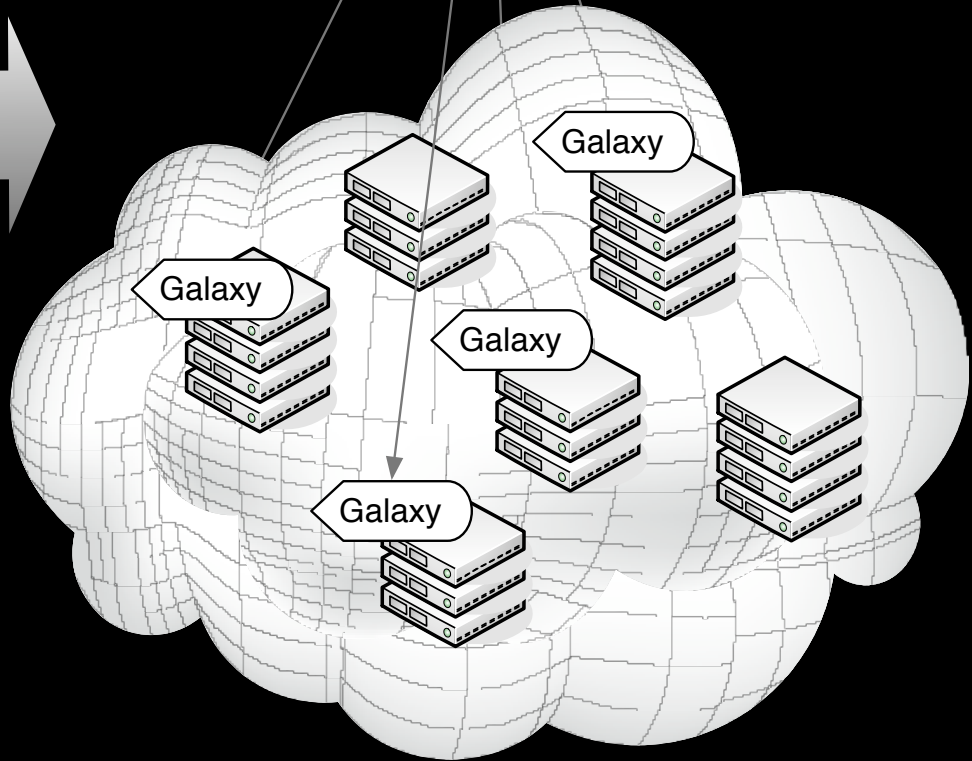
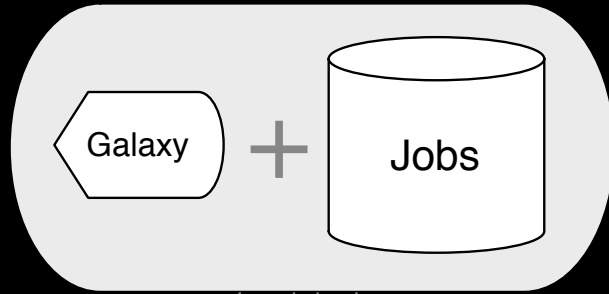
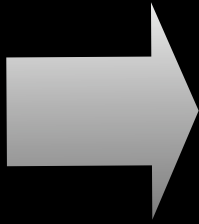
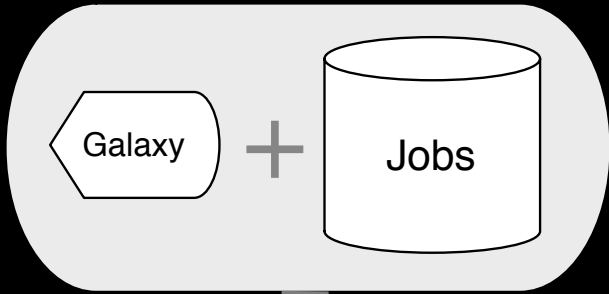
<http://usegalaxy.org/>

Enable **accessible**, **transparent**, and **reproducible** research

The shared resource problem

- Limited computational and storage capacity
- Must upload data to a shared resource
- Difficult to impossible to customize
- Lack of support for oscillating data volume





Galaxy on the Cloud

- Ideal for small labs and individual researchers
 - Labs do not have to house compute resources
 - Support variable volume of analysis data and computation requirements
 - Ready deployment with pre-configured reference genomes and tools
- Goal is to keep Galaxy use unchanged but deliver flexibility and job performance improvement while eliminating an otherwise required setup

Galaxy CloudMan

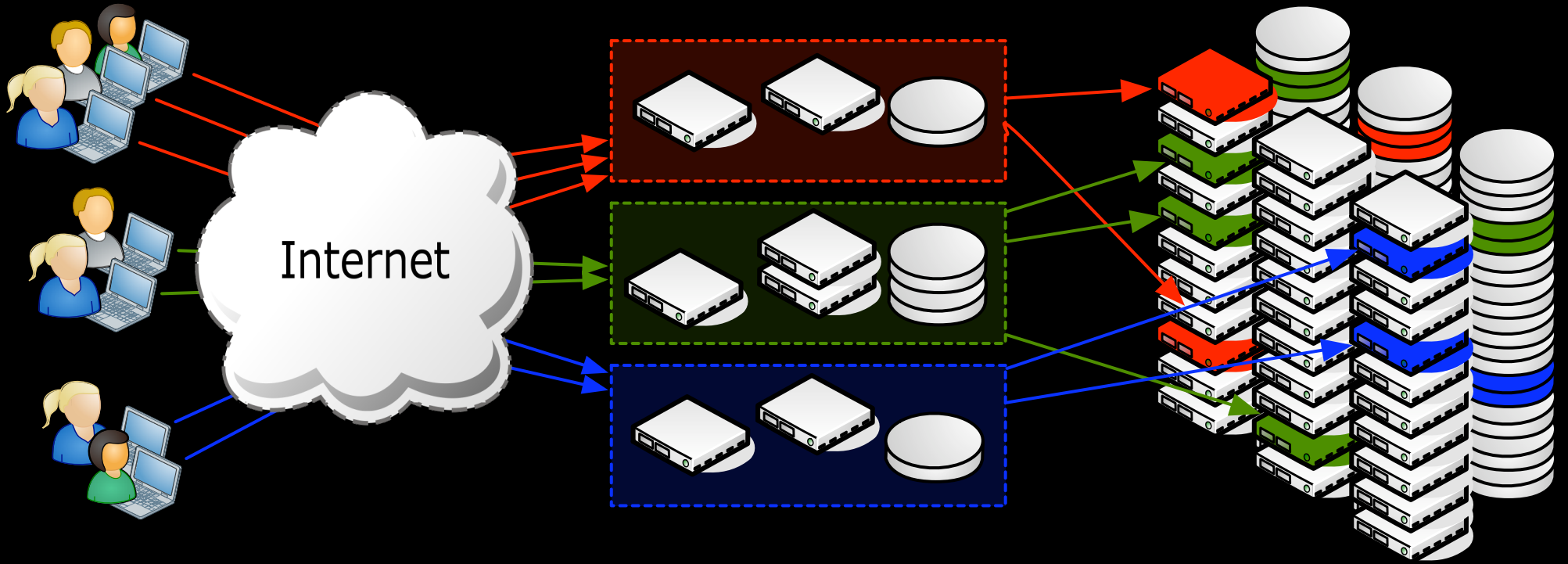
- Complete solution for instantiating, running and scaling cloud resources with an automatically configured Galaxy application
- Deployment on Amazon Web Services Cloud
- **Wizard-guided setup**: requires no computational expertise, no infrastructure, no software
- **Elastic resource scaling**: manual or automatic
- **Dynamic persistent storage**
- **Sharing** of derived cluster instances
- **Automated** configuration for machine image, tools, and data
- **Self-contained** deployment

- **Deploy a Galaxy cluster in minutes!**

A. Users in different labs

B. Isolated Galaxy instance(s)

C. Dense data center



SaaS



IaaS

Deploying Galaxy on the AWS Cloud

1. **Create an AWS account** and sign up for EC2 and S3 services
2. Use the AWS Management Console to **start a master EC2 instance**
3. **Use the Galaxy CloudMan web interface** on the master instance to manage the cluster

2. Start an EC2 Instance

The image shows a composite screenshot of the AWS Management Console. At the top, the navigation bar includes 'AWS', 'Products', 'Developers', 'Community', 'Support', and 'Account' (highlighted with a red box). Below this, the 'Your Account' section is visible on the left, with 'Security Credentials' highlighted by a red box. The main content area is the 'Amazon EC2 Console Dashboard', which includes a 'Getting Started' section with a 'Launch Instance' button and a 'My Resources' section showing 0 Running Instances, 0 Elastic IPs, 6 EBS Volumes, and 12 EBS Snapshots. A 'Request Instances Wizard' modal is open in the foreground, showing the configuration for a new instance. The wizard is in the 'REVIEW' step and displays the following details:

- AMI: Other Linux AMI ID ami-ed03ed84 (x86_64) Edit AMI
- Number of Instances: 1
- Availability Zone: No Preference
- Monitoring: Disabled
- Instance Type: Large (m1.large)
- Instance Class: On Demand Edit Instance Details
- Kernel ID: Use Default
- Ramdisk ID: Use Default
- User Data: testGC1|AKIAJKQI3RT... Edit Advanced Details
- Key Pair Name: galaxy_keypair Edit Key Pair
- Security Group(s): default, galaxyWeb Edit Firewall

At the bottom of the wizard, there are 'Back' and 'Launch' buttons.

3. Configure Your Cluster





The screenshot displays the Galaxy Cloudman web interface. At the top left, the logo 'Galaxy Cloudman' is visible. At the top right, there are links for 'Info: report bugs | wiki | screencast'. The main content area is titled 'Galaxy Cloudman Console'. A central dialog box titled 'Initial Cluster Configuration' is open, containing the following text: 'Welcome to Galaxy Cloudman. This application will allow you to manage this cluster and the services provided within. To get started, choose the type of cluster you'd like to work with and specify the size of your persistent data storage, if any.' Below this text, there is a radio button selected for 'Start a full Galaxy Cluster. Specify initial storage size (in Gigabytes)'. A text input field contains the value '100' followed by 'GB' and a green 'OK' label. A link 'Show more startup options' is also present. At the bottom of the dialog box is a 'Start Cluster' button. In the background, the console interface shows a 'Terminal' tab and a 'Status' section with labels for 'Cluster name', 'Disk status', 'Worker status', 'Service status', and 'External Logs'.

Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application allows you to manage this instance of Galaxy CloudMan. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to add and remove 'worker' nodes for running jobs.


[Terminate cluster](#) [Add nodes ▼](#) [Remove nodes](#) [Access Galaxy](#)

Status

Cluster name: share-an-instance demo 
Disk status: 84M / 10G (1%) 
Worker status: Idle: 0 Available: 0 Requested: 0
Service status: Applications  Data 
External Logs: [Galaxy_Log](#)



Autoscaling is **off**.
Turn on?

Cluster status log 



Tools Options ▾

- [Get Data](#)
- [Send Data](#)
- [ENCODE Tools](#)
- [Lift-Over](#)
- [Text Manipulation](#)
- [Filter and Sort](#)
- [Join, Subtract and Group](#)
- [Convert Formats](#)
- [Extract Features](#)
- [Fetch Sequences](#)
- [Fetch Alignments](#)
- [Get Genomic Scores](#)
- [Operate on Genomic Intervals](#)
- [Statistics](#)
- [Wavelet Analysis](#)
- [Graph/Display Data](#)
- [Regional Variation](#)
- [Multiple regression](#)
- [Multivariate Analysis](#)
- [Evolution](#)
- [Metagenomic analyses](#)
- [FASTA manipulation](#)

- NGS TOOLBOX BETA
- [NGS: QC and manipulation](#)
- [NGS: Assembly](#)
- [NGS: Mapping](#)
- [NGS: Indel Analysis](#)
- [NGS: Expression Analysis](#)
- [NGS: SAM Tools](#)
- [NGS: Peak Calling](#)
- [Human Genome Variation](#)
- [EMBOSS](#)

Welcome to Galaxy on the Cloud

History Options ▾

i Your history is empty. Click 'Get Data' on the left pane to start

Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application allows you to manage this instance of Galaxy CloudMan. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to add and remove 'worker' nodes for running jobs.

Stat

Galaxy Cloudman Console

Autoscaling Configuration

Autoscaling autoscaling
Autoscaling do. The clust grow larger!
While respect automaticall nodes reduc
Once turned

Minimum
Maximum
Type of N

Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application allows you to manage this instance of Galaxy CloudMan. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to add and remove 'worker' nodes for running jobs.

Status

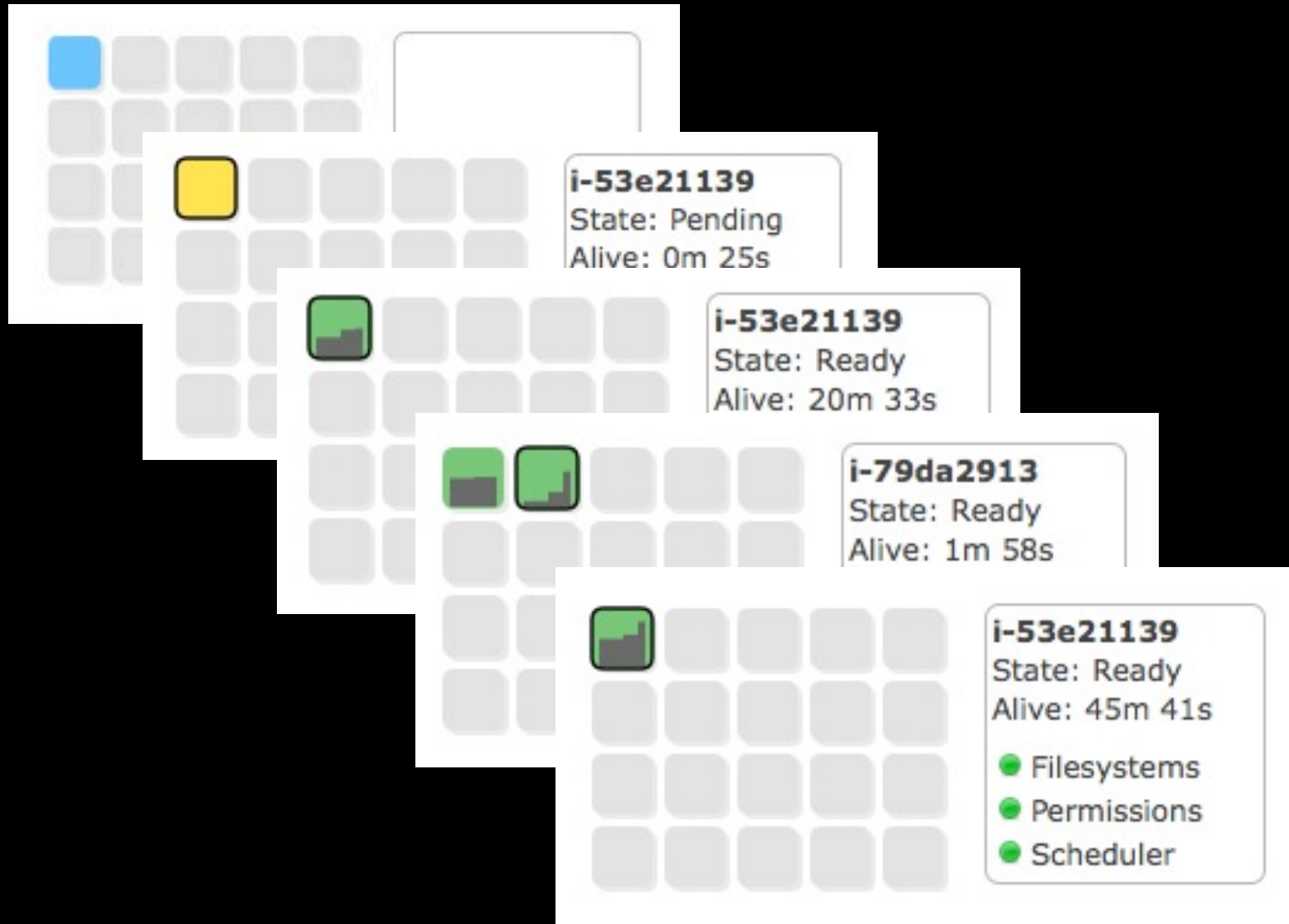
Cluster name: share-an-instance demo
Disk status: 84M / 10G (1%)
Worker status: Idle: 0 Available: 0 Requested: 0
Service status: Applications Data
External Logs: [Galaxy Log](#)



Autoscaling is on.
Turn off?
Min nodes: 0
Max nodes: 15
[Adjust limits?](#)

Cluster status log

4. Grow and Shrink



Once an analysis is complete

The image shows a screenshot of the Galaxy Cloudman web interface. The main content area is partially obscured by a modal dialog box titled "Initial Cluster Configuration".

Initial Cluster Configuration

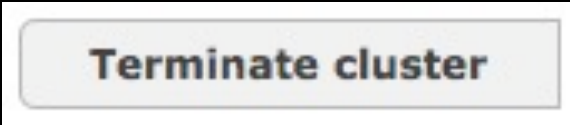
Welcome to Galaxy Cloudman. This application will allow you to manage this cluster and the services provided within. To get started, choose the type of cluster you'd like to work with and specify the size of your persistent data storage, if any.

- Start a full Galaxy Cluster. Specify initial storage size (in Gigabytes)
 GB
- Share-an-instance
 Shared instance bucket path
- Data volume and SGE only. Specify initial storage size (in Gigabytes)
 GB
- SGE Only. No persistent storage created.

[Hide extra options](#)

The background interface shows the "Galaxy Cloudman Console" with a "Status" section containing fields for "Cluster name", "Disk status", "Worker status", "Service status", and "External Logs". A "Termination" button is also visible.

Don't waste the resources

- Once the need for a given cluster subsides,
 - you can always start it
back up
- Data is preserved while a cluster is down

Benefits of the CloudMan architecture

- Minimum setup time and cost
 - No need for an external broker
- Data persistence
- Built-in support for managing the oscillating data volume
- Self-contained deployment
 - Customizable instances: CloudMan as PaaS
 - Versioning of tools, data, and configurations

Acknowledgements

Funding from NHGRI, NSF, Penn State, Emory,
and Pennsylvania Department of Health

The Galaxy Team at Penn State and Emory



Two full days of presentations and discussion

Early registration ends April 24

galaxy.psu.edu/gcc2011

Questions & Comments

Try your own cluster; it takes only 5 minutes and less than \$1.

Complete instructions available at <http://usegalaxy.org/cloud>