

PENNSTATE®



Deploying Galaxy on the Cloud

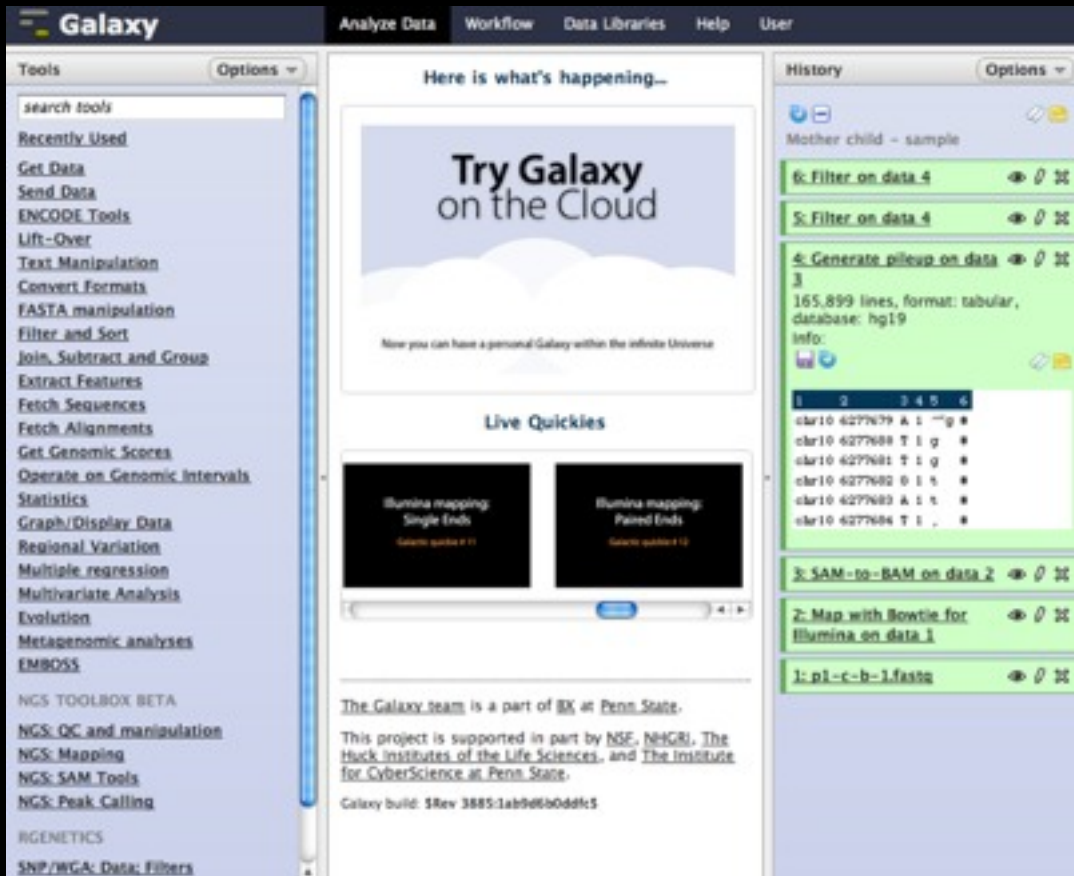
Enis Afgan, Dannon Baker, Nate Coraor, Anton
Nekrutenko, James Taylor

Bioinformatics Open Source Conference, July 9, 2010, Boston, MA



EMORY

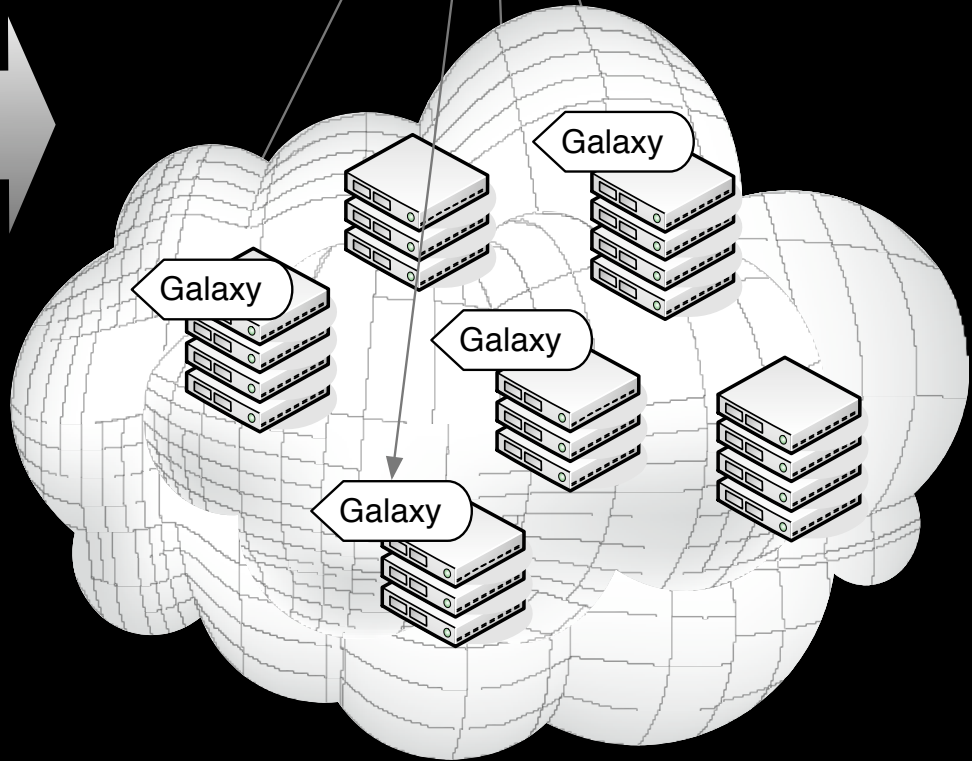
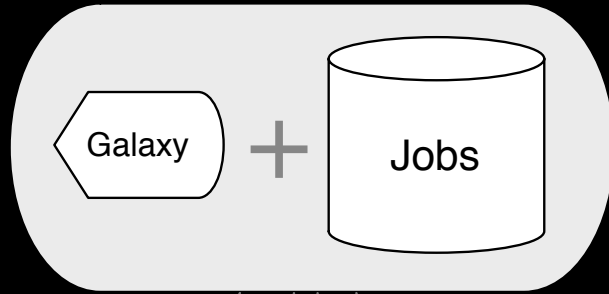
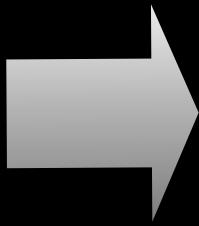
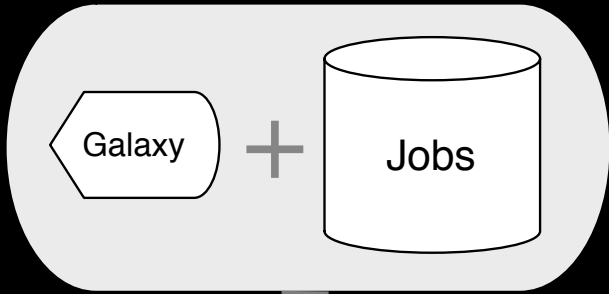
Galaxy: accessible analysis system



The screenshot displays the Galaxy web interface. On the left is a 'Tools' sidebar with a search bar and a list of tool categories including 'Recently Used', 'Get Data', 'Send Data', 'ENCODE Tools', 'Text Manipulation', 'Convert Formats', 'FASTA manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Extract Features', 'Fetch Sequences', 'Fetch Alignments', 'Get Genomic Scores', 'Operate on Genomic Intervals', 'Statistics', 'Graph/Display Data', 'Regional Variation', 'Multiple regression', 'Multivariate Analysis', 'Evolution', 'Metagenomic analyses', 'EMBOSS', 'NGS TOOLBOX BETA', 'NGS: QC and manipulation', 'NGS: Mapping', 'NGS: SAM Tools', 'NGS: Peak Calling', 'RGENETICS', and 'SNP/WGA: Data: Filters'. The main content area is titled 'Here is what's happening...' and features a 'Try Galaxy on the Cloud' banner, 'Live Quickies' for 'Illumina mapping: Single Ends' and 'Illumina mapping: Paired Ends', and a footer with project information and a Galaxy build ID. On the right is a 'History' panel showing a workflow for 'Mother child - sample' with steps: 1: pl-c-b-1.fastq, 2: Map with Bowtie for Illumina on data 1, 3: SAM-to-BAM on data 2, 4: Generate pileup on data 3 (165,899 lines, format: tabular, database: hg19), 5: Filter on data 4, and 6: Filter on data 4.

- 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
- Workflow system

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



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

Current Status

- Deployment of Galaxy on Amazon Web Services Cloud
 - Requires no computational expertise, no infrastructure, no software
- Support for dynamic resource scaling
- Support for dynamic storage
- Automated configuration of the Galaxy Cloud machine image

- Deploy a Galaxy cluster in minutes!

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 Cloud web interface on the master instance to manage the cluster size

2. Start an EC2 Instance

The image shows a composite screenshot of the AWS Management Console. At the top left is the Amazon Web Services logo. The navigation bar includes links for AWS, Products, Developers, Community, Support, and Account (highlighted with a red box). Below the navigation bar, the 'Your Account' section is visible, with 'Security Credentials' highlighted by a red box. The main content area is the 'Amazon EC2 Console Dashboard' for the US East region. It features 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 settings:

- 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 Cloud Console interface. At the top left is the 'Galaxy' logo, and at the top right are links for 'Info: report bugs | wiki | screencasts'. The main heading is 'Galaxy Cloud Console'. Below it is a welcome message: 'Welcome to the Galaxy Cloud Console. This application allows you to manage this instance of Galaxy. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, Galaxy will start and you will be able to see its standard interface and you will be able to add and remove 'worker' nodes on which jobs are run.'

There are two buttons: 'Terminate cluster' on the left and 'Access Galaxy' on the right. Below these is a 'Status' section with the following information:

- Cluster name: gc_dev1
- Disk status: 0Gb / 0Gb
- Instance status: Idle: 0 Available

A modal dialog box titled 'Initial Volume Configuration' is centered on the screen. It contains the text: 'This initial configuration is required to start Galaxy. The data volume created will be used for storing all of data uploaded and analyzed through Galaxy.' Below this text is a label 'Permanent storage size (1-1000GB):' followed by an empty input field. At the bottom of the dialog is a button labeled 'Create Data Volume'.

At the bottom of the console, there is a 'Cluster status log' section with a green plus icon to its right.

Galaxy Cloud Console

Welcome to the Galaxy Cloud Console. This application allows you to manage this instance of Galaxy. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, Galaxy will start and you will be able to see its standard interface and you will be able to add and remove 'worker' nodes on which jobs are run.

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

Status

Cluster name: gc_dev1

Disk status: 49M / 1014M (5%) 

Instance status: Idle: 0 Available: 0 Requested: 0



 Filesystems  Database  Scheduler  Galaxy

Cluster status log

```
18:18:13 - Configuring SGE...
18:18:13 - Setting up SGE.
18:18:19 - Successfully setup SGE; configuring SGE
18:18:19 - Completed initial cluster configuration.
18:18:53 - Creating user data volume of size '1'GB.
18:18:54 - Saving newly created user data volume ID (vol-d03154b9) to user's bucket
'gc-42d4f99232c4b8060942debdcf76bd3d' within file 'persistent-volumes-latest.txt'.
18:18:54 - Attaching user data volume 'vol-d03154b9' to instance as device '/dev/sdd'.
18:19:00 - Volume 'vol-d03154b9' attached to instance 'i-2b9c6f41' as device '/dev/sdd'
18:19:00 - Creating user data file system 'galaxyData' on device '/dev/sdd'.
18:19:02 - Configuring PostgreSQL with a database for Galaxy...
18:19:20 - Setting up Galaxy
18:19:20 - Starting Galaxy...
```


Tools

- [Get Data](#)
- [Text Manipulation](#)
- [Filter and Sort](#)
- [Join, Subtract and Group](#)
- [Operate on Genomic Intervals](#)
- [Graph/Display Data](#)



NGS TOOLBOX BETA

- [NGS: QC and manipulation](#)
- [NGS: Mapping](#)
- [NGS: SAM Tools](#)

Welcome to Galaxy on the Cloud

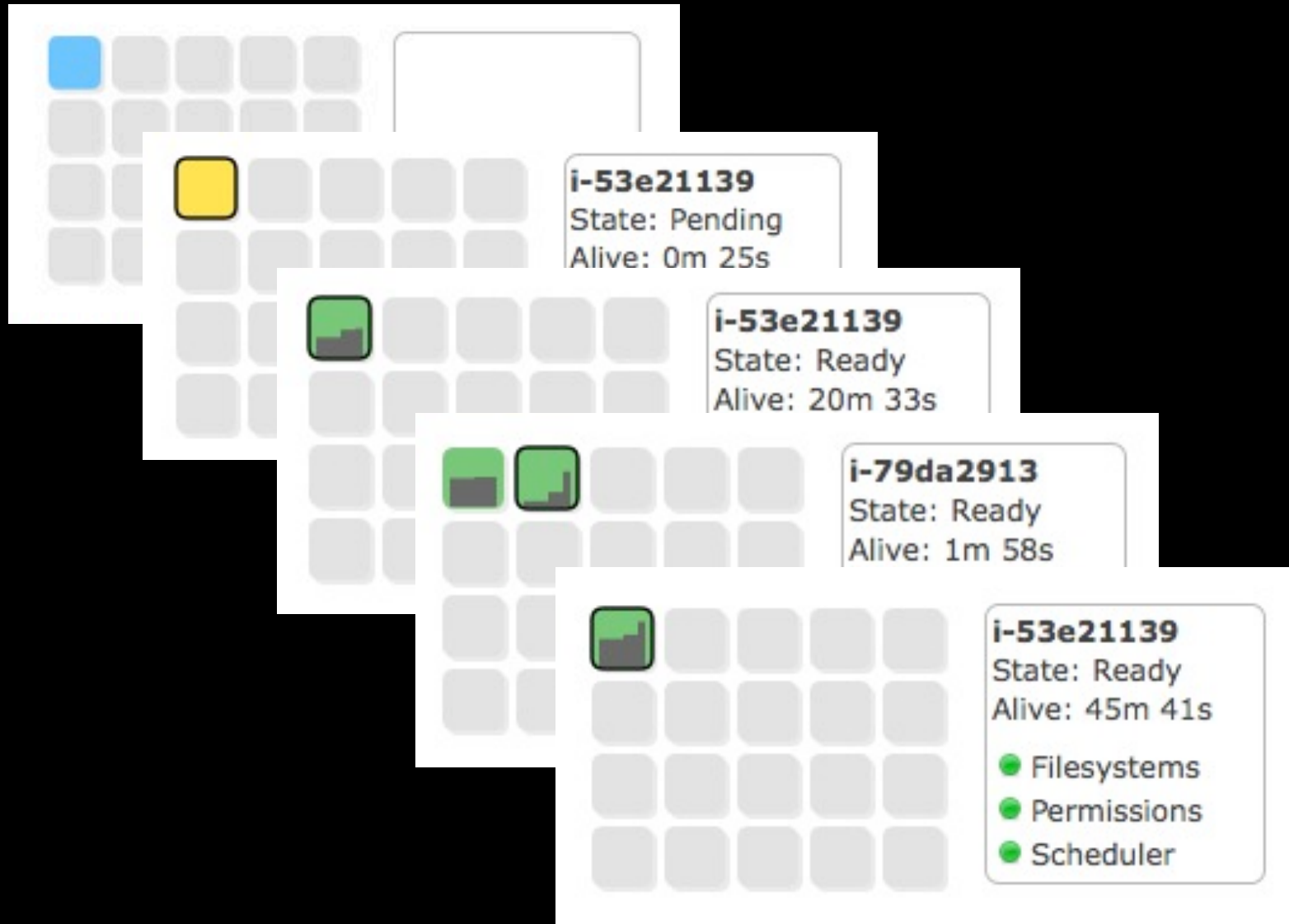


History Options

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

4. Grow and Shrink



Grow Storage

Status

Cluster name: gc_dev1

Disk status: 832M / 1014M (83%)

Instance status: Idle: 0 Available: 1 Requested: 1

i-53e21139
State: Ready
Alive: 45m 41s

- Filesystems
- Permissions
- Scheduler

● Filesystems ● Database ● Scheduler ● Galaxy
● galaxyData:0 ● galaxyTools ● galaxyIndices

1. Stop services
2. Detach volume
3. Snapshot

Status

Cluster name: gc_dev1

Disk status: 244M / 5.0G (5%)

Instance status: Idle: 1 Available: 1 Requested: 1


i-53e21139
State: Ready
Alive: 49m 42s

- Filesystems
- Permissions
- Scheduler

● Filesystems ● Database ● Scheduler ● Galaxy

4. New volume
5. Grow file system
6. Resume services

Clean Up

- Once the need for a given cluster subsides,
 - you can always start it back up
- Data is preserved while a cluster is down
- Complete the shut down process by terminating the master instance from the AWS console

What is Coming

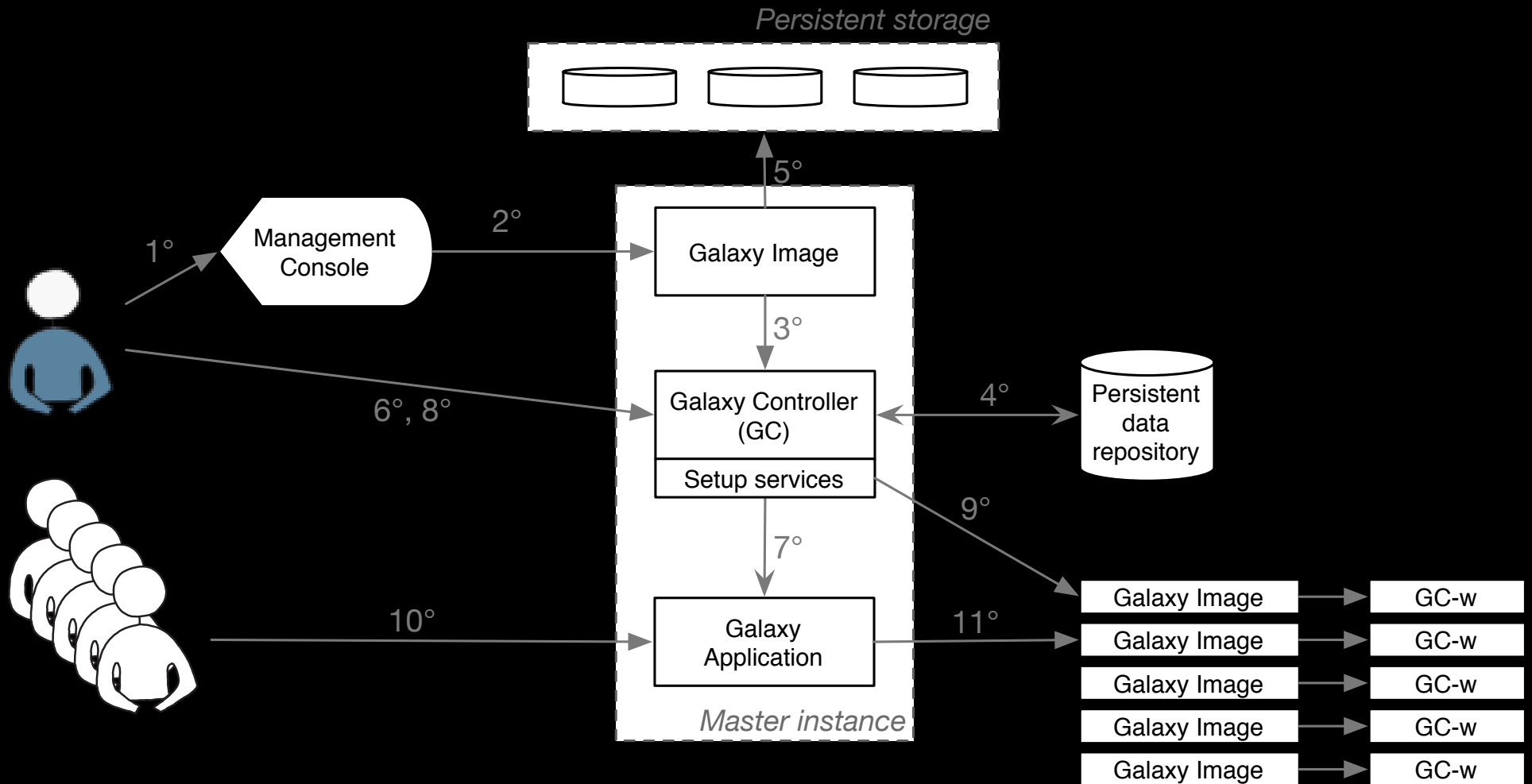
- Automatic cluster scaling
 - Based on workload customization
- Automatic job splitting/parallelization

Questions & Comments

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

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

A Little More GC Details



Cloud or No Cloud?

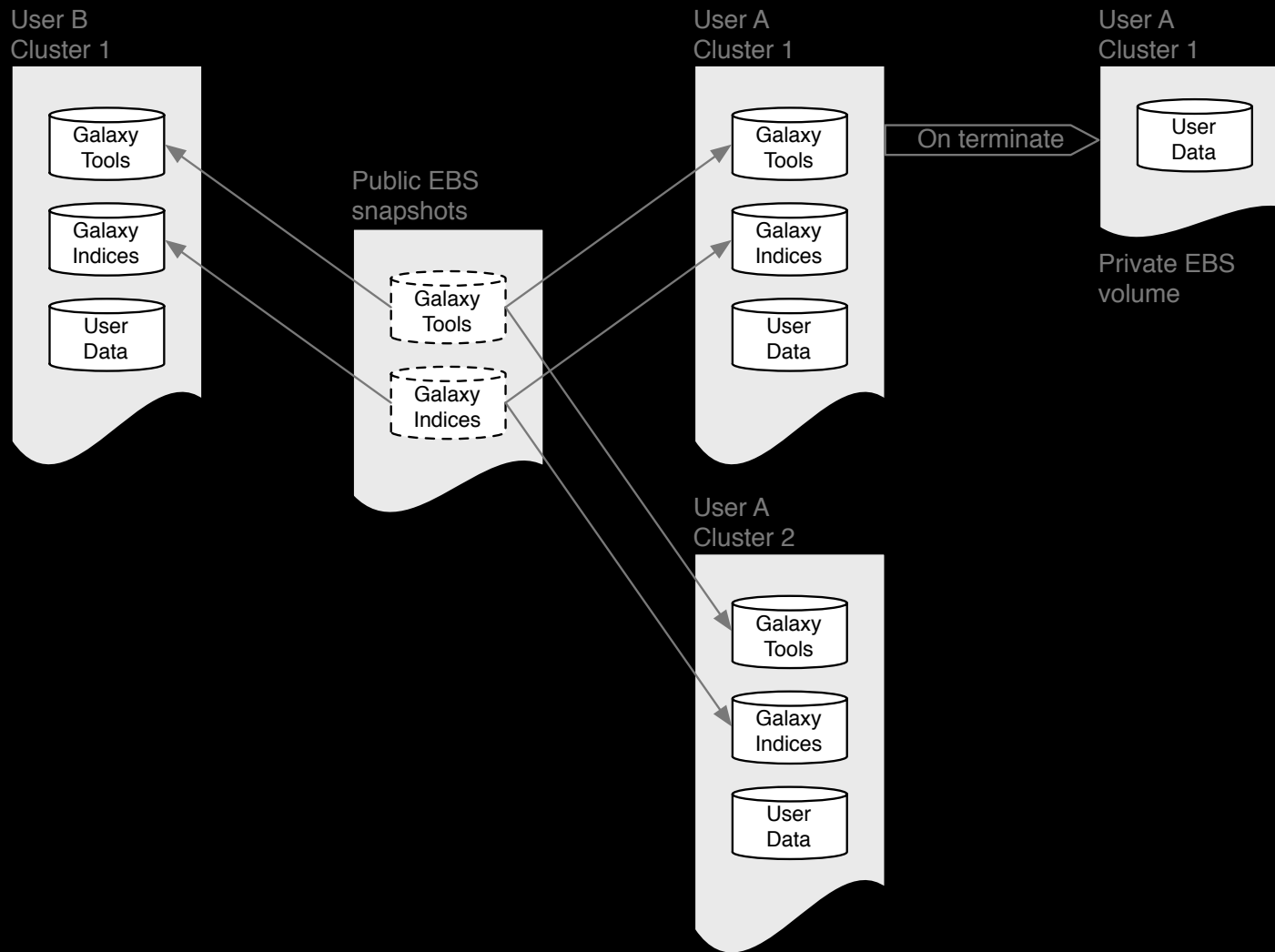
Pros

- Consumption based cost - cost reduction?
- Better utilization of resource
- Management done by cloud provider
- Faster deployment time
- Dynamic scalability

Cons

- Not a silver bullet
- Expensive for 24/7 use
- Offers scalability in terms of infrastructure, applications are still sequential
- The data transfer problem?
- Security?

Enabling Persistence



Enabling Versioning

