

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



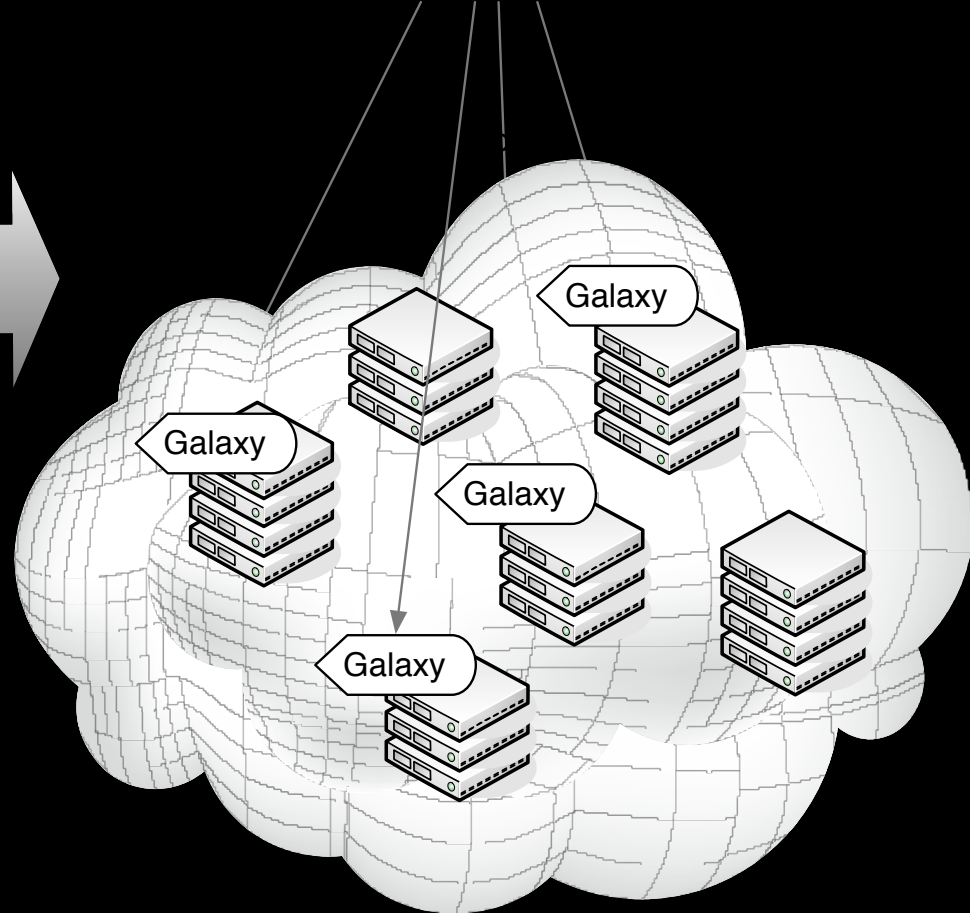
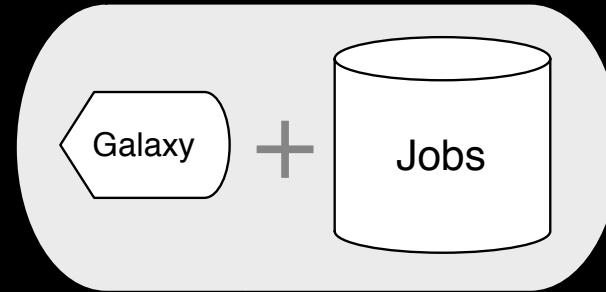
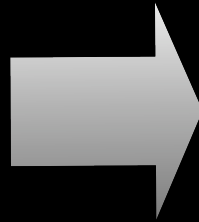
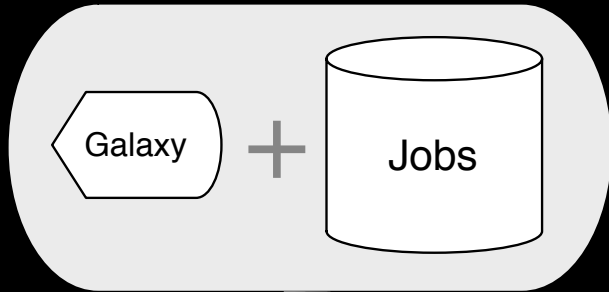
EMORY

Galaxy: accessible analysis system



- 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

amazon web services

Sign in to the AWS Management Console | Create an AWS Account

AWS | Products | Developers | Community | Support | **Account**

Your Account

Account Activity
View current charges and account activity, by service and usage type.

Consolidated Billing
Sign up to receive one bill for multiple AWS accounts, and add or remove accounts from your bill.

DevPay Activity
View revenue and costs for your manageable Amazon DevPay products.

Payment Method
View and edit current payment method, as well as add new payment methods.

Personal Information
View and edit personal or self-communication profile subscriptions.

Security Credentials
AWS uses two types of authentication requests to protect your account.

Usage Reports
Download customizable usage reports for the services you are subscribed to.

Navigation

Region: US East

EC2 Dashboard

INSTANCES

- Instances
- Spot Requests

IMAGES

- AMIs
- Bundle Tasks

ELASTIC BLOCK STORE

- Volumes
- Snapshots

NETWORKING & SECURITY

- Elastic IPs
- Security Groups
- Key Pairs
- Load Balancers

Introducing Amazon Relational Database Service

Amazon RDS is a fully managed relational database in the cloud.

Learn More...

aws.amazon.com | AWS | Products | Developers | Community | Support | Account

Welcome, Enis Afgan | Settings | Sign Out

Amazon S3 | **Amazon EC2** | Amazon Elastic MapReduce | Amazon CloudFront | Amazon RDS

Navigation

Region: US East

Amazon EC2 Console Dashboard

Getting Started

To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.

Launch Instance

Note: Your instances will be launched in the US East (Virginia) region.

Service Health

Current Status

Amazon EC2 (US East - N. Virginia)

My Resources

You are using the following Amazon EC2 resources in the US East (Virginia) region:

- 0 Running Instances
- 0 Elastic IPs
- 6 EBS Volumes
- 12 EBS Snapshots

Request Instances Wizard

CHOOSE AN AMI | INSTANCE DETAILS | CREATE KEY PAIR | CONFIGURE FIREWALL | REVIEW

Please review the information below, then click Launch.

AMI: Other Linux AMI ID ami-ed03ed84 (x86_64) Edit AMI

Number of Instances:

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

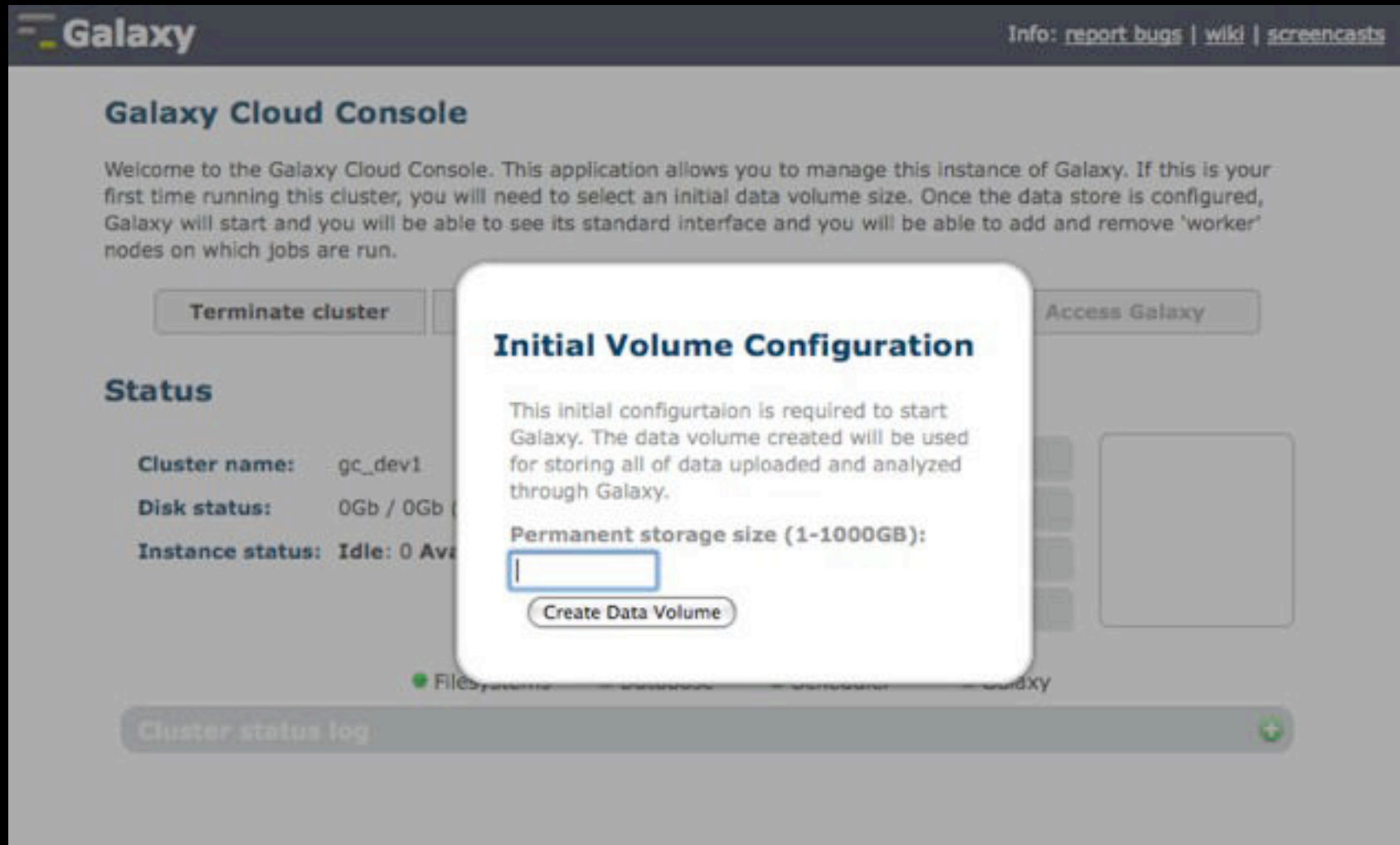
User Data: testGC1IAKIAJKQ13RT... Edit Advanced Details

Key Pair Name: galaxy_keypair Edit Key Pair

Security Group(s): default, galaxyWeb Edit Firewall

Back Launch

3. Configure Your Cluster



The screenshot displays the Galaxy Cloud Console interface. At the top, the 'Galaxy' logo is on the left, and links for 'Info: report bugs | wiki | screencasts' are on the right. The main heading is 'Galaxy Cloud Console'. Below it, a welcome message states: '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.'

Two buttons are visible: 'Terminate cluster' on the left and 'Access Galaxy' on the right. Below these, the 'Status' section shows cluster details: 'Cluster name: gc_dev1', 'Disk status: 0Gb / 0Gb', and 'Instance status: Idle: 0 Available'. A modal dialog 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 text input field. At the bottom of the dialog is a 'Create Data Volume' button. In the background, a 'Cluster status log' section is partially visible at the bottom.

Galaxy Info: [report bugs](#) | [wiki](#) | [screencasts](#)

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 **Access Galaxy**

Status

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

Initial Volume Configuration

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.

Permanent storage size (1-1000GB):

Create Data Volume

Cluster status log

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.


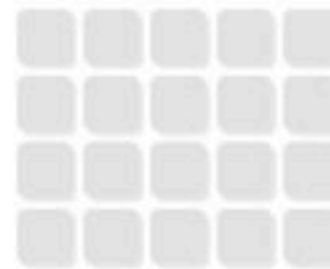
[Terminate cluster](#)[Add instances ▼](#)[Remove instances](#)[Access Galaxy](#)

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...
```

Galaxy

http://ec2-75-101-213-19.compute-1.amazonaws.com/

Google

Galaxy

Analyze DataWorkflowData LibrariesHelpUser

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](#)

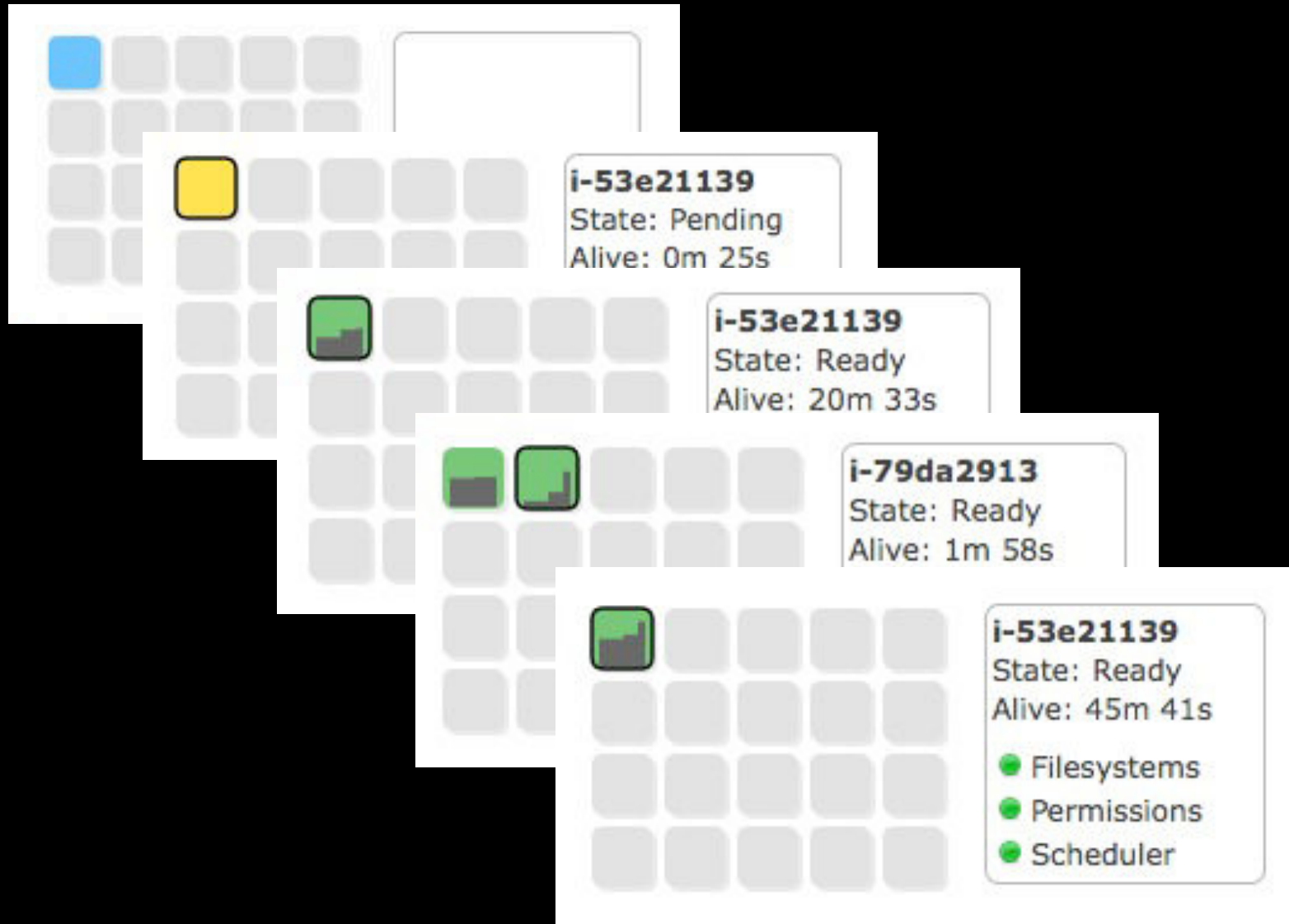
[NGS: SAM Tools](#)

Welcome to Galaxy on the Cloud

HistoryOptions

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

Status

Cluster name: gc_dev1

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

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



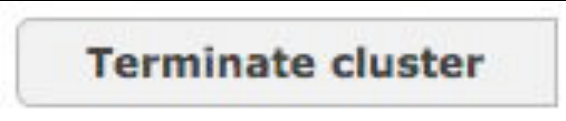
i-53e21139
State: Ready
Alive: 49m 42s

● Filesystems
● Permissions
● Scheduler

● Filesystems ● Database ● Scheduler ● Galaxy

1. Stop services
2. Detach volume
3. Snapshot
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

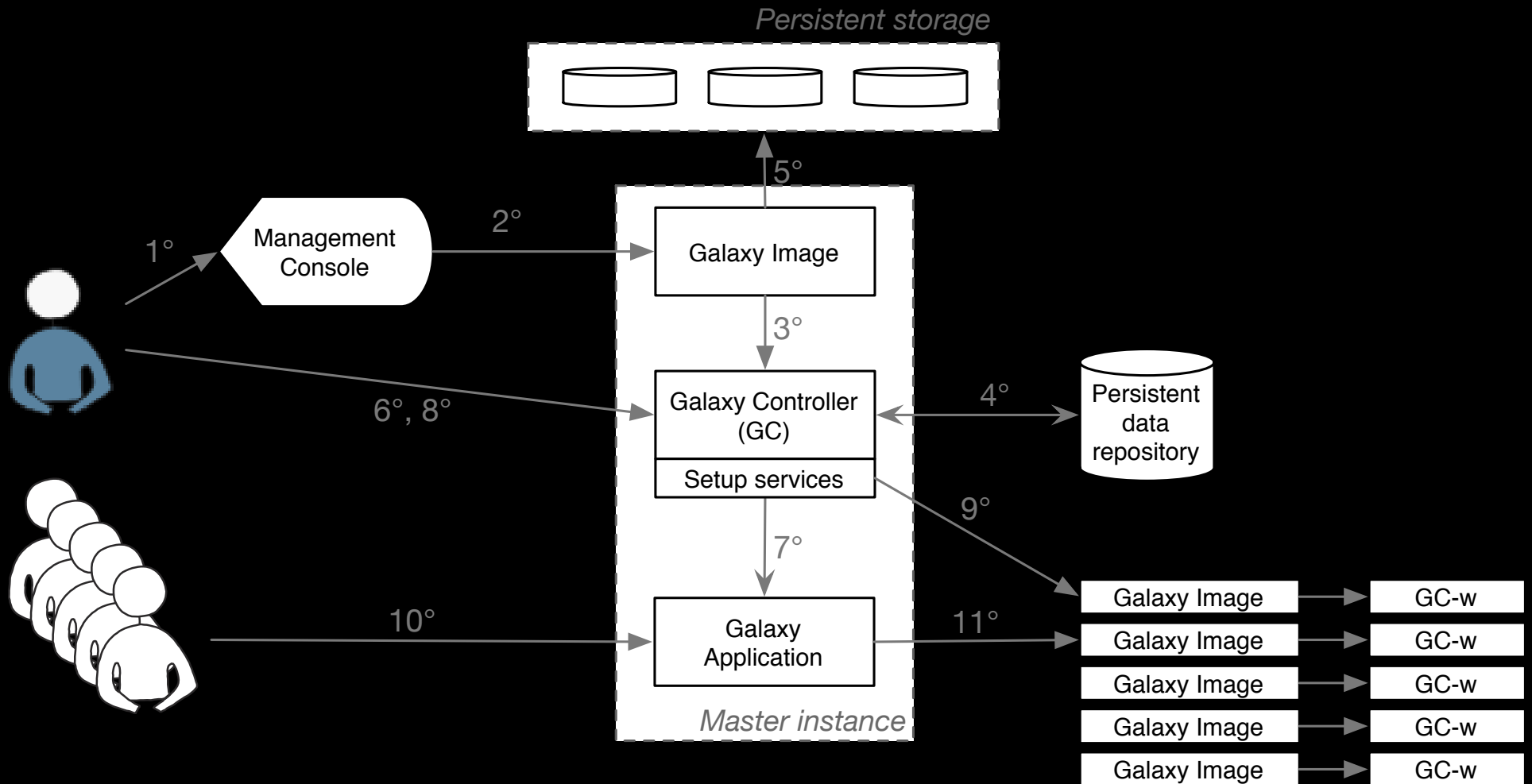
- Automatic cluster scaling
 - Based on workload customization
- Automatic job splitting/parallelization

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?

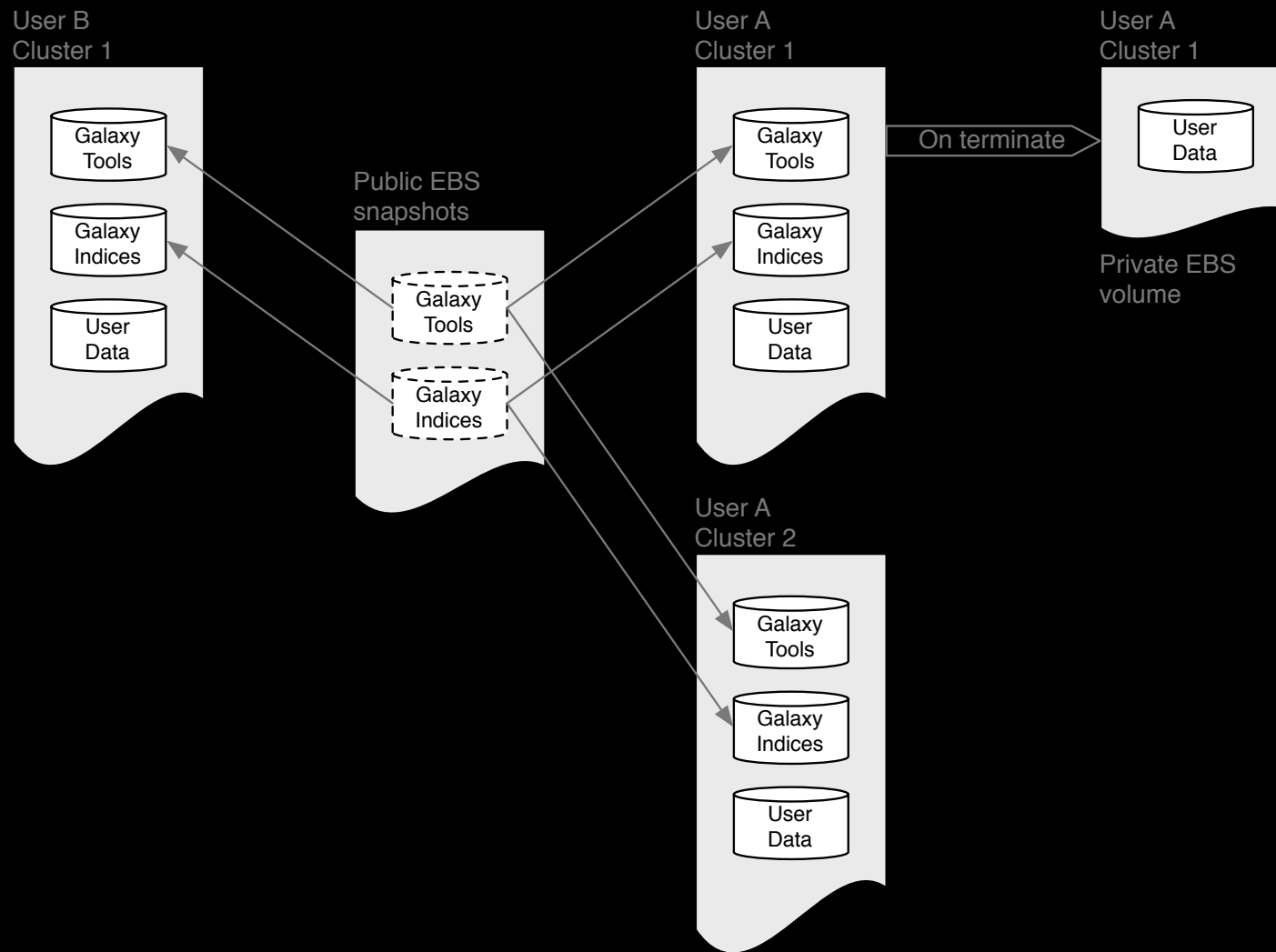
Pros

- 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

