Deploying Galaxy on the Cloud
Enis Afgan and Dannon Baker
Cloud Computing

- Dynamically scalable shared resources accessed over a network
- Private, public, or hybrid
- Virtually unlimited resources: storage, computing, services
- Only pay for what you use
Approaches to Cloud Computing

- **SalesForce, Google Docs, Zoho, web email**
- **RightScale, Google App Engine, Microsoft Azure**
- **Amazon Web Services (AWS), GoGrid, Rackspace**
 Galaxy on the Cloud

• Enable execution of Galaxy on cloud infrastructures
  • Labs do not have to house compute resources
  • Support variable analysis data volume
  • Web-based Galaxy instantiation
• Goal is to keep Galaxy use unchanged but deliver flexibility and job performance improvement
So, who is it for and how does it work?

- Ideal for small labs and individual researchers

A. Users in different labs
B. Isolated Galaxy instance(s)
C. Dense data center
## Cloud or No Cloud?

<table>
<thead>
<tr>
<th><strong>Pros</strong></th>
<th><strong>Cons</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Consumption based cost - cost reduction?</td>
<td>• Not a silver bullet</td>
</tr>
<tr>
<td>• Better utilization of resource</td>
<td>• Expensive for 24/7 use</td>
</tr>
<tr>
<td>• Management done by cloud provider</td>
<td>• Offers scalability in terms of infrastructure, applications are still</td>
</tr>
<tr>
<td>• Faster deployment time</td>
<td>sequential</td>
</tr>
<tr>
<td>• Dynamic scalability</td>
<td>• The data transfer problem?</td>
</tr>
<tr>
<td></td>
<td>• Security?</td>
</tr>
</tbody>
</table>
Current Status

- Deployment of Galaxy on AWS
- Requires no computational expertise, no infrastructure, no additional software
- Support for dynamic resource scaling
- Deploy a Galaxy cluster in minutes!
Deploying Galaxy on the AWS Cloud

1. Create an AWS account and sign up for Elastic Compute Cloud (EC2) and Simple Storage Service (S3) services
2. Use AWS Management Console to start an EC2 instance
3. Use GC web interface on started EC2 instance to start a desired number of compute instances
4. Enjoy your personal instance of Galaxy on the cloud
2. Start an EC2 Instance

User data: user account info and cluster name
3. Configure Your Cluster

The Galaxy cloud console allows you to manage this instance of Galaxy. From here you can start the main Galaxy interface (including an initial set of "worker" nodes on which jobs will be run), as well as add and remove workers while the main interface is running.

Scale

- Add more instances
- Remove idle instances

Status

- Cluster name: testGC1
- Cluster status: Waiting for user action
- Instance status: Idle: 0 Available: 0 Requested: 0

Cluster status log:

17:33:12 - Volume 'vol-2553ee4c' attached to instance 'i-7f661c14' as device '/dev/sdi'
17:33:12 - 'galaxyData' data volume not there yet; will get it at worker instance start time.
17:33:18 - Successfully mounted data volume 'galaxyTools' attached as device '/dev/sdg' to file system
/mnt/galaxyTools'
17:33:19 - Successfully mounted data volume 'galaxyIndices' attached as device '/dev/sdi' to file system
/mnt/galaxyIndices'
17:33:19 - Configuring PostgreSQL with a database for Galaxy...
17:33:19 - '/mnt/galaxyData' directory doesn't exist yet; will configure PostgreSQL later.
17:33:24 - Configuring SGE...
17:33:24 - Setting up SGE.
17:33:32 - Successfully setup SGE; configuring users' SGE profiles.
17:33:32 - Completed initial cluster configuration.
(Starting Workers)
(Couple of Minutes Later)
4. Grow and Shrink
Clean Up

• Once need for given cluster subsides, power it off - you can always start it back up

• Data is preserved while cluster is down

• Complete the shut down process by terminating the master instance from the AWS console
A Little More GC Details

- Management Console
- Galaxy Controller (GC)
- Galaxy Application
- Persistent data repository
- Master instance
- Galaxy Image
- GC-w
- GC-w
- GC-w
- GC-w

Sunday, May 16, 2010
The Summary Picture
What is Coming

- Data storage expansion
- Software update
- Automatic cluster scaling
  - Based on workload customization
- Support for private and hybrid clouds
- Automatic job splitting/parallelization
Questions & Comments

Try your own cluster. Complete instructions available at http://usegalaxy.org/cloud
On the More Technical Side

- **Machine Image (MI)**: a pre-configured unit that can be instantiated - any changes to it are lost at next boot
  - Amazon Machine Image (AMI)
- (EC2) instance
- Live instance of a MI
- Amazon Simple Storage Service (S3)
  - Web-based interface to unlimited data storage
  - Catch: 5GB limit on size of any one file; overall, slow
- Elastic Block Store (EBS volume)
  - Block level storage for use with EC2 instances
  - Acts like a regular disk, persistent across invocations
  - Multiple EBSs can be attached to 1 instance, but each EBS can be attached only to one instance at a time
Enabling Persistence

- User B Cluster 1
  - Galaxy Tools
  - Galaxy Indices
  - User Data
  - Public EBS snapshots

- User A Cluster 1
  - Galaxy Tools
  - Galaxy Indices
  - User Data

- User A Cluster 2
  - Galaxy Tools
  - Galaxy Indices
  - User Data

- On terminate

- User A Cluster 1
  - User Data
  - Private EBS volume

Sunday, May 16, 2010
Enabling Versioning

Public S3 buckets:
- GC-default
- GC-snaps

Private S3 bucket:
- GC-User A, Cluster1
- GC-User A, Cluster2

GC source:
- latest
- prev. versions

Public snap IDs:
- latest
- prev. versions

Public S3 buckets:
- GC-default
- GC-snaps

GC-User A, Cluster1:
- latest GC used
- snaps IDs

GC-User A, Cluster2:
- latest GC used
- snaps IDs