PENNSTATE

1855

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

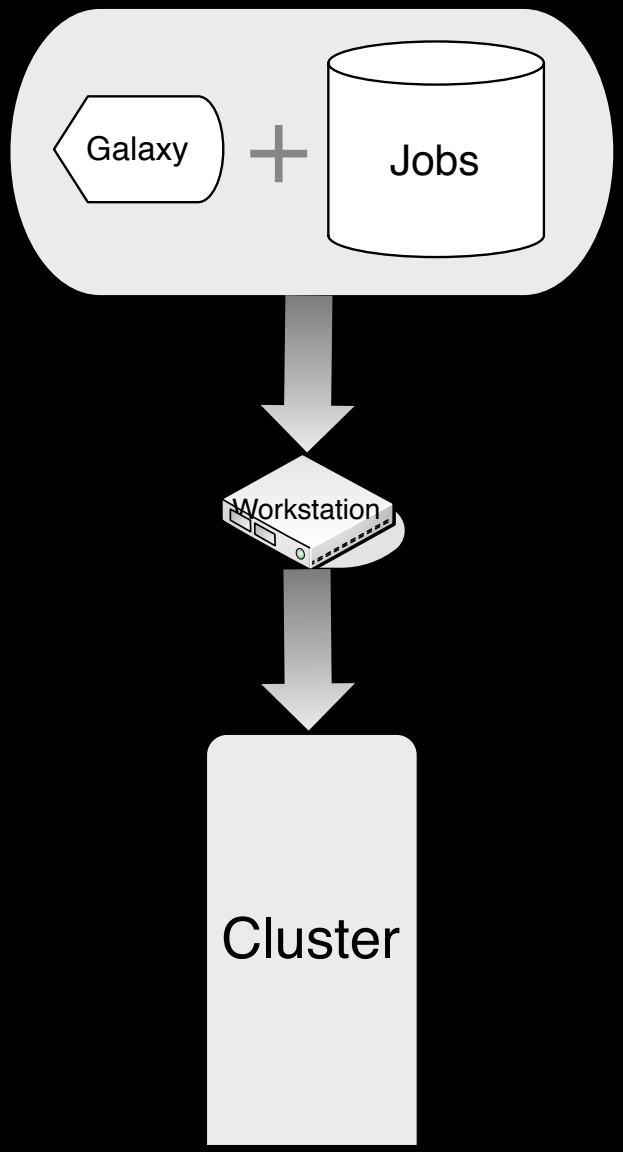
http://usegalaxy.org/

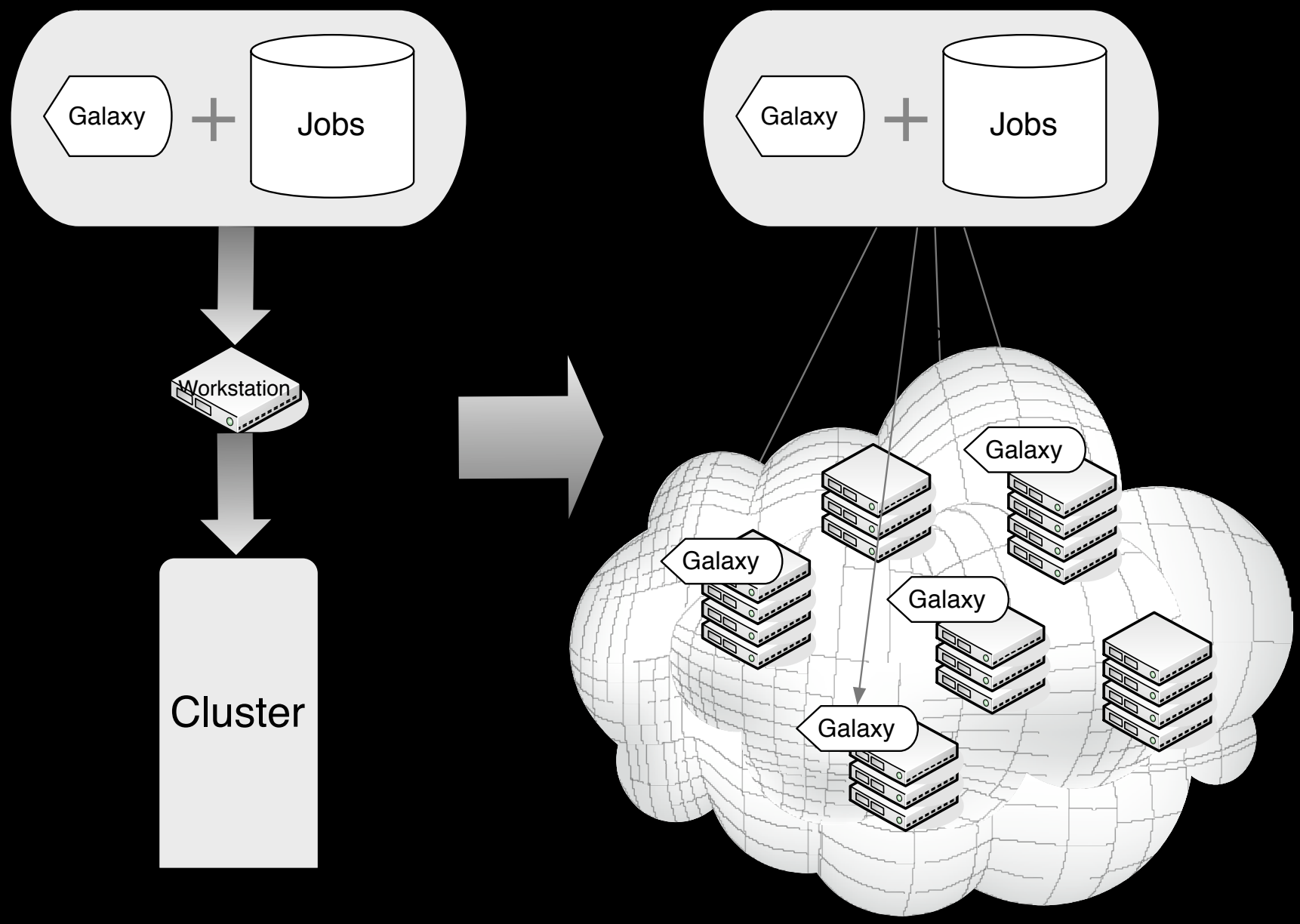
- Easily integrate new tools
- Consistent tool user interfaces automatically generated
- History system facilitates and tracks multistep analyses
- Exact parameters of a step can always be inspected, and easily rerun
- Sharing: analyses, processes, tools, results
- Workflow system
- Data visualization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

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
Jobs
Workstation
Cluster


Galaxy
Jobs
Workstation
Cluster
Galaxy
Jobs
Galaxy
Galaxy
Galaxy
Galaxy

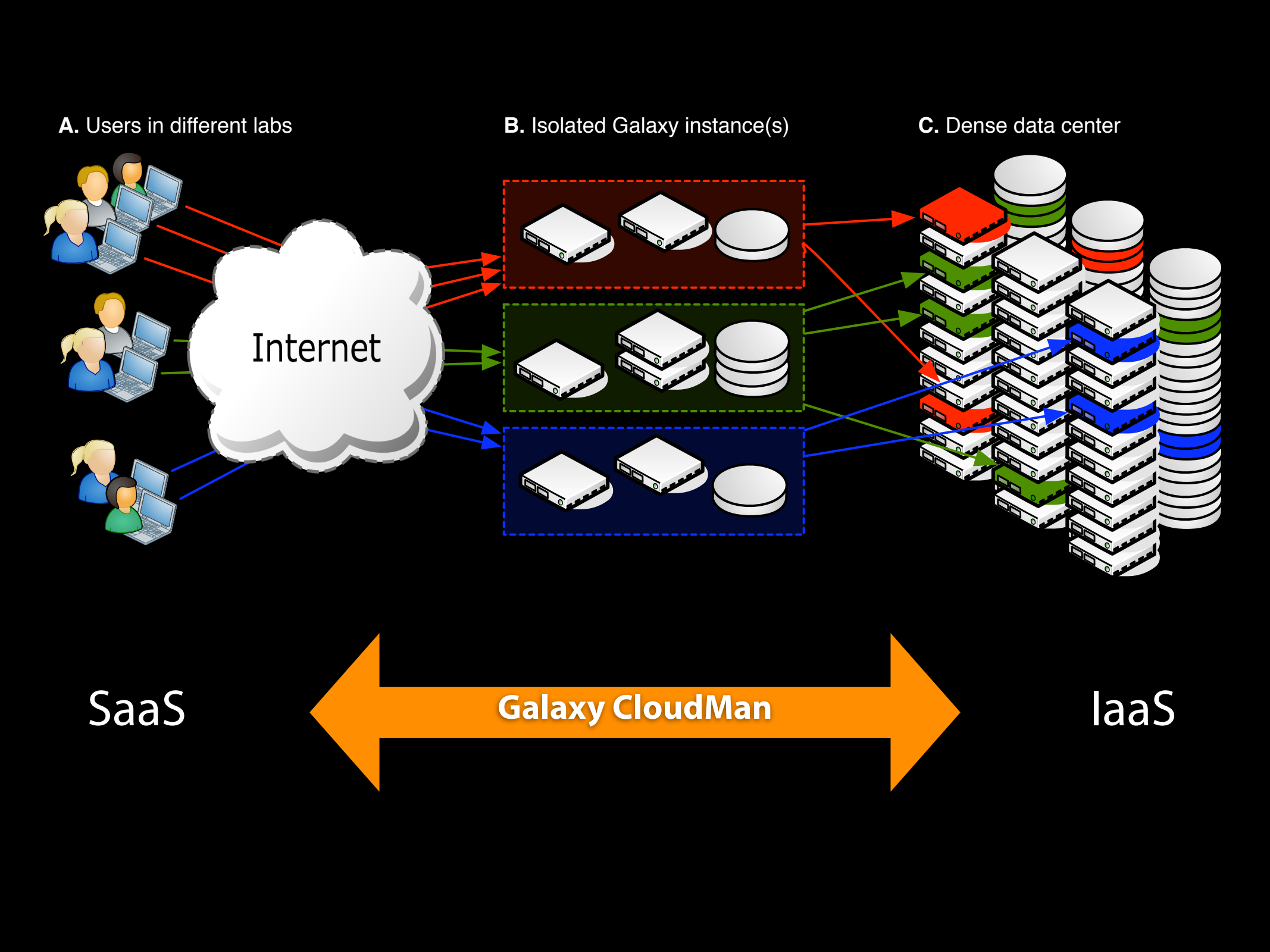
Galaxy on the Cloud

- 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

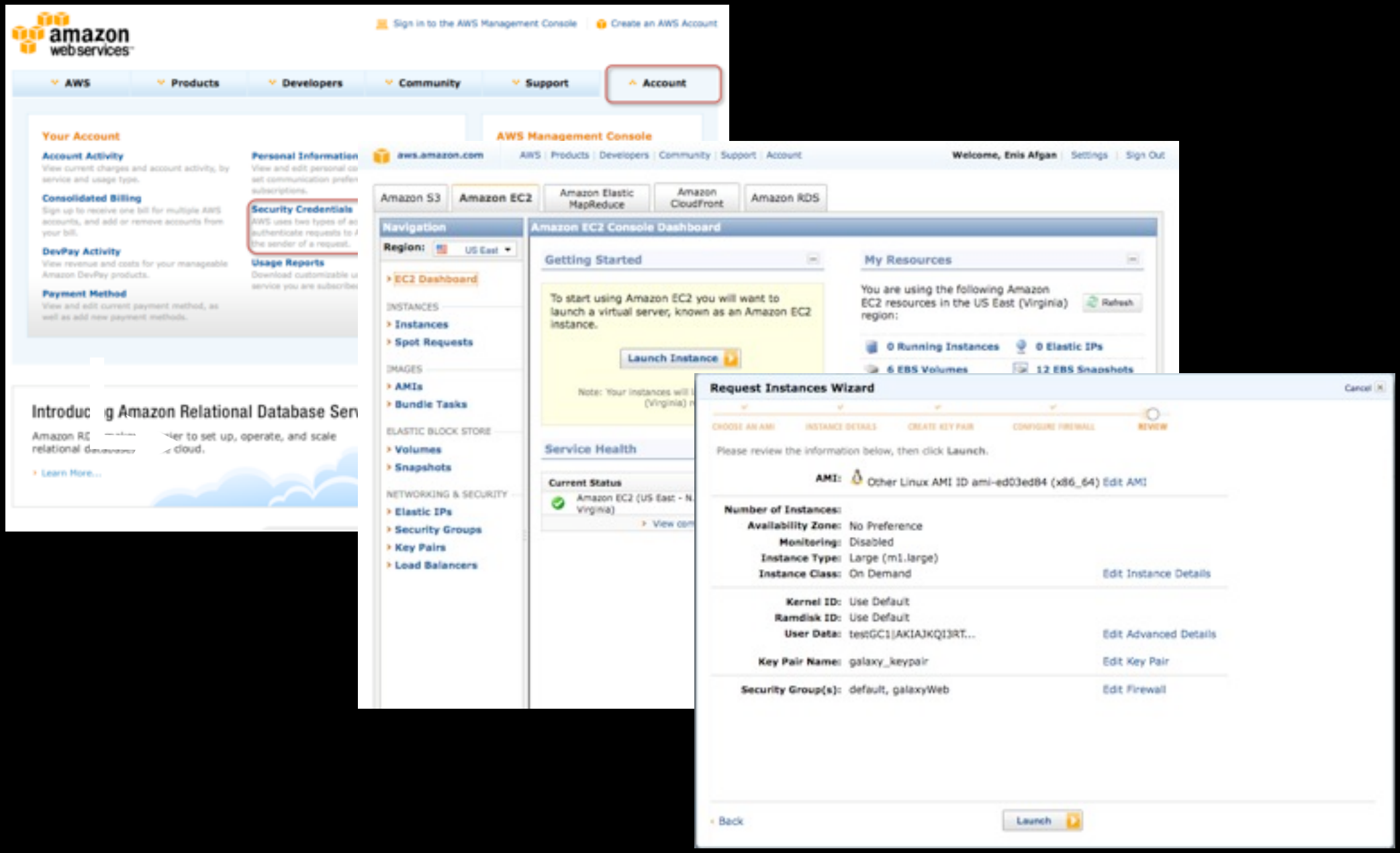
- **Deploy a Galaxy cluster in minutes!**


A. Users in different labs
B. Isolated Galaxy instance(s)
C. Dense data center
Internet
SaaS
Galaxy CloudMan
IaaS

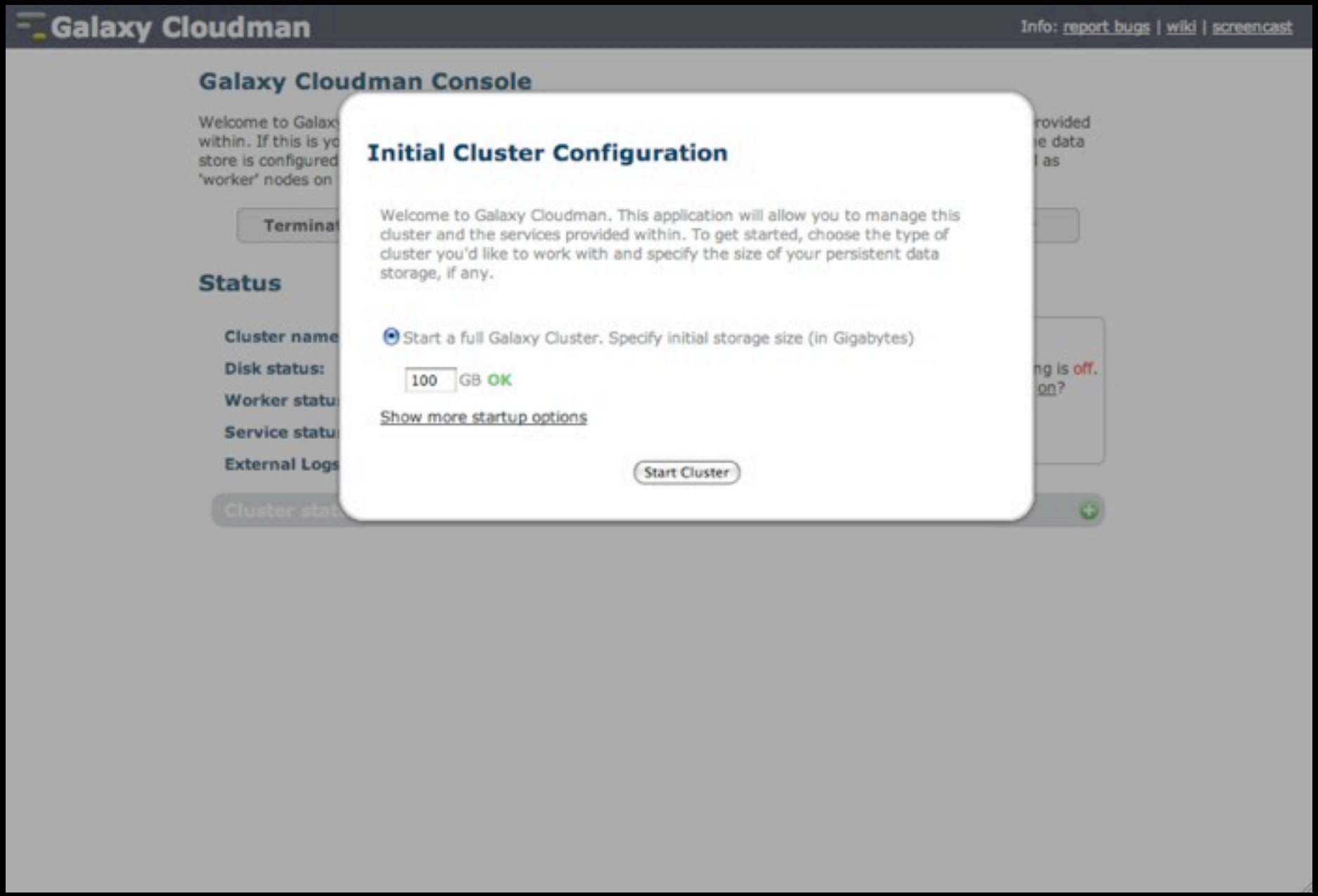
Deploying Galaxy on the AWS Cloud

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



3. Configure Your Cluster



ec2-50-17-119-106.compute-1.amazonaws.com/cloud

Galaxy Cloudman

Info: report bugs | wiki | screencast

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

ec2-50-17-119-106.compute-1.amazonaws.com/root

Galaxy

Analyze Data | Workflow | Shared Data | Visualization | Help | User

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

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

ec2-50-17-119-106.compute-1.amazonaws.com/cloud

Galaxy Cloudman

Info: report bugs | wiki | screencast

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 on.
Turn off?

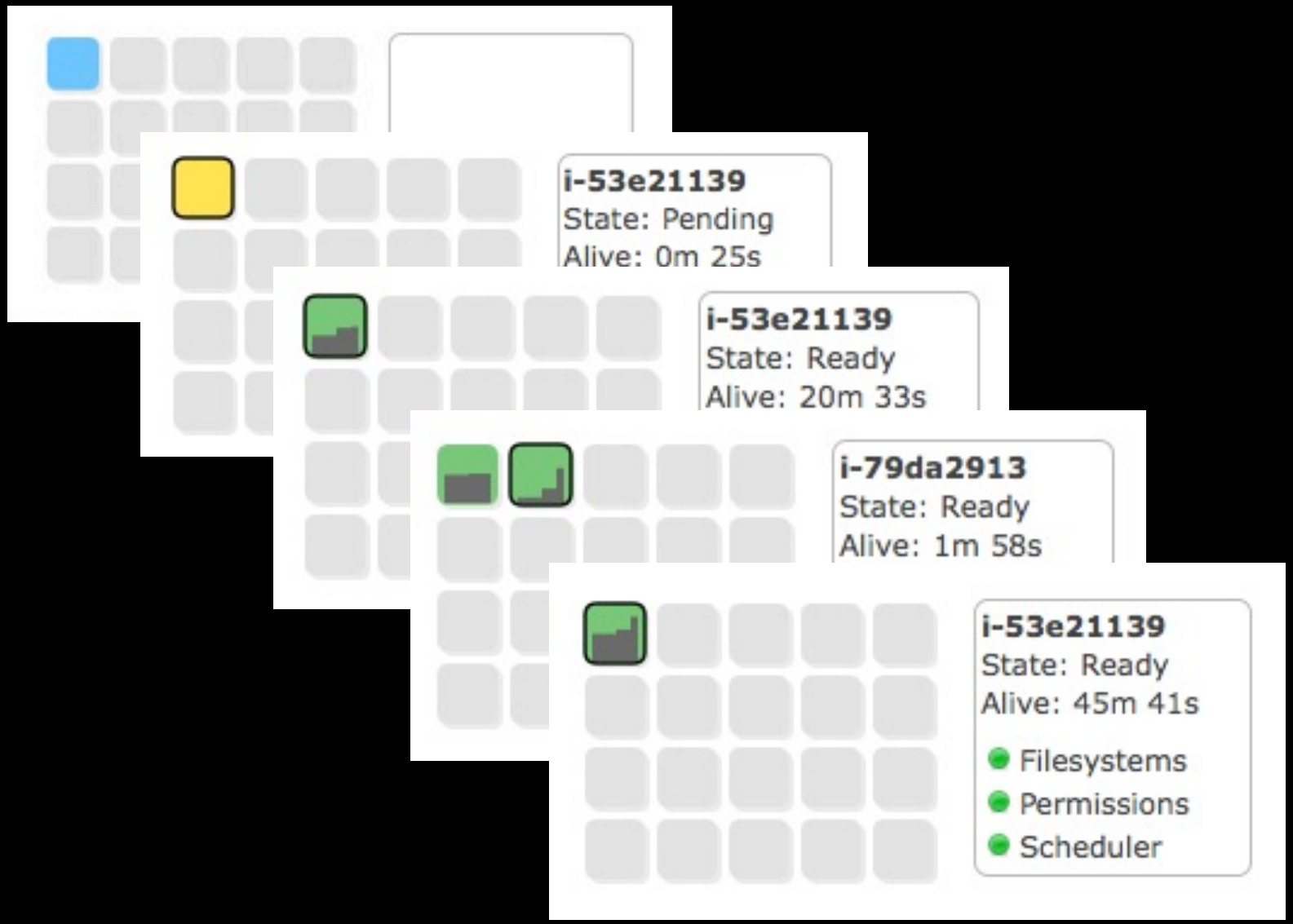
Min nodes: 0
Max nodes: 15
Adjust limits?

Cluster status log

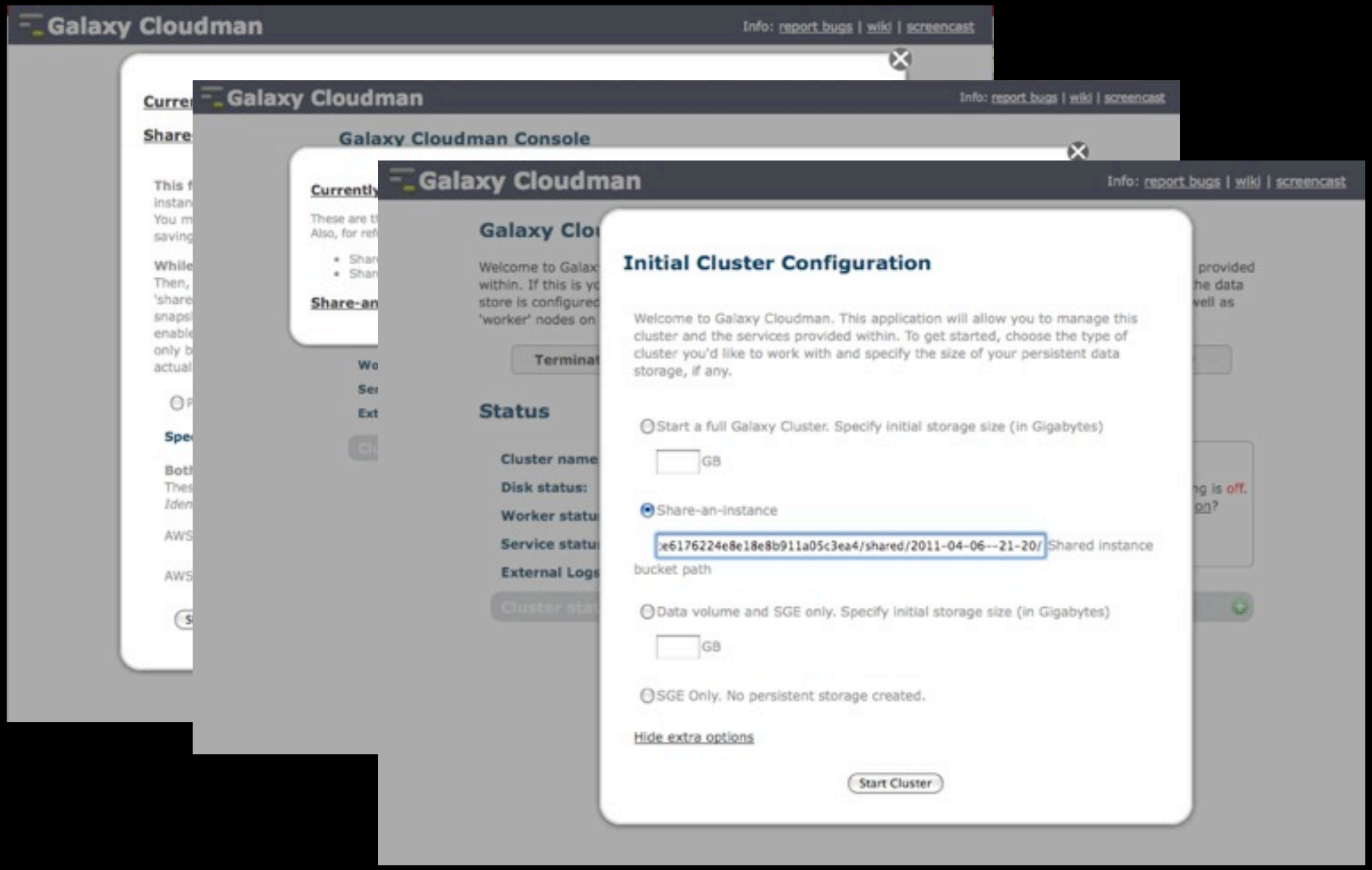
Autoscaling Configuration

Minimum
2
Maximum
15
Type of N
Same as M
Turn au

4. Grow and Shrink



Once an analysis is complete



Don’t waste the resources

- Once the need for a given cluster subsides, Terminate cluster - 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