

CloudMan

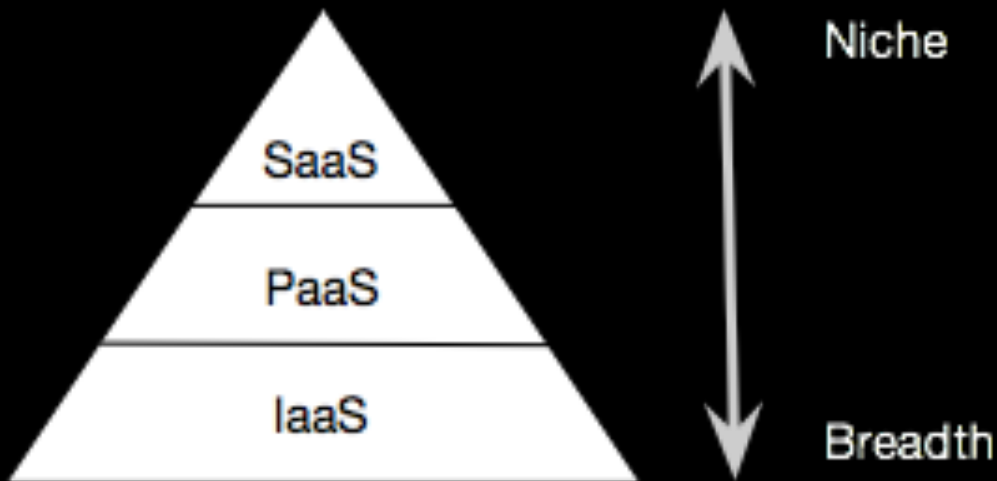
Enabling Ubiquitous, Accessible, Reproducible Research

usecloudman.org

Cloud Computing

- Dynamically scalable shared resources accessed over a network
- Control infrastructure via API
- Private, public, or hybrid
- Virtually unlimited resources: storage, computing, services
 - Only pay for what you use

Approaches to Cloud Computing

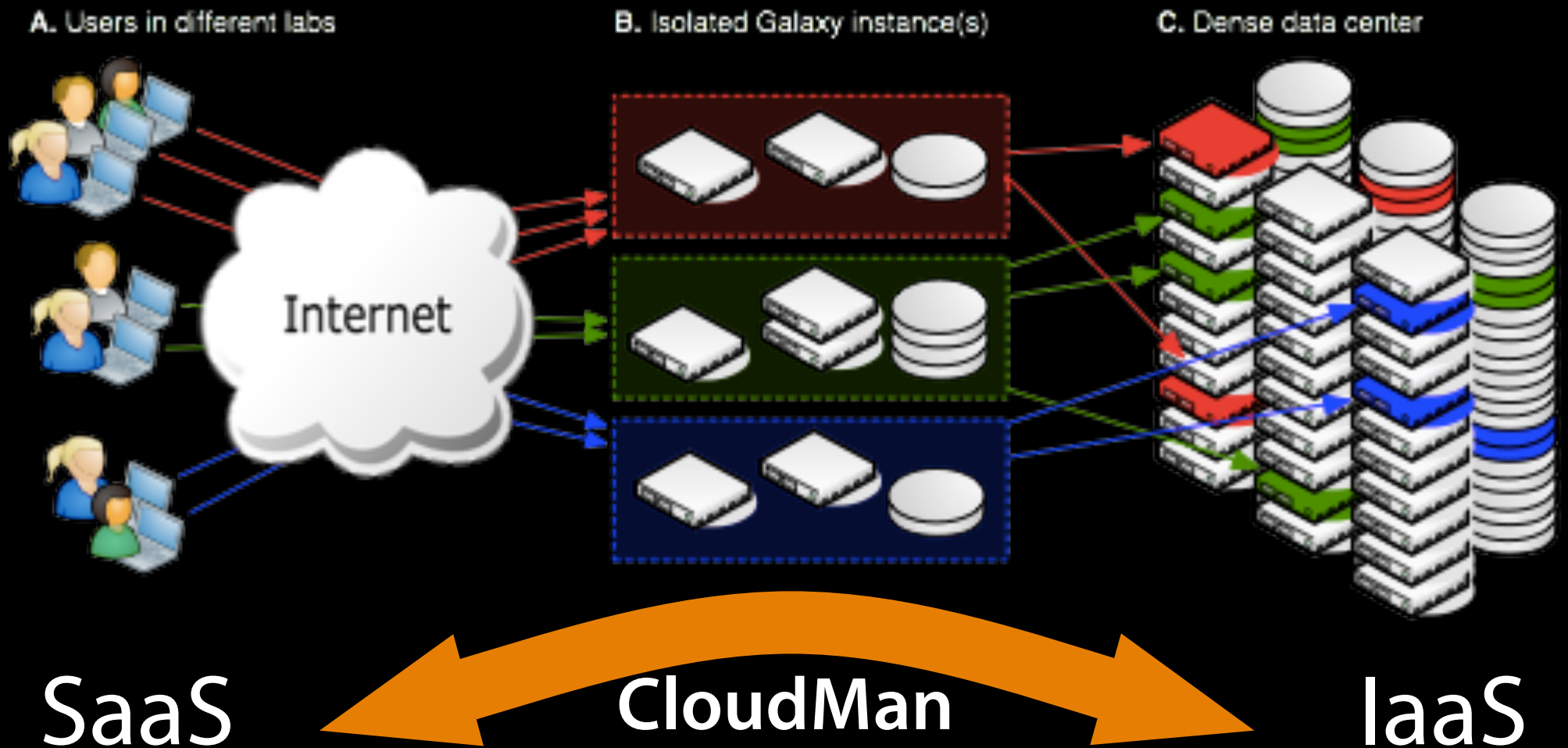


- Salesforce, Google Docs, Zoho, web email
- RightScale, Google App Engine, Microsoft Azure
- Amazon Web Services (AWS), Rackspace, NeCTAR

When to use the cloud?

- Don't have informatics expertise or the infrastructure to run and maintain
- Have variable or particular resource needs
- Cannot upload data to a shared resource
- Need for customization
- Have oscillating data volume

The big picture



What is CloudMan?

A **cloud manager** that orchestrates all of the steps required to provision, manage, and share a compute platform on a cloud infrastructure, all through a web browser.

Deploying a CloudMan Platform

1. An **account** on the supported cloud
2. Start a **master instance** via BioCloudCentral.org or the cloud web console
3. Use the **CloudMan web interface** on the master instance to manage the platform

Start an Instance

bioCloudCentral.org



BioCloudCentral

https://biocloudcentral.herokuapp.com/launch

BioCloudCentral

Easily launch [CloudMan](#), [CloudBioLinux](#) and [Galaxy](#) platforms on Cloud Computing resources (including [Amazon Web Services](#)).

Cluster name

Current Protocols Demo

Name of your cluster used for identification. This can be any name you choose.

Password

.....

Your choice of password, for the CloudMan web interface and accessing the instance via ssh or FreeNX.

Cloud

Amazon (AWS EC2)

Choose from the available clouds. The credentials you provide below must match (ie, exist on) the chosen cloud.

Access key

AKIAJKMSM6GLSW7V2CPA

Your Access Key ID. For the Amazon cloud, available from the [security credentials page](#).

Secret key

EjvlMvij9SLVuxvb9OgaD58qiUXLkEZLa1

Your Secret Access Key. For the Amazon cloud, also available from the [security credentials page](#).

Instance type

Large (4 ECUs / 7.5GB RAM)

Type (ie, virtual hardware configuration) of the instance to start.

[Show advanced startup options](#)

Start an instance

This website is an open service developed by the [CloudBioLinux](#) and [CloudMan](#) communities. The goal is to make it easy to get started doing scalable biological analysis on cloud resources. See [this guide](#) for a detailed usage example when using the Amazon cloud. The [open source code](#) is available on GitHub allowing you to also run this service locally.

This site can be used for any of the available clouds. Note that you must have appropriate credentials for the chosen cloud. If a desired cloud is not available and you would like to see it there, please [contact us](#).

Launching servers on the Amazon cloud will incur [usage fees](#) from Amazon for their resources. By using this service you acknowledge your sole responsibility for any costs accrued.

Manage Your Cluster

CloudMan Console

Welcome to [CloudMan](#). This application allows you to manage this instance cloud cluster and the services provided within. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to manage services provided by the application.

Terminate cluster

Add nodes ▼

Remove nodes ▼

Access Galaxy

Status

Cluster name: ghem

Disk status: 0 / 0 (0%)

Worker status: Idle: 4 Available: 2 Requested: 5

Service status: Applications  Data 



Autoscaling is **off**.
Turn on?

Cluster status log



Tools

[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](#)[Motif Tools](#)[Multiple Alignments](#)[Metagenomic analyses](#)[FASTA manipulation](#)[NCBI BLAST+](#)[NGS: QC and manipulation](#)[NGS: Picard \(beta\)](#)[NGS: Mapping](#)[NGS: Indel Analysis](#)[NGS: RNA Analysis](#)[NGS: SAM Tools](#)[NGS: GATK Tools \(beta\)](#)

Welcome to Galaxy on the Cloud

managed by CloudMan

History



0 bytes

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

Galaxy – Ready for Use

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Test table browser
- UCSC Archaea table browser
- BX main browser
- EBI SRA ENA SRA
- Get Microbial Data

NGS: Mapping

- Lastz map short reads against reference sequence
- Lastz paired reads map short paired reads against reference sequence

NGS: RNA Analysis

RNA-SEQ

- Tophat for Illumina Find splice junctions using RNA-seq data
- Tophat for SOLiD Find splice junctions using RNA-seq data

NGS: GATK Tools (beta)

ALIGNMENT UTILITIES

- Depth of Coverage on BAM files
- REALIGNMENT
- Realigner Target Creator for use in local realignment
- Indel Realigner – perform local

Tophat for Illumina (version 1.5.0)

RNA-Seq FASTQ file:

6: C2_R3_1.chr4.fq

Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Will you select a reference genome from your history or use a built-in index?:

Use a built-in index

Built-ins were indexed using default options

Select a reference genome:

Arabidopsis thaliana (TAIR9)

If your genome of interest is not listed, contact the Galaxy team

Is this library mate-paired?:

Single-end

Tophat settings to use:

Use Defaults

You can use the default settings or set custom values for

Execute

Tophat Overview

Tophat is a fast splice junction mapper for RNA-Seq read aligner Bowtie, and then analyzes the mapping results. S.L. TopHat: discovering splice junctions with RNA-Seq.

Know what you are doing

There is no such thing (yet) as an automated gearshift running this tool with default parameters will probably require carefully reading the documentation and experiment

Input formats

Tophat accepts files in Sanger FASTQ format. Use the FASTQ format

Outputs

Tophat produces two output files:

junctions -- A UCSC BED track of junctions reported by

History

Options

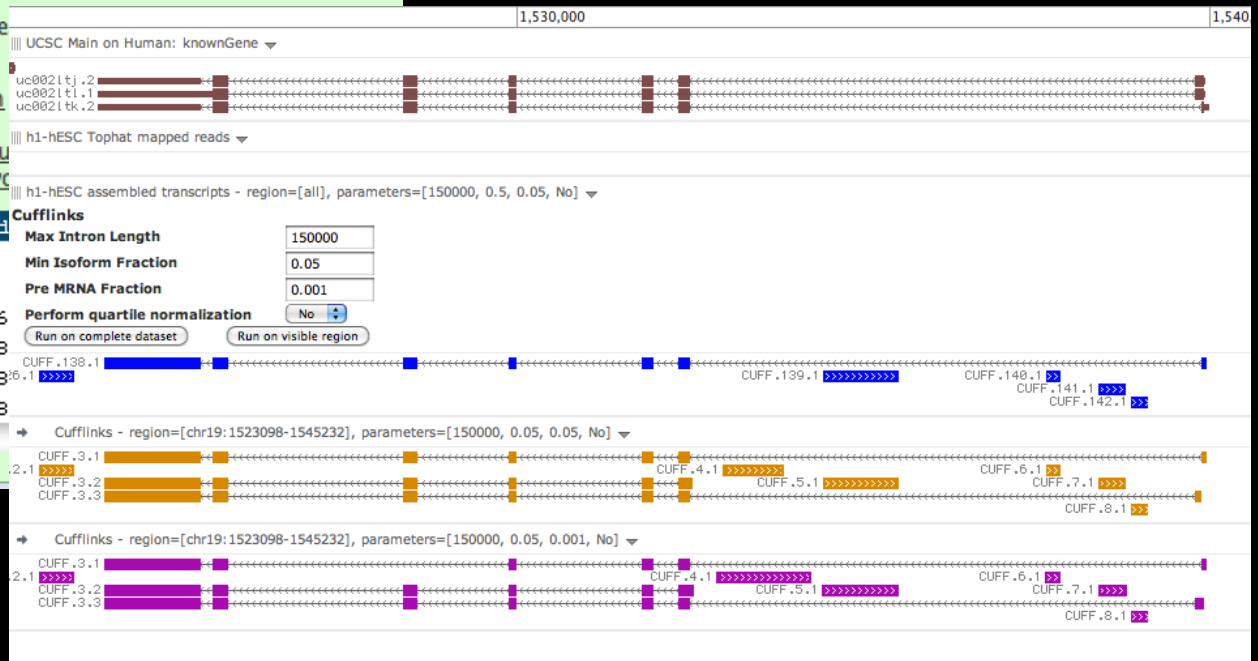
Worm genes

2: UCSC Main on C. elegans: sangerGene (chrII:1-15279323)

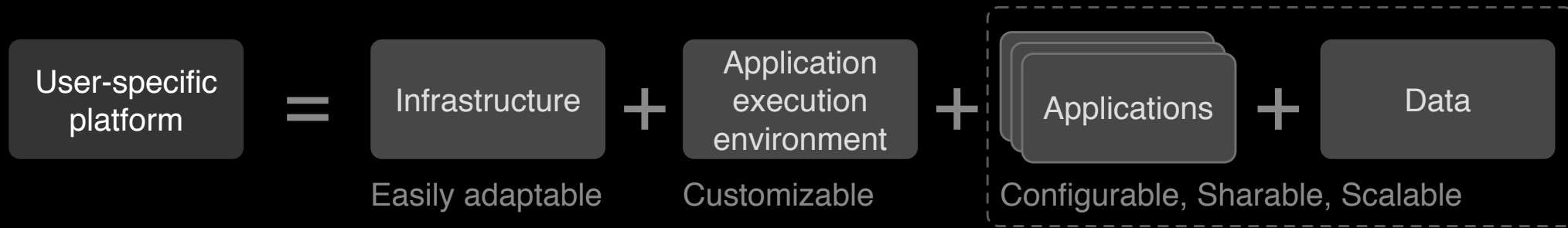
5,224 regions
format: bed, database

display at UCSC main
view in GeneTrack
display at Ensembl
display at GBrowse

| 1.Chrom | 2.Start | 3.End |
|---------|---------|-------|
| chrII | 1866 | 4663 |
| chrII | 6663 | 9233 |
| chrII | 9807 | 11826 |
| chrII | 12984 | 15998 |
| chrII | 19537 | 22158 |
| chrII | 23328 | 24428 |



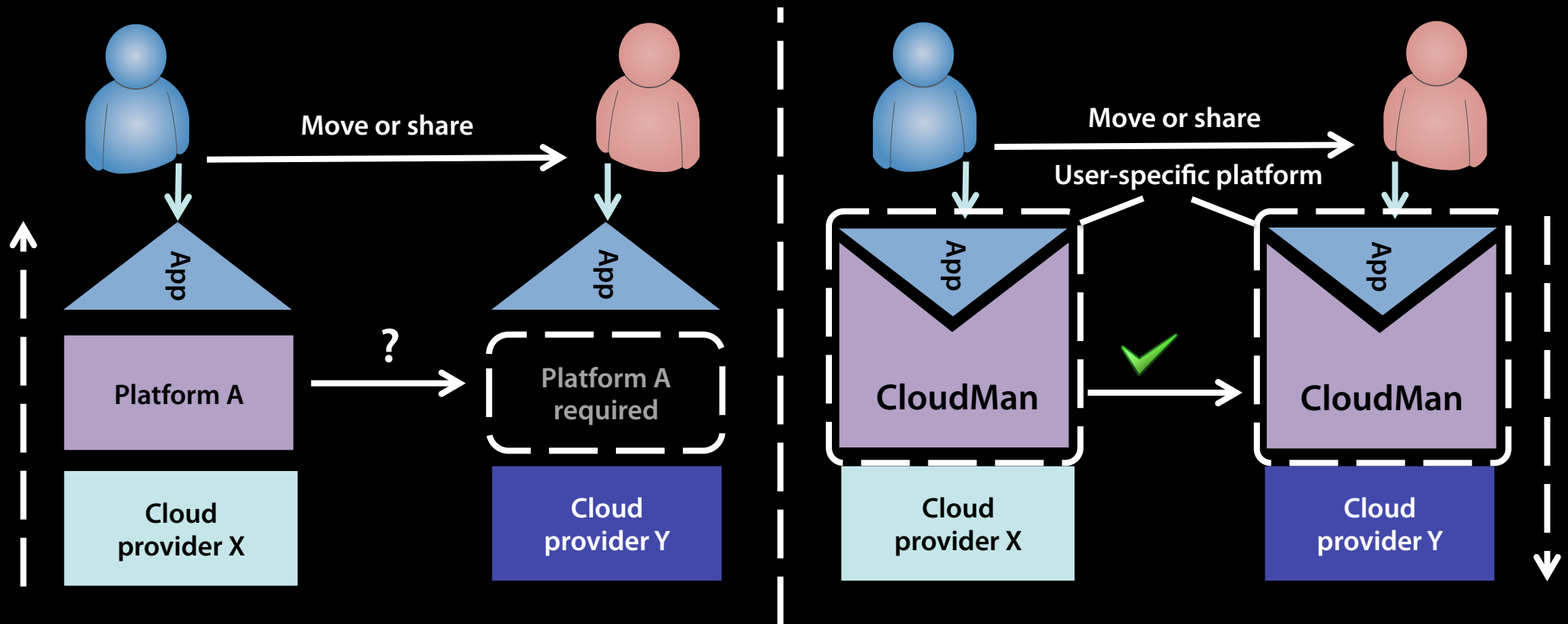
CloudMan-as-a-Platform



Enable easy creation of **user-specific cloud platforms**

Couple the infrastructure, complex and functional application execution environments, applications, and data into a single unit that can easily be used and manipulated by a user.

Packaged Platform Enables Reproducibility



CloudMan Platform Features

- A complete solution for instantiating, running and scaling cloud resources
 - Get a scalable **compute cluster** (SGE)
- Get an automatically **configured Galaxy application**
 - Scope of tools and reference datasets exceed Galaxy Main
- Deployment on **AWS, OpenStack, and OpenNebula** clouds
 - **Wizard-guided setup**: requires no computational expertise, no infrastructure, no software
- **Automated** configuration for machine image, tools, and data
 - Replicate EXACT environment anywhere (cloud, local, VM) & quickly
- **Self-contained** deployment
- **Elastic resource scaling**: manual or automatic
 - On AWS, support for **Spot** instances
- **Dynamic persistent storage**
- **Use any S3 bucket** as a local file system
- **Share** your instance: including all customizations (data, tools & configurations)
- Deploy a (Galaxy) cluster in minutes!

Value Added Features

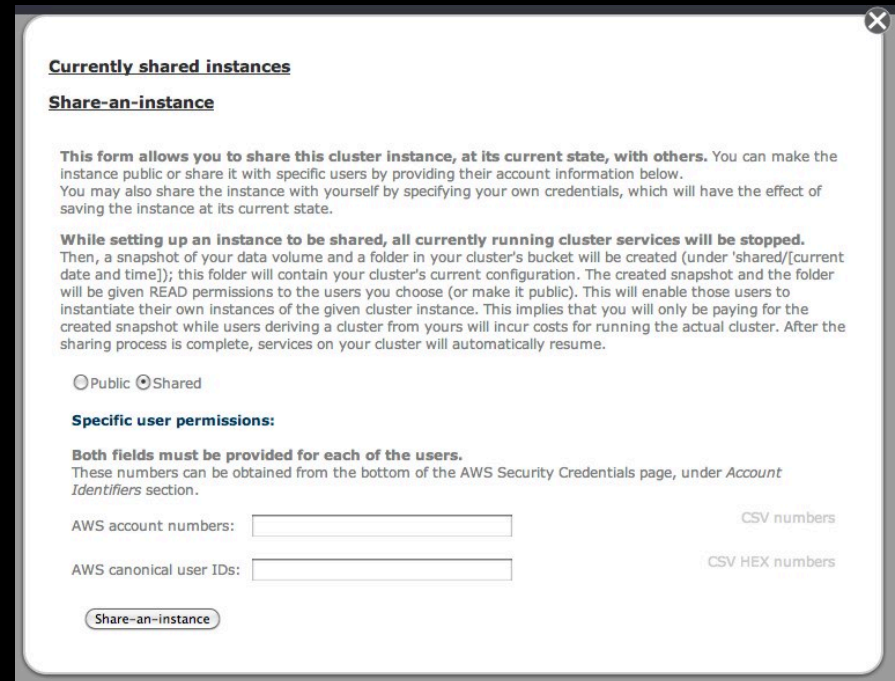
Customizing, Sharing, Scaling

Customize Your Instance

- Each CloudMan **instance is self-contained**, meaning that it can be built upon
- Deploy a tool and make it available
 - With all the configurations and sample data
- Upload data and share it with others
- Snapshot your own instance to capture:
 - Data
 - Configuration

Share Your Instance

- Share entire (Galaxy) CloudMan platform
 - Even the customized ones (including data and/or tools)
 - Fully automated solution
 - Publish a self-contained analysis
 - In progress or otherwise



The screenshot shows a web form titled "Share-an-instance" with a close button (X) in the top right corner. The form is divided into several sections:

- Currently shared instances**: A header section.
- Share-an-instance**: A sub-header section.
- Form text**: A paragraph explaining that the form allows sharing the cluster instance at its current state, with options to make it public or share with specific users. It also mentions that sharing with oneself by specifying credentials will save the instance at its current state.
- Warning**: A bolded section stating that while setting up an instance to be shared, all currently running cluster services will be stopped. It explains that a snapshot of the data volume and a folder in the cluster's bucket will be created, containing the current configuration. The snapshot and folder will be given READ permissions to the users chosen (or made public). This will enable those users to instantiate their own instances of the given cluster instance, implying they will only be paying for the created snapshot while users deriving a cluster from yours will incur costs for running the actual cluster. After the sharing process is complete, services on your cluster will automatically resume.
- Options**: Two radio buttons labeled "Public" and "Shared", with "Shared" selected.
- Specific user permissions:**: A section header.
- Instructions**: A paragraph stating that both fields must be provided for each of the users, and that these numbers can be obtained from the bottom of the AWS Security Credentials page, under the "Account Identifiers" section.
- Input fields**: Two text input fields. The first is labeled "AWS account numbers:" and the second is labeled "AWS canonical user IDs:". To the right of the first field is the text "CSV numbers" and to the right of the second field is "CSV HEX numbers".
- Submit button**: A button labeled "Share-an-instance" at the bottom.

| Name | Instance ID | References |
|---------------------------|--|--------------------------------------|
| Exome sequencing pipeline | cm-b53c6f1223f966914df347687f6fc818/shared/2011-10-07--14-00 | Pipeline description |

Scaling the Infrastructure with the Computation

The image shows a browser window at `ec2-50-17-119-106.compute-1.amazonaws.com/cloud` displaying the Galaxy Cloudman interface. The browser's address bar and the page's header both show the URL and the Galaxy Cloudman logo. The header also includes links for [Info](#), [report bugs](#), [wiki](#), and [screencast](#).

The main content area is titled "Galaxy Cloudman Console" and contains a welcome message: "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."

Below the welcome message are four buttons: "Terminate cluster", "Add nodes", "Remove nodes", and "Access Galaxy".

The "Status" section displays the following information:

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

An "Autoscaling Configuration" dialog box is open, showing the following text:

Autoscaling Configuration

Autoscaling attempts to automate the elasticity offered by cloud computing for this particular cluster. **Once turned on, autoscaling takes over the control over the size of your cluster.**

Autoscaling is simple, just specify the cluster size limits you want to work within and use your cluster as you normally do. The cluster will not automatically shrink to less than the minimum number of worker nodes you specify and it will never grow larger than the maximum number of worker nodes you specify.

While respecting the set limits, if there are more jobs than the cluster can comfortably process at a given time autoscaling will automatically add compute nodes; if there are cluster nodes sitting idle at the end of an hour autoscaling will terminate those nodes reducing the size of the cluster and your cost.

Once turned on, the cluster size limits respected by autoscaling can be adjusted or autoscaling can be turned off.

At the bottom of the dialog, there is a grid of 15 squares (3 rows by 5 columns). The top-left square is green, and the rest are grey. To the right of the grid, a box indicates "Autoscaling is on. Turn off?" and lists "Min nodes: 0" and "Max nodes: 15" with a link to "Adjust limits?".

At the bottom of the main console, there is a "Cluster status log" button with a plus icon.

Exercising Elasticity with AutoScaling

Fixed cluster size

5 nodes

Computation time: 9 hrs

Computation cost: \$20

20 nodes

Computation time: 6 hrs

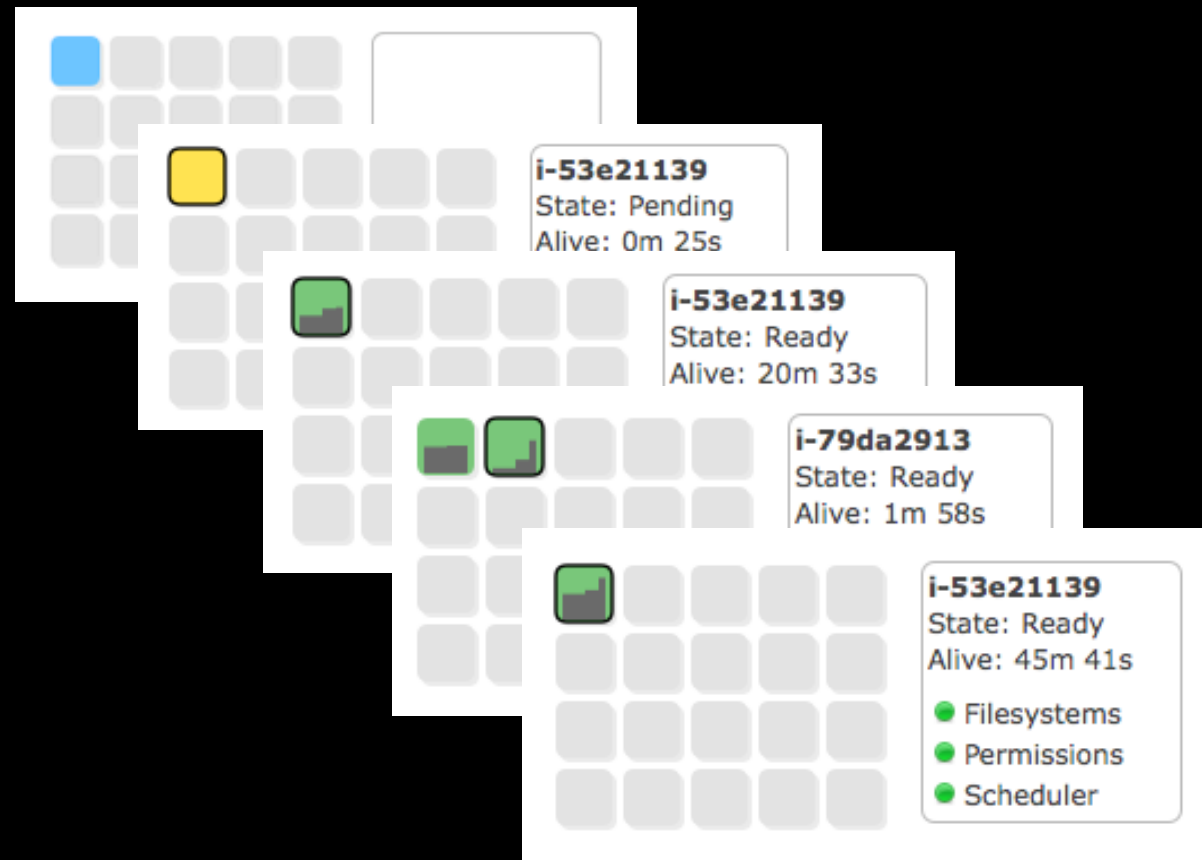
Computation cost: \$50

Dynamic cluster size

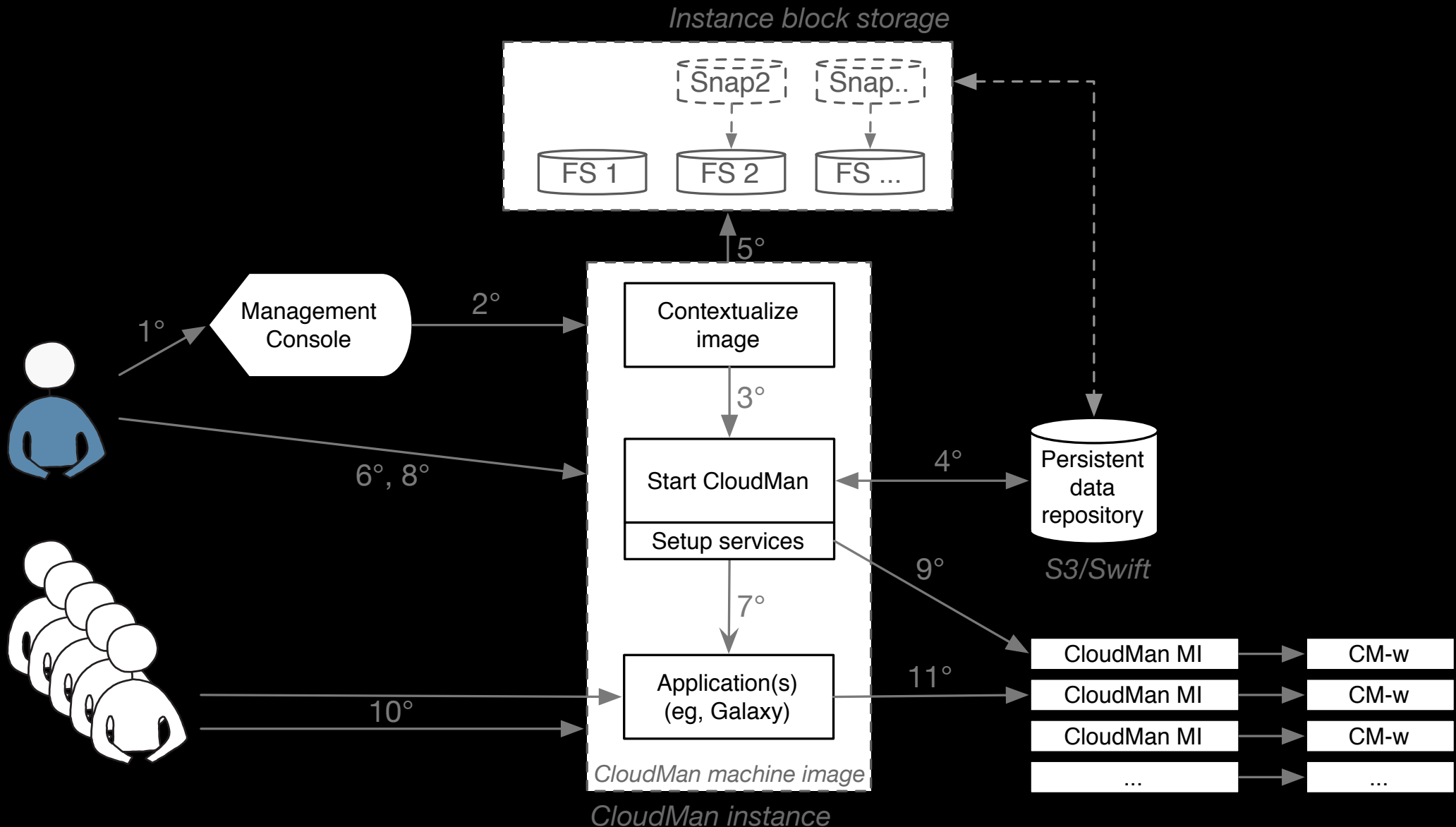
1 to 16
nodes

Computation time: 6 hrs

Computation cost: \$20



Flexible Architecture



What the Future Holds?

Standards, APIs, Apps

CloudMan and interoperability

Interoperable w/
multiple cloud
platforms

Use open
standards (OCCI,
CDMI)

Become the
default cloud
manager for
research clouds

CloudMan as a platform

Enable easy use of
value added
services via
automation

Advanced
autoscaling, data
management

Deployment
customizations

(REST) API

CloudMan and applications

A range of
application
execution
environments

Embedded
integration for
support for a range
of applications

Reproducible,
customizable,
sharable cloud
environments

Significance of CloudMan

An **accessible and reproducible cloud environment** that enables **decentralization** of services and realizes a **scalable model**, thus supporting some of the core pillars of science.

usecloudman.org