

Horizontal scaling with Galaxy

Enis Afgan

Galaxy Team

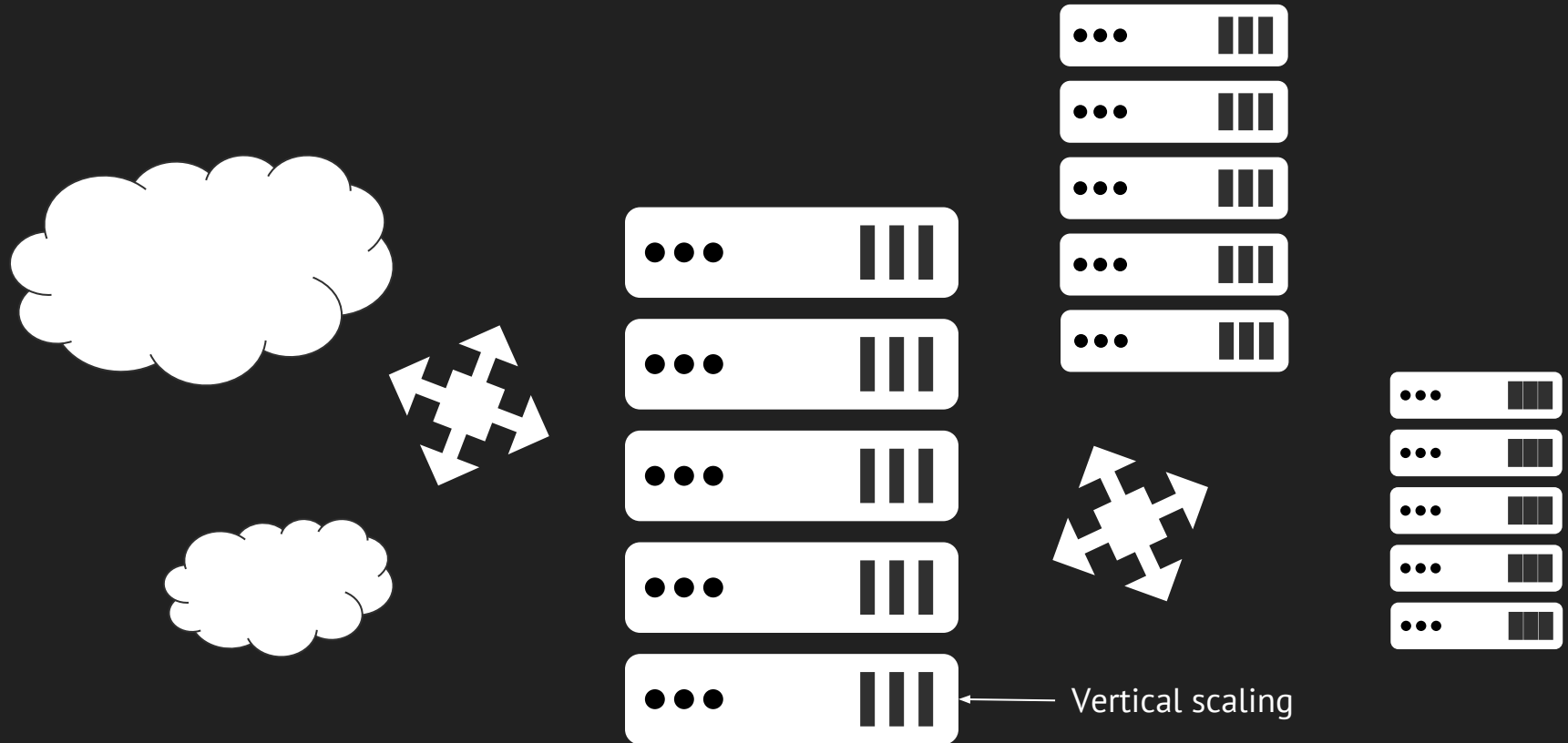
Johns Hopkins University

Galaxy Africa

Cape Town

April 5, 2018.

Horizontal scaling: service growth



Horizontal scaling #2: service replication



Scaling Galaxy: overview



Public
servers



Cloud
clusters



Local
installations

Usability

Flexibility

Want a local Galaxy?

```
$ git clone -b release_18.01 https://github.com/galaxyproject/galaxy.git
$ sh run.sh
...
http://localhost:8080
```

getgalaxy.org

Galaxy beyond the development server

Galaxy needs a complex ecosystem of software to operate effectively:



Robust
database



Job
manager



FTP
server



Shared data



Container
manager

Automating installations

Automate the process of building each component

Codify knowledge about the system → easier to build, easier to reproduce

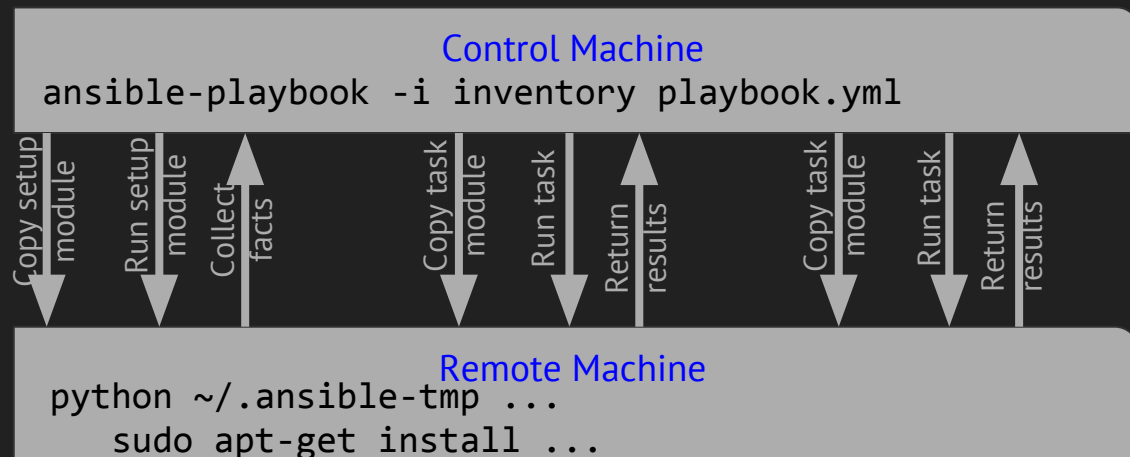
We use Ansible as the technology of choice

Roles:

- [galaxy-os](#)
- [nginx](#)
- [postgresql](#)
- [postgresql_objects](#)
- [galaxy](#)
- [interactive-environments](#)
- [trackster](#)
- [pulsar](#)
- [galaxy-tools](#)
- [galaxy-extras](#)

Playbooks:

- [usegalaxy-playbook](#)
- [infrastructure-playbook](#)
- [galaxy-cloudman-playbook](#)
- [GalaxyKickStart](#)



Closer look at the **Galaxy Main** installation

Galaxy Main requires scale

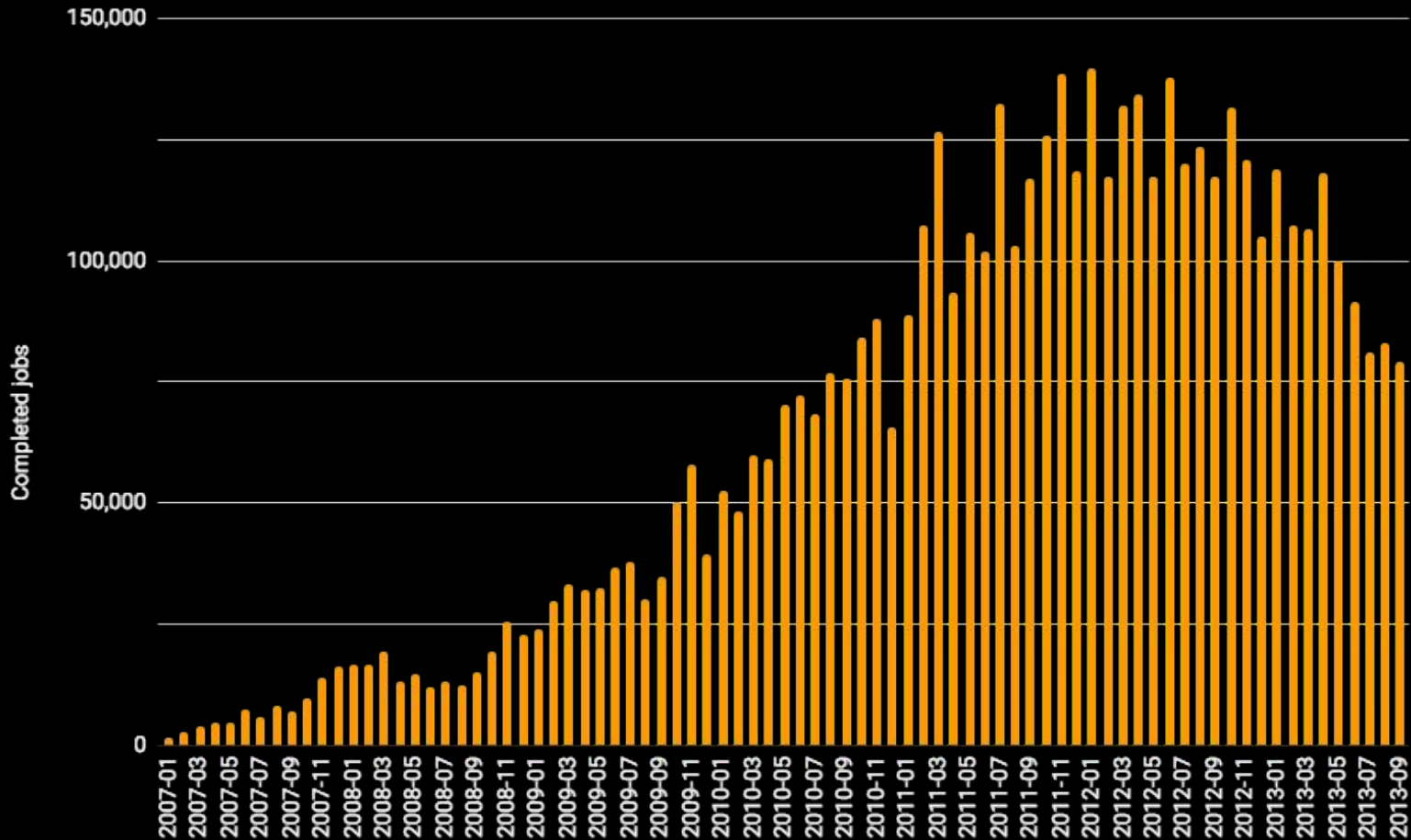
125,000
registered users

2PB
user data

18M
jobs run

100
training events
(2017 & 2018)

Running into scalability issues...



Decentralizing the installation

- Traditionally, Galaxy was designed for local installation and required a shared file system
- Implement a pluggable interface to compute resources to readily connect to external cluster(s)
- Leverage Pulsar for data staging and Galaxy resource/job manager

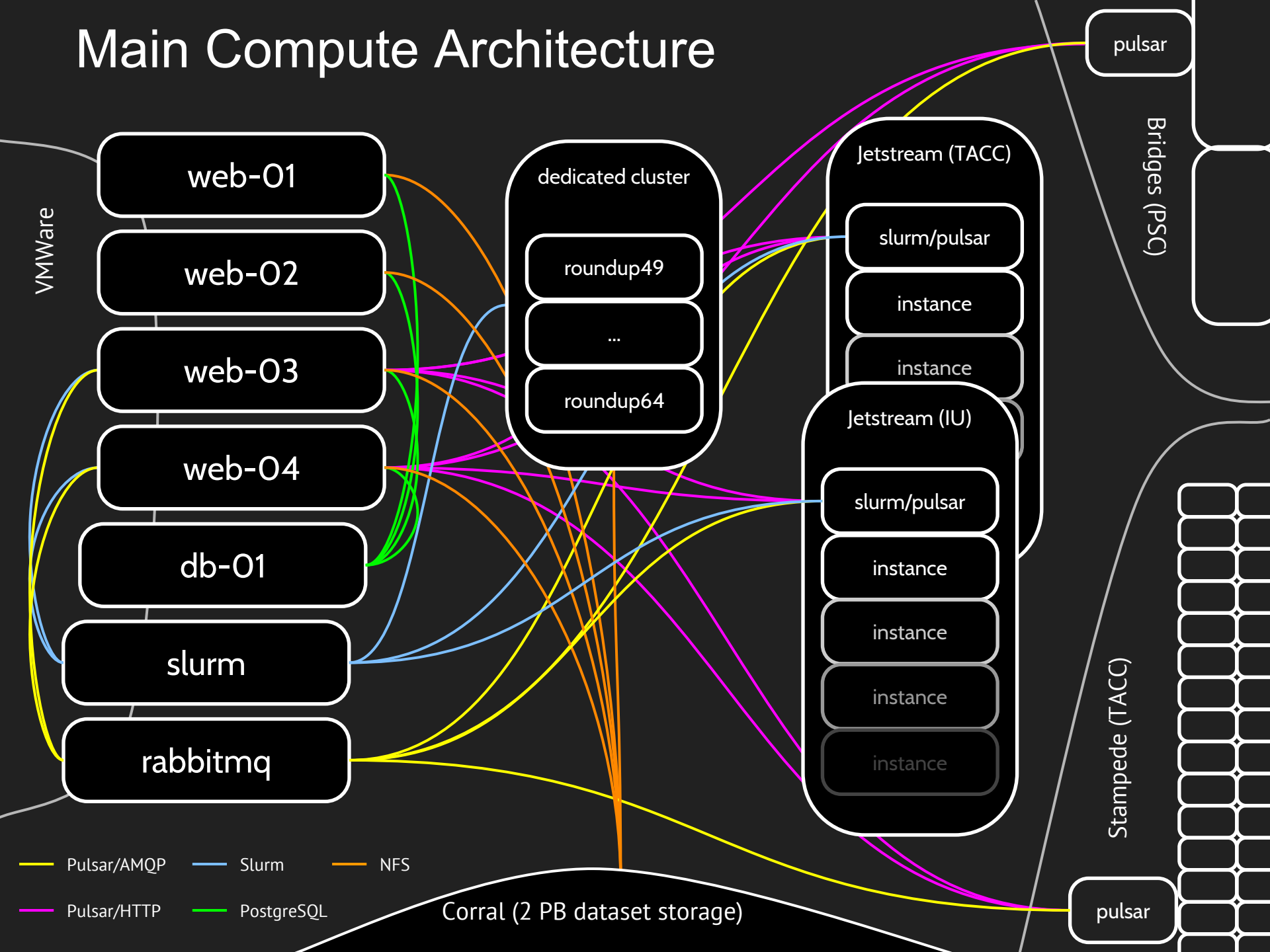
Expanding compute capacity

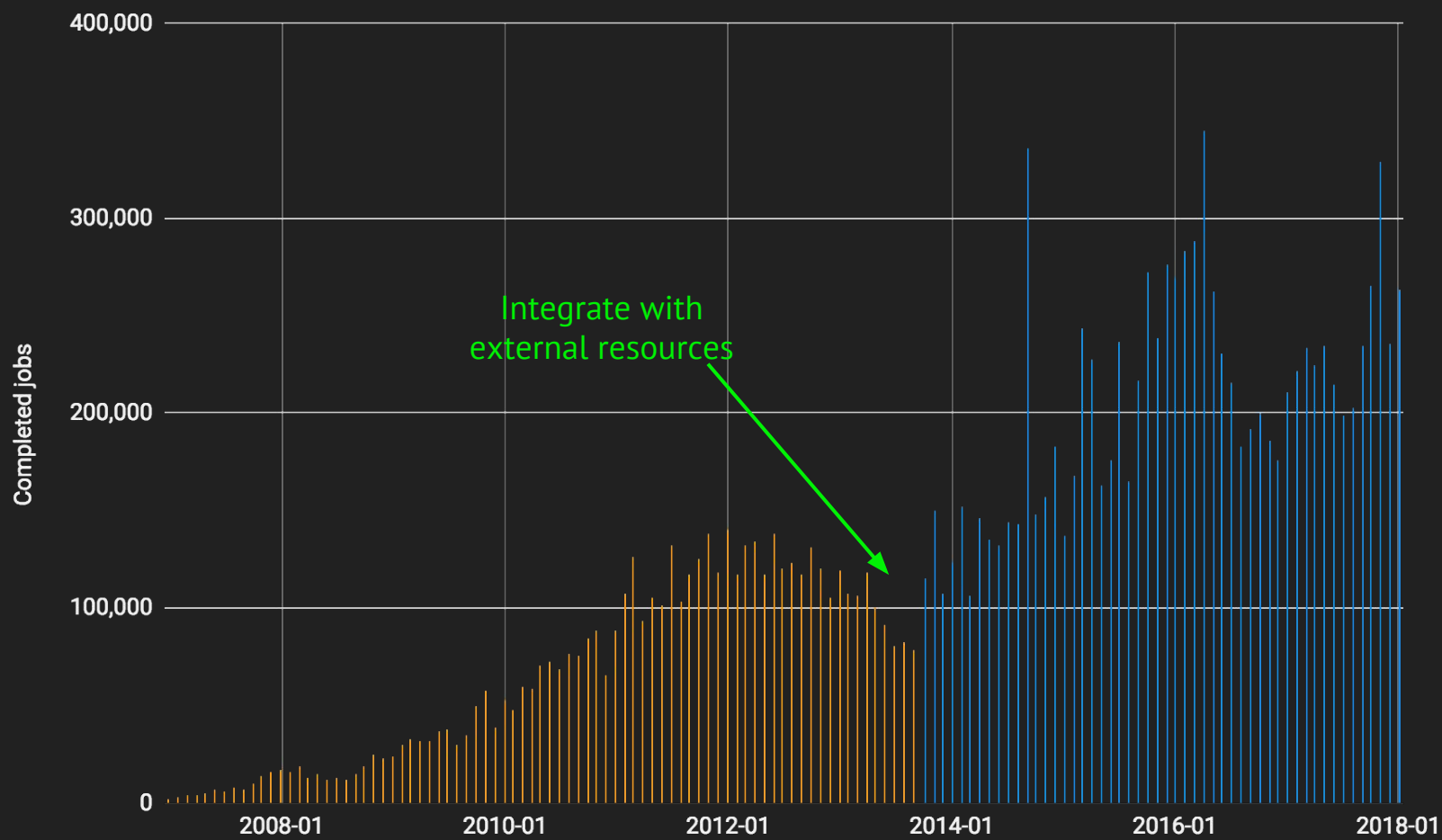
Leveraging National Cyberinfrastructure: Galaxy/XSEDE Gateway



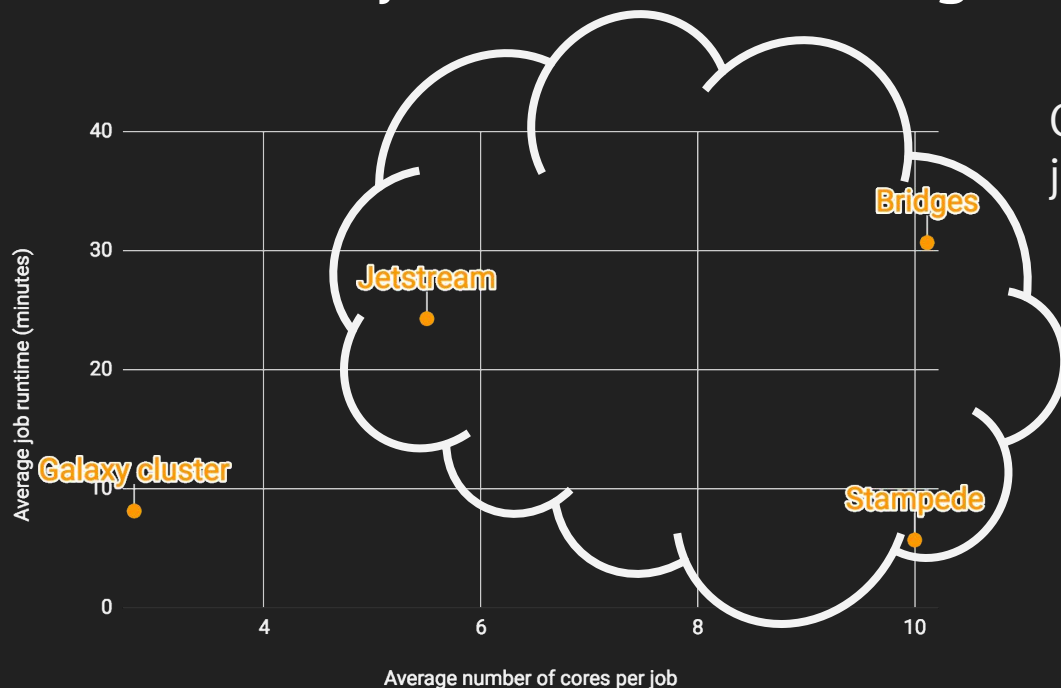
Shared XSEDE resources
Dedicated resources

Main Compute Architecture





More than job counts, scaling moved the horizon

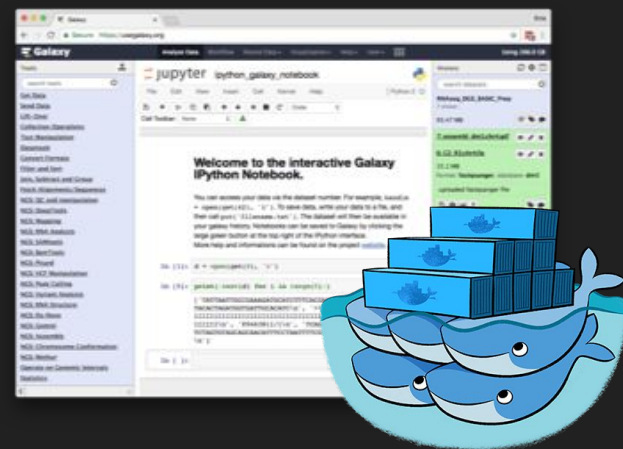


Can now **run larger jobs**, and more jobs:

On Jetstream, 325,000 jobs run on behalf of 12,000 users

Can run **new types of jobs**:

Galaxy Interactive
Environments: Jupyter, RStudio



Scaling Galaxy: overview



Public
servers



Cloud
clusters



Individual
installations

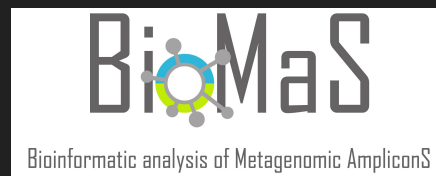
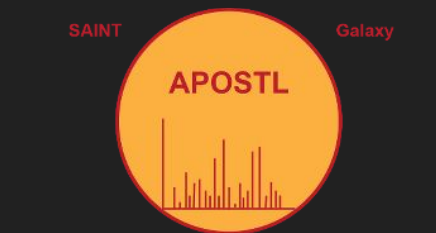
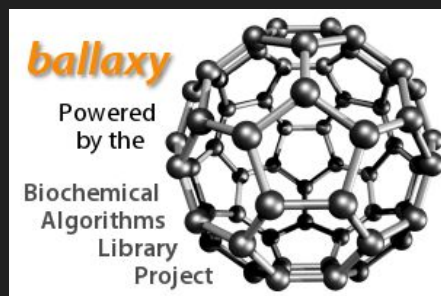
Usability

Flexibility

100+ public Galaxy Servers



bit.ly/gxyServers



Scaling with consistency

- Multiple Galaxy servers are wonderful for accessibility and versatility
- Globally, they do lead to a good bit of repeated effort → not very scalable
- Focus on **usegalaxy.* federation**
 - A set of coordinated Galaxy instances with a set of common core tools and reference genomes
 - .org / .eu / .org.au domains exist today
 - Leverage common reference data and, in future, tool and Galaxy binaries
 - Serve as a model for other local instances to reuse installation components

Galaxy federation components

Reference data

- Leverage CernVM file system (CVMFS) as a distributed, read-only file system
- A centrally updated and automatically replicating global set of servers
- Stratum-0: master copy
- Stratum-1: read-only replicas: 3 in the US, 1 in EU, 1 in AU
- Anyone can connect to these: bit.ly/gxyCVMFS

Tools

- Current list of tools (for usegalaxy.eu) is published at bit.ly/gxyEUtools
- Use Ephemeris command line tool to install locally
- Still work in progress; eventually will be able to use CVMFS directly

Scaling Galaxy: overview



Public
servers



Cloud
clusters



Individual
installations

Usability

Flexibility

Galaxy on the cloud

- Launch your own instance of Galaxy on the cloud
 - Within minutes, using a web browser, including the infrastructure and configurations with ability to scale
- Based on **CloudMan**: a cloud manager for deploying Galaxy on a variety of cloud providers



CLOUDMAN



Welcome to Galaxy Melbourne

The GVL paper has recently been published. [View the paper](#)

We provide the Galaxy instance to:

- provide bioinformatics infrastructure to the Melbourne life sciences community
- give researchers tools to do their own analysis
- help improve general bioinformatics skills
- provide a platform for tutorials and other training

IMPORTANT NOTE:

RNA-Seq data is available for download. If you require more data, please contact the GVL team.

GVL Dashboard Home Admin About

GVL 4.1.0

Welcome to the GVL Dashboard! The GVL Dashboard is a portal through which you can access all services on your GVL instance.

Instance Services

for abripi-mGVL

Service Name	Description	Status	Access Link
Galaxy	Galaxy is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research.	●	http://abripi.genome.edu.au/galaxy Username: manual sign up Password: c!outum password
Cloudman	CloudMan is a cloud manager that orchestrates the steps required to provision and manage compute clusters on cloud infrastructure. Use Cloudman to start and manage your Galaxy service and to add additional nodes to your compute cluster.	●	http://abripi.genome.edu.au/cloud Username: ubuntu Password: <cluster password>
Lubuntu Desktop	Lubuntu is a lightweight desktop environment through which you can run desktop applications on your virtual machine. You can also access the GVL commandline utilities through the desktop.	●	http://abripi.genome.edu.au/vnc Username: ubuntu Password: <cluster password>
SSH	You can login to your virtual machine remotely through an SSH client.	●	ssh://abripi1.genome.edu.au Username: ubuntu Password: <cluster password>
JupyterLab	JupyterLab can be used to access your personal (Python) Notebook. Python Notebook is a web-based interactive computational environment where you can combine code execution, text, mathematics, plots and rich media into a single document.	●	http://abripi.genome.edu.au/jupyter Username: researcher Password: <cluster password>
RStudio	RStudio IDE is a powerful and productive user interface for R.	●	http://abripi.genome.edu.au/rstudio Username: researcher Password: <cluster password>
Public HTML	This is a shared web-accessible folder. Any files you place in this directory will be publicly accessible.	●	http://abripi.genome.edu.au/public/researcher/ Username: researcher Password: <cluster password>
SMRT Portal	SMRT Portal is PacBio's open source software suite for single molecule, real-time sequencing.	●	http://abripi.genome.edu.au/smportal Username: manual sign up Password: c!outum password

Jupyter Untitled Last Check

```
In [31]: from math import log
from math import sqrt

In [32]: x=[1,2,3,4,5,6,7,8,9]

In [33]: y=[log10(i) for i in x]

Out[34]: [0.0,
0.3010299956639812,
0.47712125471966244,
0.6020599913279624,
0.6989700043360189,
0.7781512503836436,
0.8450980400142569,
0.9030899869919435,
0.9542425094393249,
1.0]
```

SMRT Portal Home Help About

DESIGN JOB **MONITOR JOBS** **VIEW DATA**

Open Existing Create New Import and Manage

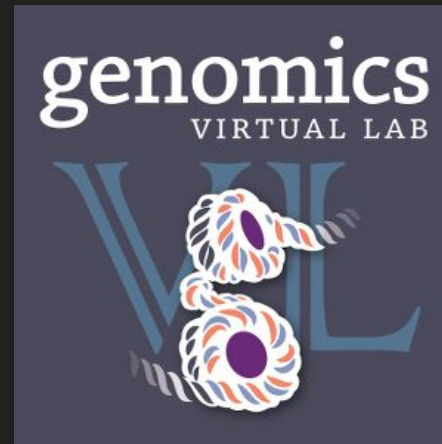
RECENT JOBS

Job Name	Protocol	Reference Sequence	Started	Status	User

Genomics Virtual Lab

GVL: a middleware layer of machine images, cloud management tools, and online services for cloud bioinformatics

➔ **A superset of Galaxy-on-the-cloud:** build arbitrarily sized compute clusters on demand, pre-populated with fully configured bioinformatics tools, reference datasets, workflows and visualisation options.

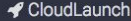


gvl.org.au



THE UNIVERSITY OF
MELBOURNE

Launch-your-own, via CloudLaunch

 CloudLaunch

Q Catalog

Public Appliances


My Appliances

User (enis)


Appliance Catalog

An online depot to discover and launch pre-configured software for a variety of clouds.


Search for an appliance




Genomics Virtual Lab (GVL)
A versatile genomics workbench with Galaxy, RStudio and Jupyter.
USE THIS FOR LATEST GALAXY.




Galaxy CloudMan
Pre-configured Galaxy instance on a scalable cluster-in-the-cloud.
DEPRECATED - USE THE GVL INSTEAD.




Ubuntu
Ubuntu operating system




CentOS
Stock CentOS



BioDocklet
Abstract the complex data operations of multi-step, bioinformatics pipelines for NGS data analysis.




Galaxy Standalone VM
A standalone Galaxy virtual machine, configured and ready for use.

This website allows you to create, monitor and access a range of virtual appliances: pre-configured application(s) that can be launched in just a few clicks. You can create new appliances or access public ones that others have made freely available. Watch the intro video below and click the  icon throughout the page to show detailed help.

Appliance Marketplace

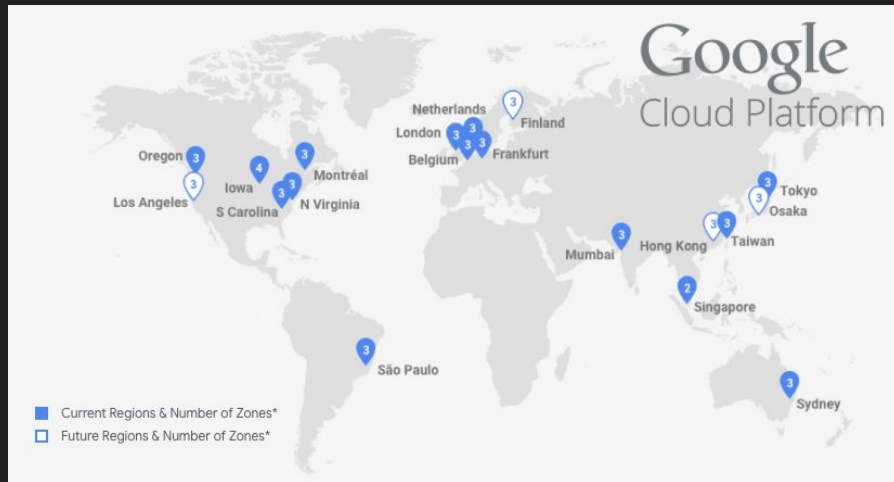
CloudLaunch introduction



Click on any appliance listed on the left to proceed, or access public appliances from the link at the top.

Demo using <https://launch.usegalaxy.org/>

Scaling across clouds



Towards Galaxy-on-the-cloud 2.0

Goals:

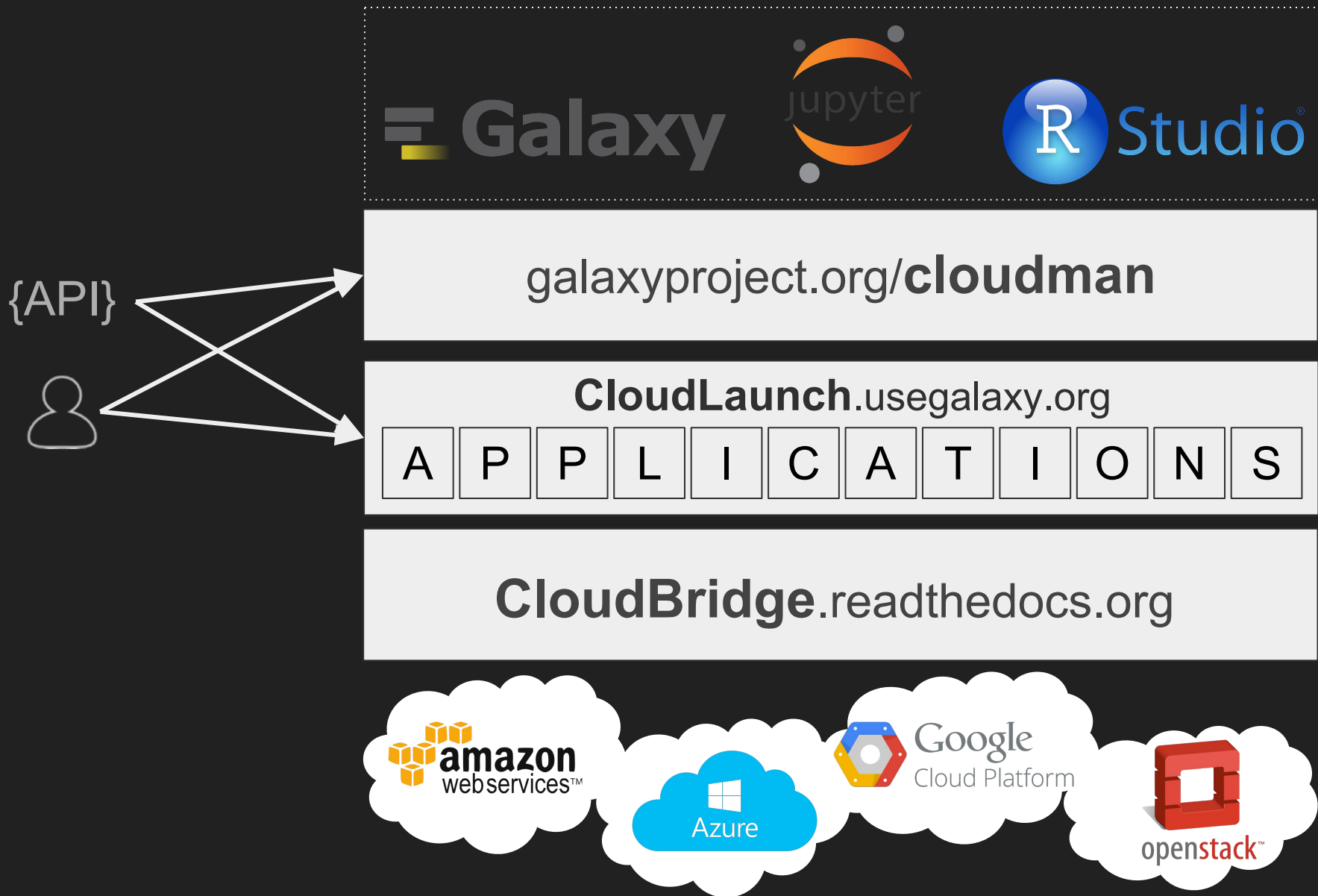
- Uniform availability without infrastructure-specific builds
- Well-defined upgrade path for users
- Focus on institutions vs. individual users

Approach:

- Abstract provider differences
- Containerize everything
- Leverage federated infrastructure components (e.g., CVMFS, containers)

How is the launch-your-own enabled?

Genomics Virtual Lab



CloudBridge

A Simple Cross-Cloud Python Library

Goonasekera, N., Lonie, A., Taylor, J., Afgan, E., “**CloudBridge – a Simple Cross-Cloud Python Library**”, *XSEDE 16*, Miami, FL, July 2016.

Multi-cloud computing with CloudBridge

CloudBridge: a simple, open-source Python multi-cloud library.

Uniform API irrespective of the underlying provider

No special casing of application code, unlike Apache Libcloud
Simpler code

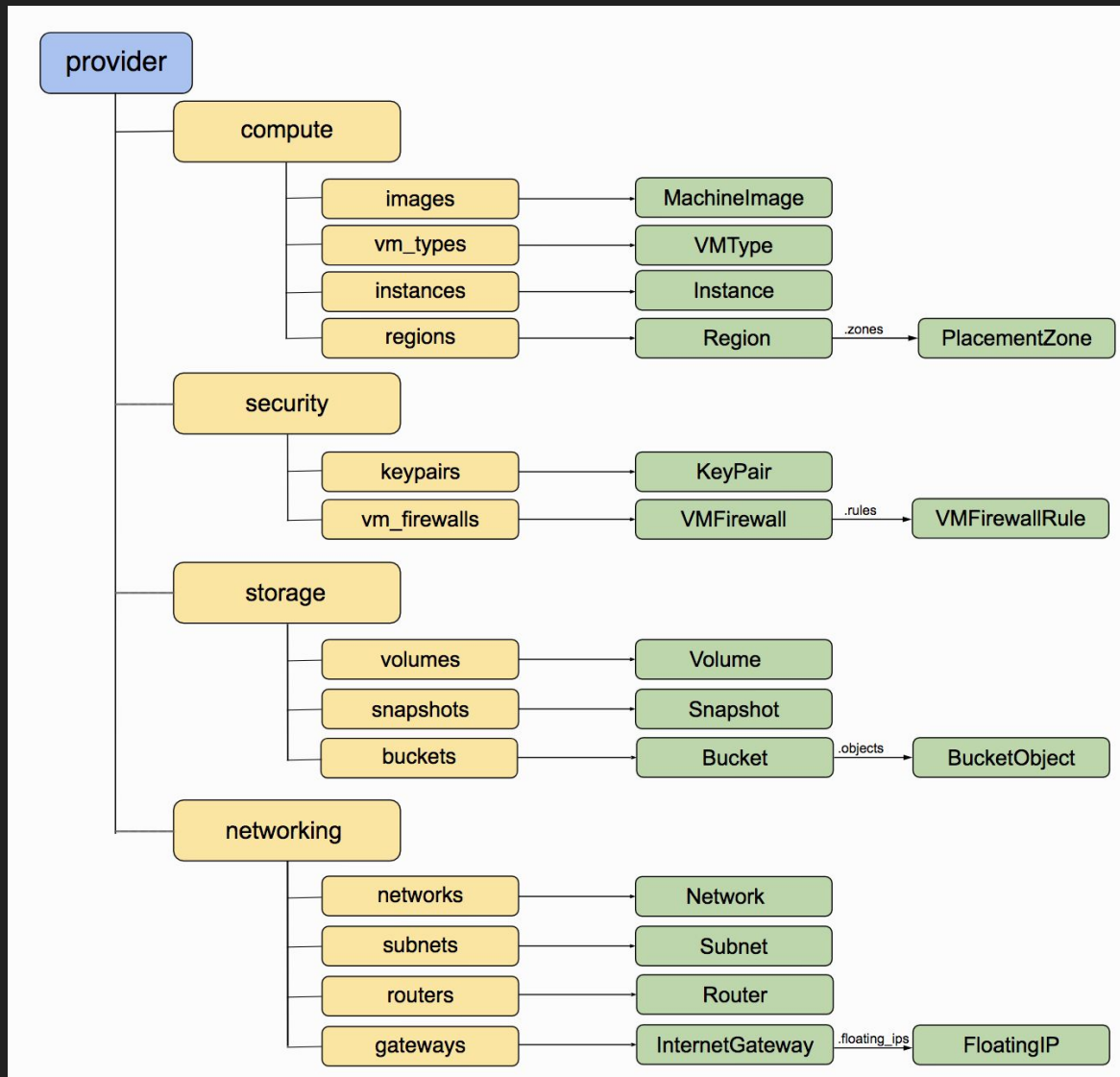
Provide a set of conformance tests for all supported clouds

No need to test against each cloud, unlike Terraform
“Write-once-run-anywhere”

Supports AWS, Azure, and OpenStack right now

GCE support is forthcoming

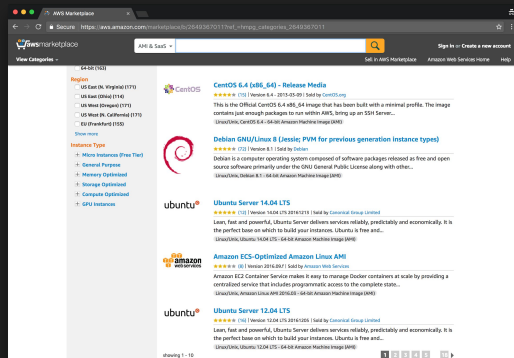
<http://cloudbridge.readthedocs.org>



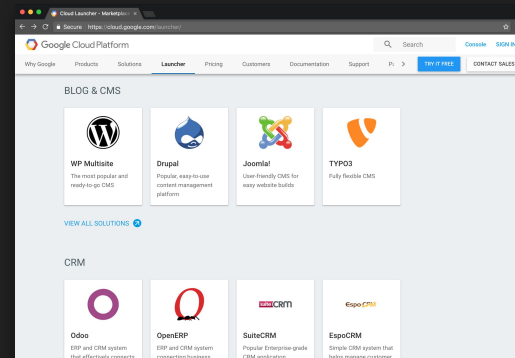
CloudLaunch

A gateway for discovering and launching applications on a variety of clouds.

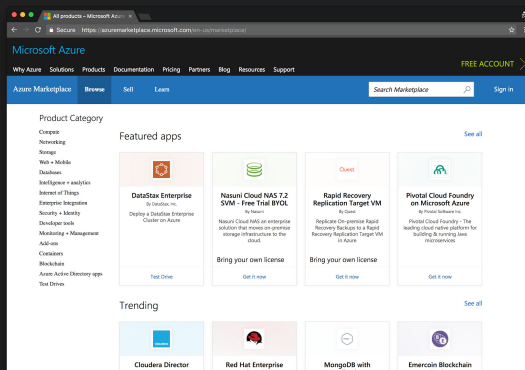
CloudLaunch-as-a-Service



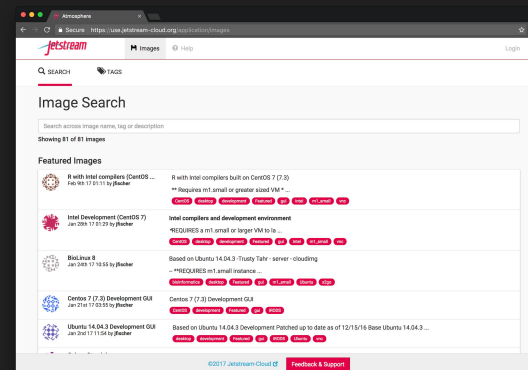
AWS Marketplace



GCE Solutions



Azure Marketplace



Jetstream Atmosphere VMs

CloudLaunch features

Cloud-agnostic

Backed by CloudBridge, use native cloud capabilities for infrastructure management

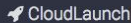
Pluggable and extensible

Arbitrary launch process and UI are supported, via an isolated plug-in mechanism

UI and REST API

UI available for end-users but it is all API driven for integration into external apps

CloudLaunch demo #2: multi-cloud, multi-app, API

 CloudLaunch

Q Catalog

Public Appliances


My Appliances

User (enis)


Appliance Catalog

An online depot to discover and launch pre-configured software for a variety of clouds.


Q Search for an appliance




Genomics Virtual Lab (GVL)
A versatile genomics workbench with Galaxy, RStudio and Jupyter.
USE THIS FOR LATEST GALAXY.




Galaxy CloudMan
Pre-configured Galaxy instance on a scalable cluster-in-the-cloud.
DEPRECATED - USE THE GVL INSTEAD.




Ubuntu
Ubuntu operating system




CentOS
Stock CentOS



BioDocklet
Abstract the complex data operations of multi-step, bioinformatics pipelines for NGS data analysis.




Galaxy Standalone VM
A standalone Galaxy virtual machine, configured and ready for use.

This website allows you to create, monitor and access a range of virtual appliances: pre-configured application(s) that can be launched in just a few clicks. You can create new appliances or access public ones that others have made freely available. Watch the intro video below and click the  icon throughout the page to show detailed help.

Appliance Marketplace

CloudLaunch introduction



Click on any appliance listed on the left to proceed, or access public appliances from the link at the top.

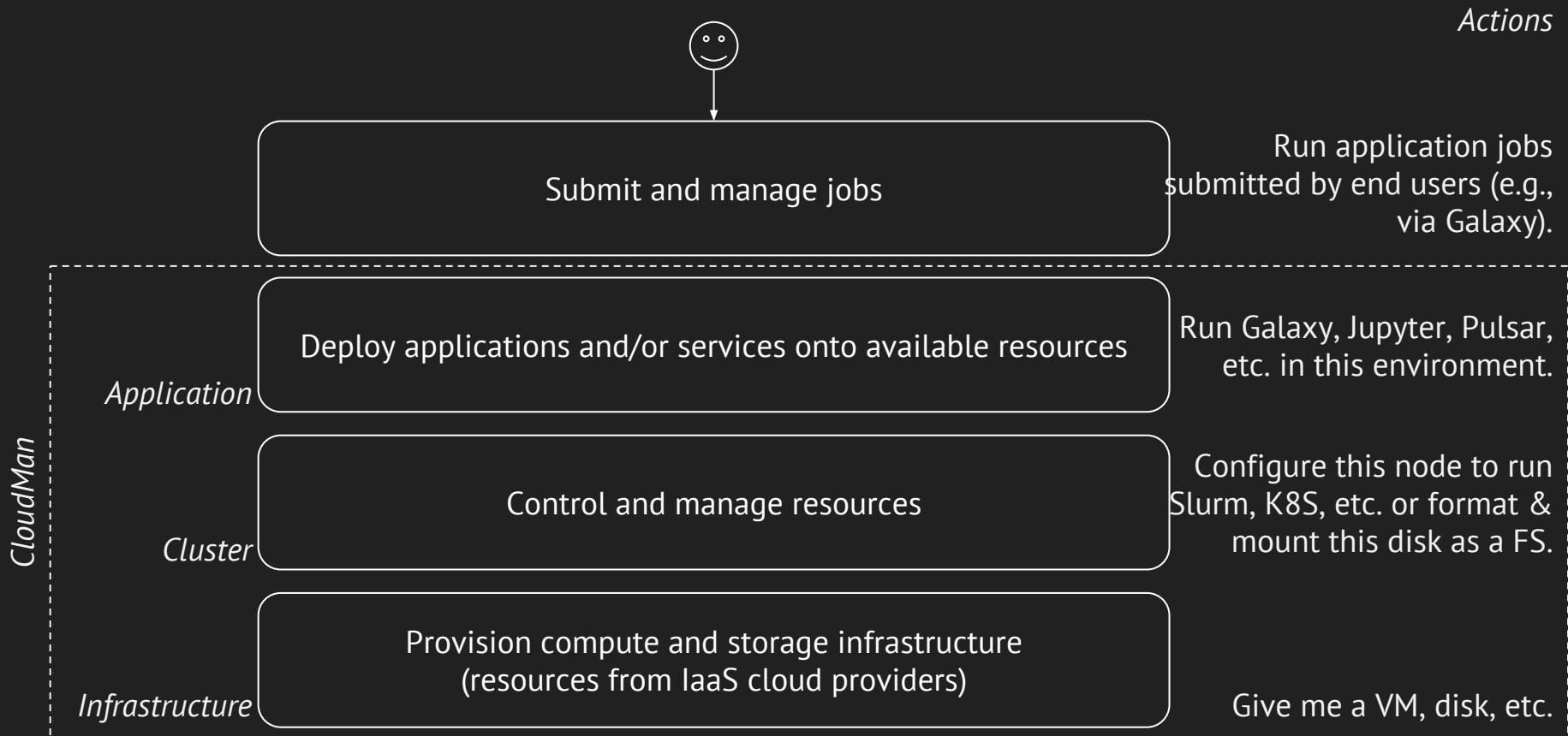
Demo using <https://launch.usegalaxy.org>

CloudMan and beyond

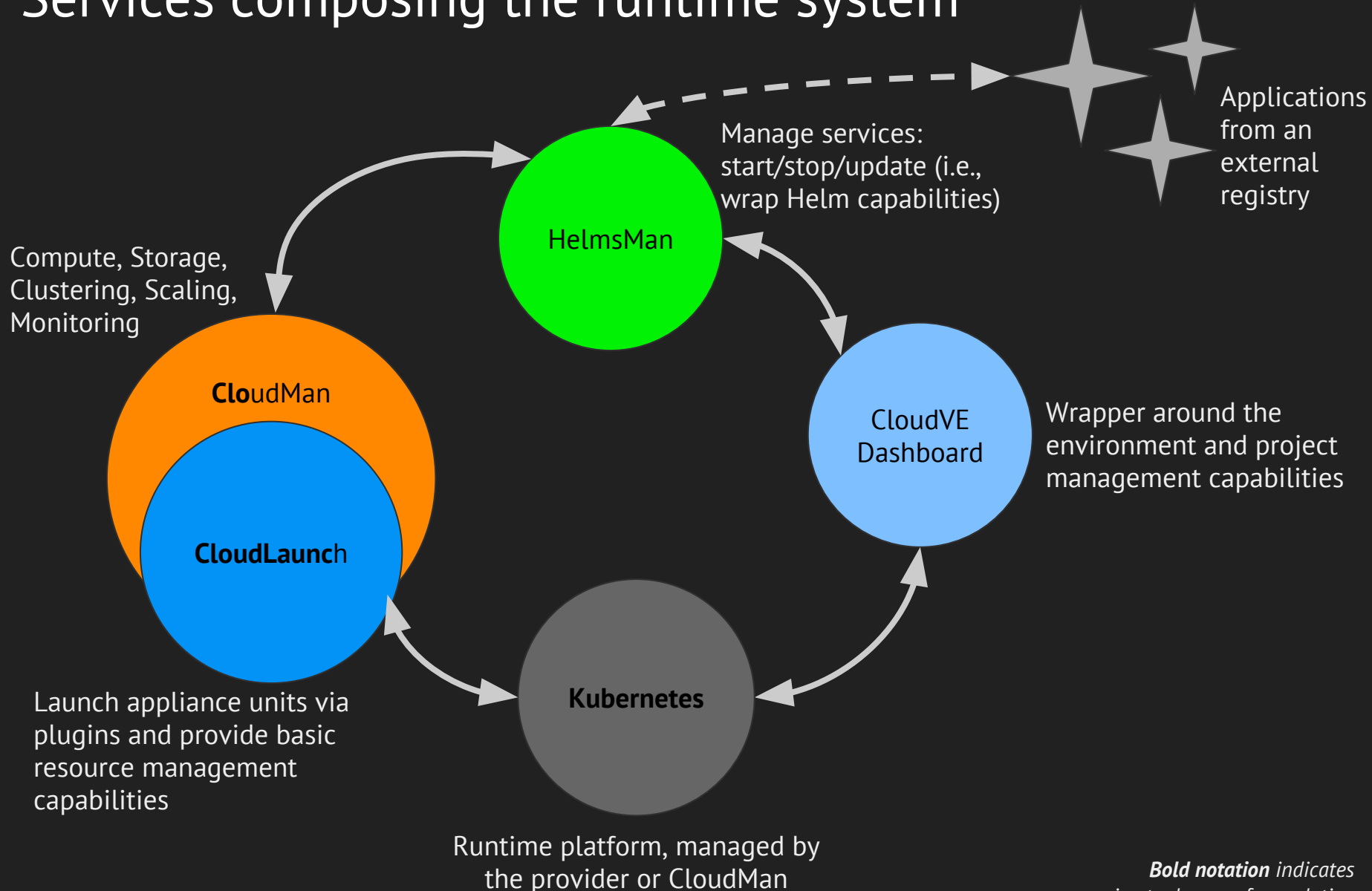
Manage deployed infrastructure and applications

Managing deployed infrastructure

- Once deployed, an application needs management, and so does the infrastructure

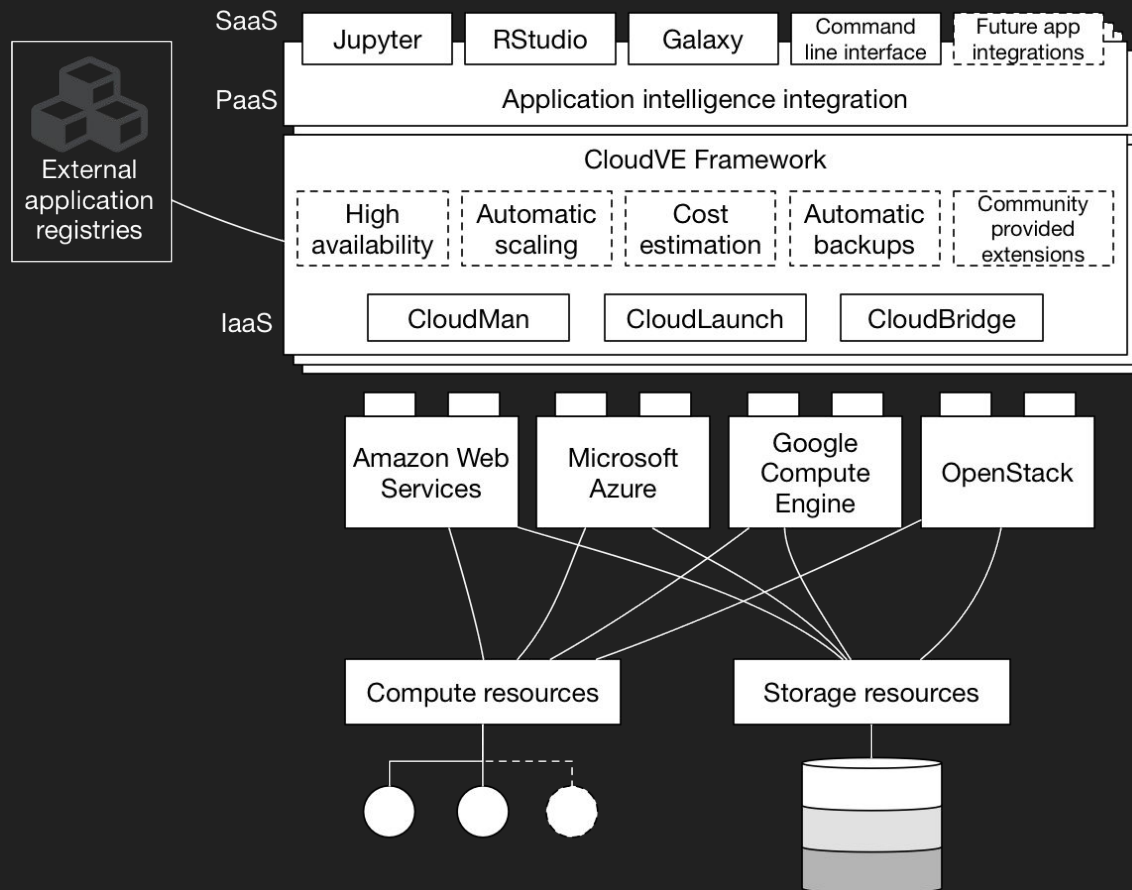


Services composing the runtime system

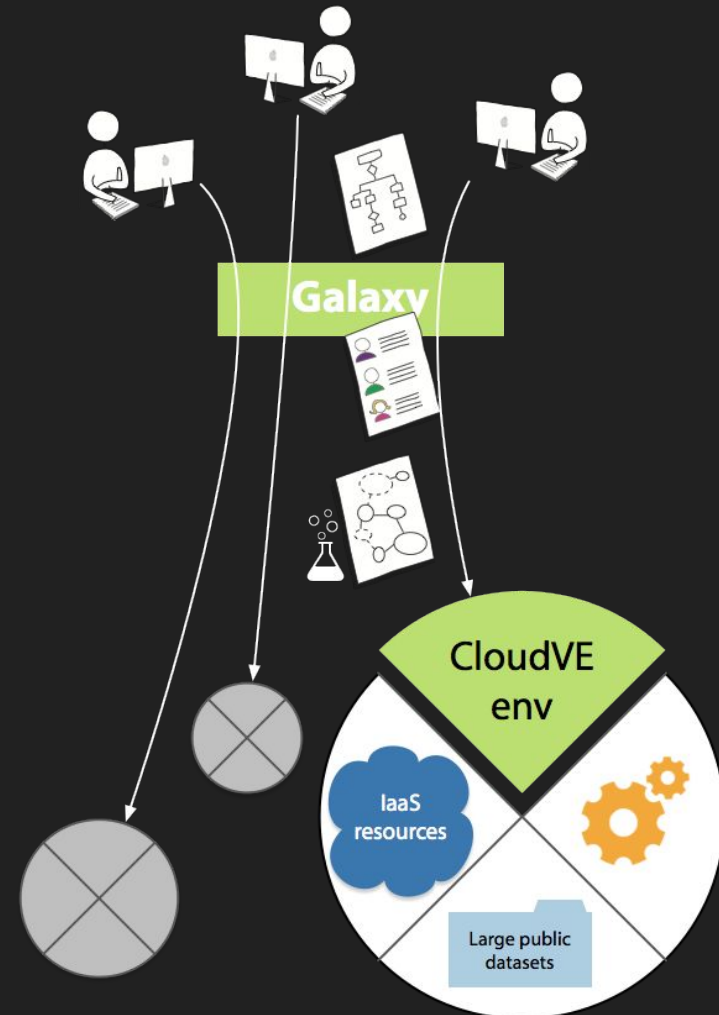


Looking forward: two models of usage

Virtual laboratory



Native application integration



Conclusions

We've seen three models of scaling Galaxy:



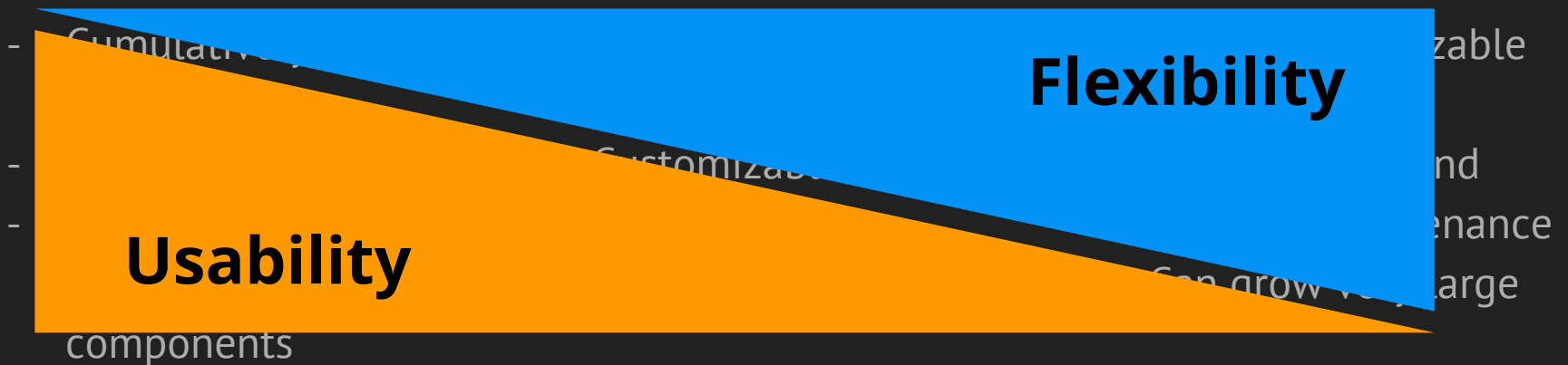
Public
servers



Cloud
clusters



Individual
installations



Acknowledgments

Institutions



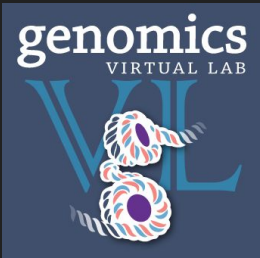
JOHNS HOPKINS
UNIVERSITY

PennState



THE UNIVERSITY OF
MELBOURNE

Projects



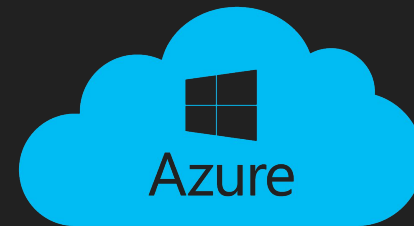
Galaxy | scc |

Science Gateways
Community Institute

Infrastructure



XSEDE
Jetstream



nectar
cloud