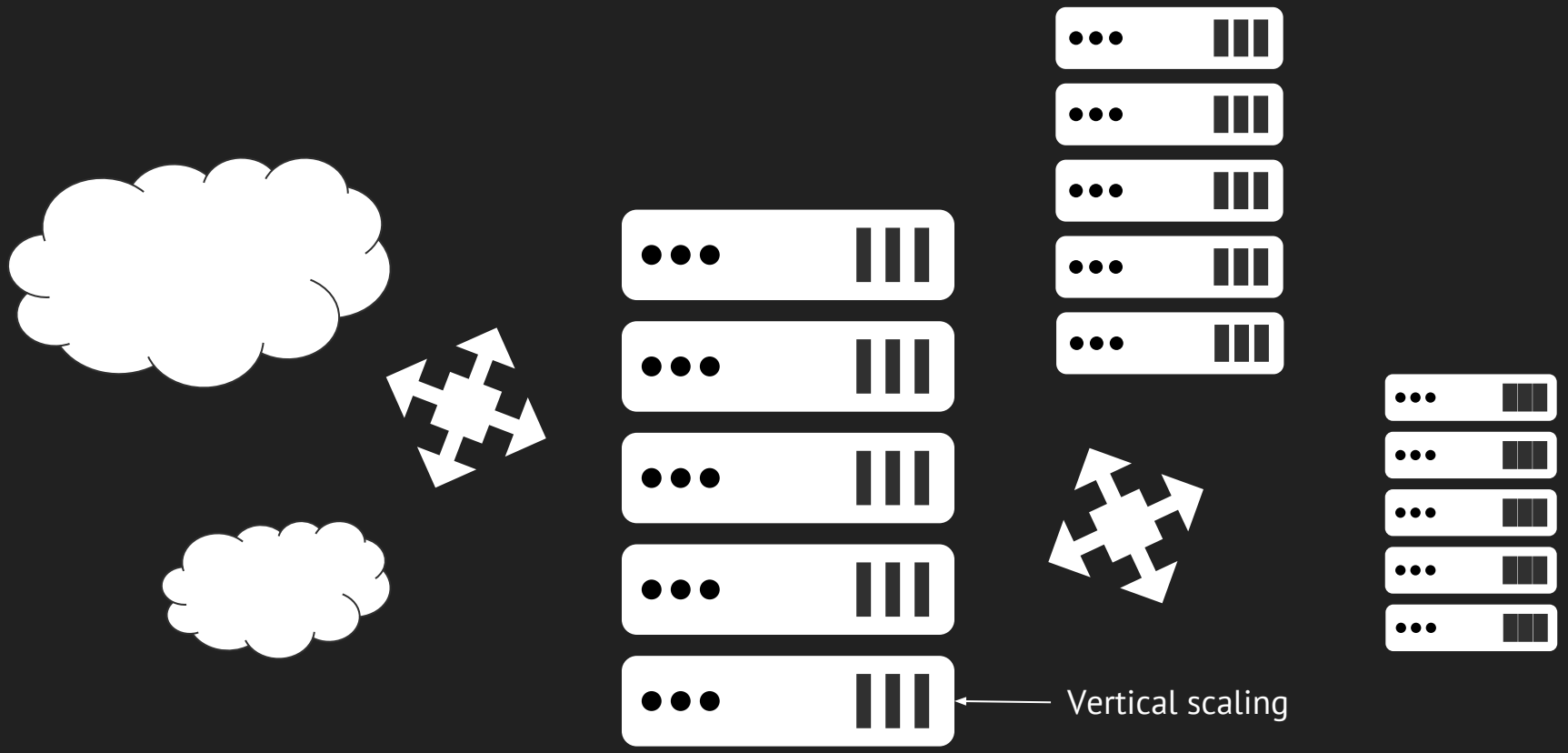
Horizontal scaling with Galaxy

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

Flexibility

Usability

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

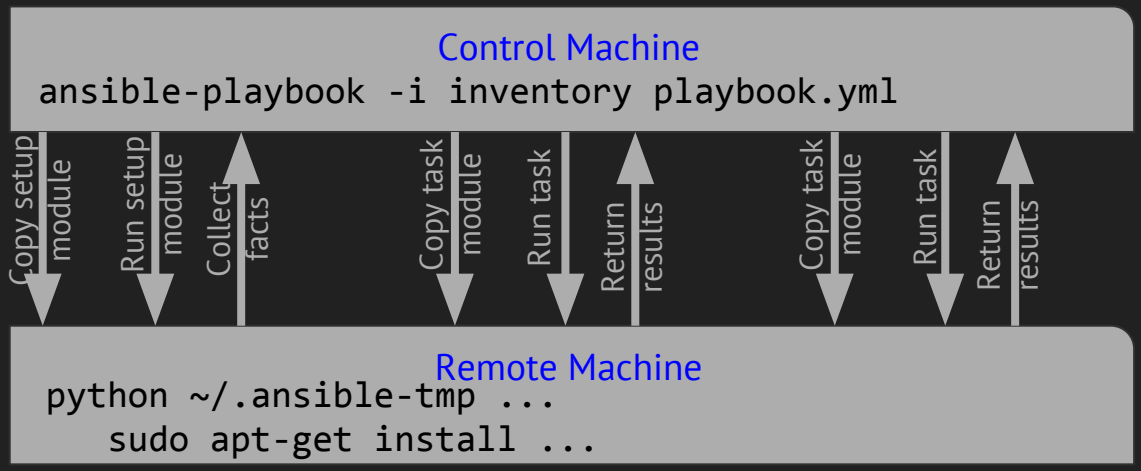
We use Ansible as the technology of choice

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

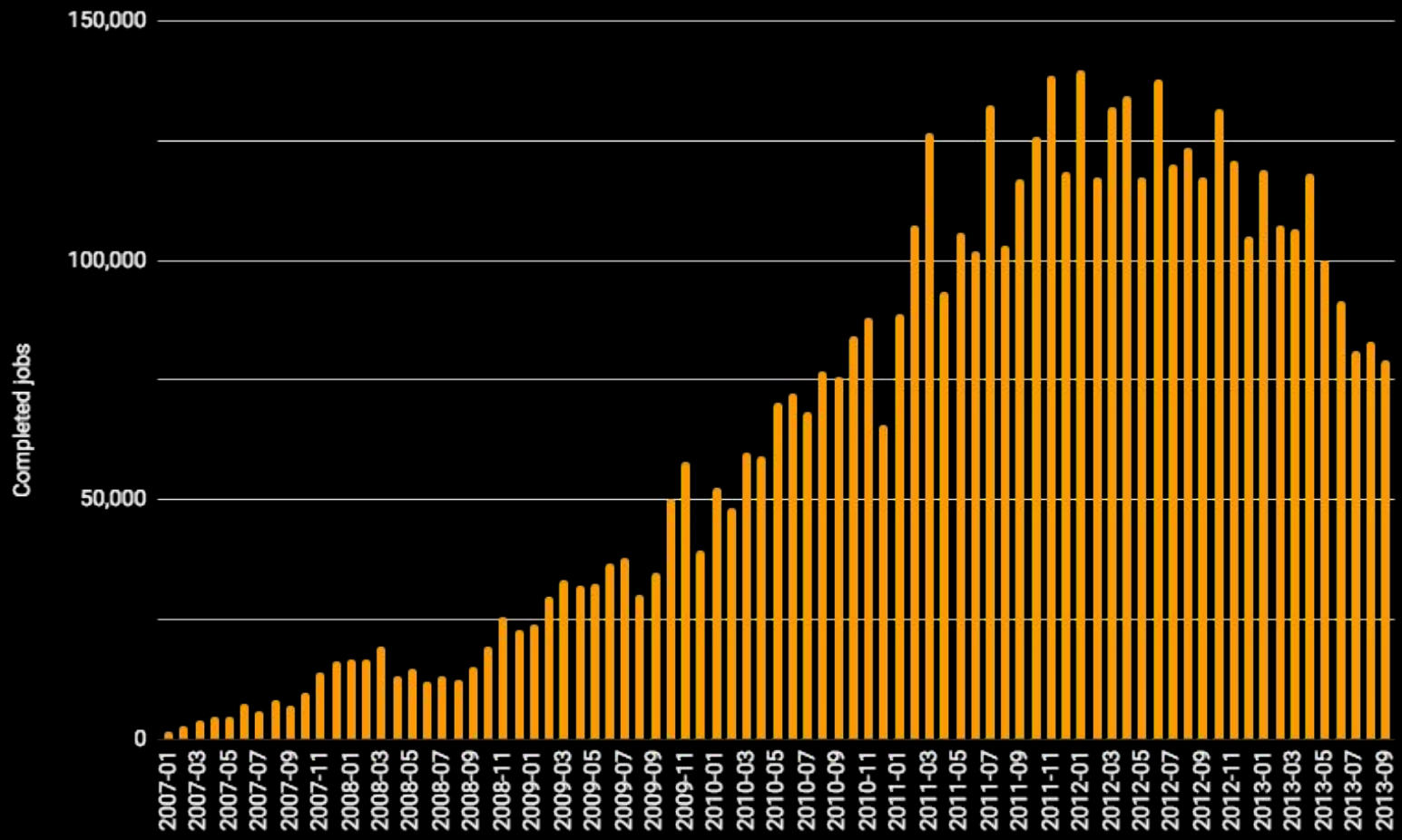
2PB
user data

18M
jobs run

100
training events
(2017 & 2018)

Stats for Galaxy Main (*usegalaxy.org*) in March 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)

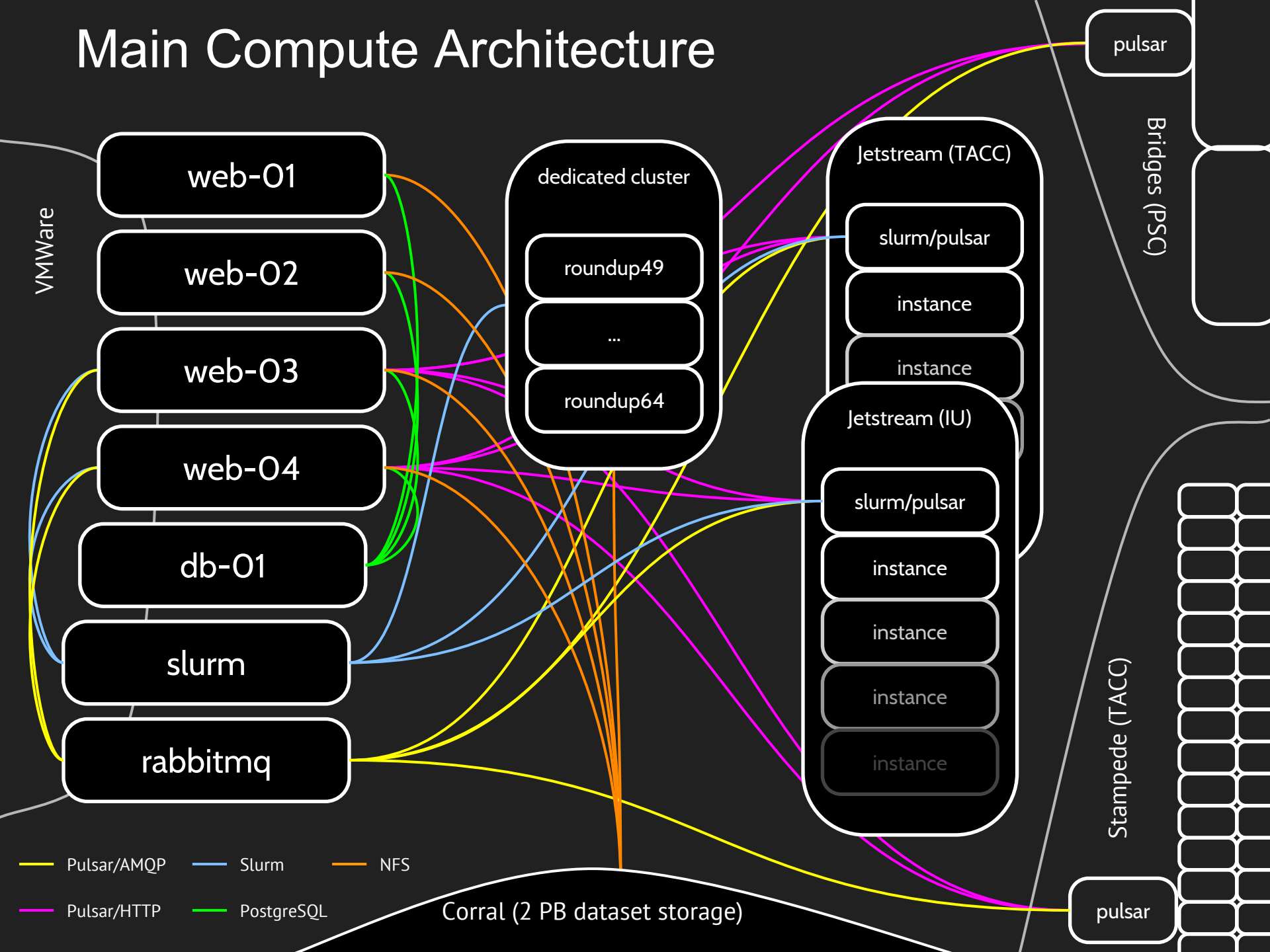
- Leverage Pulsar for data staging and Galaxy resource/job manager

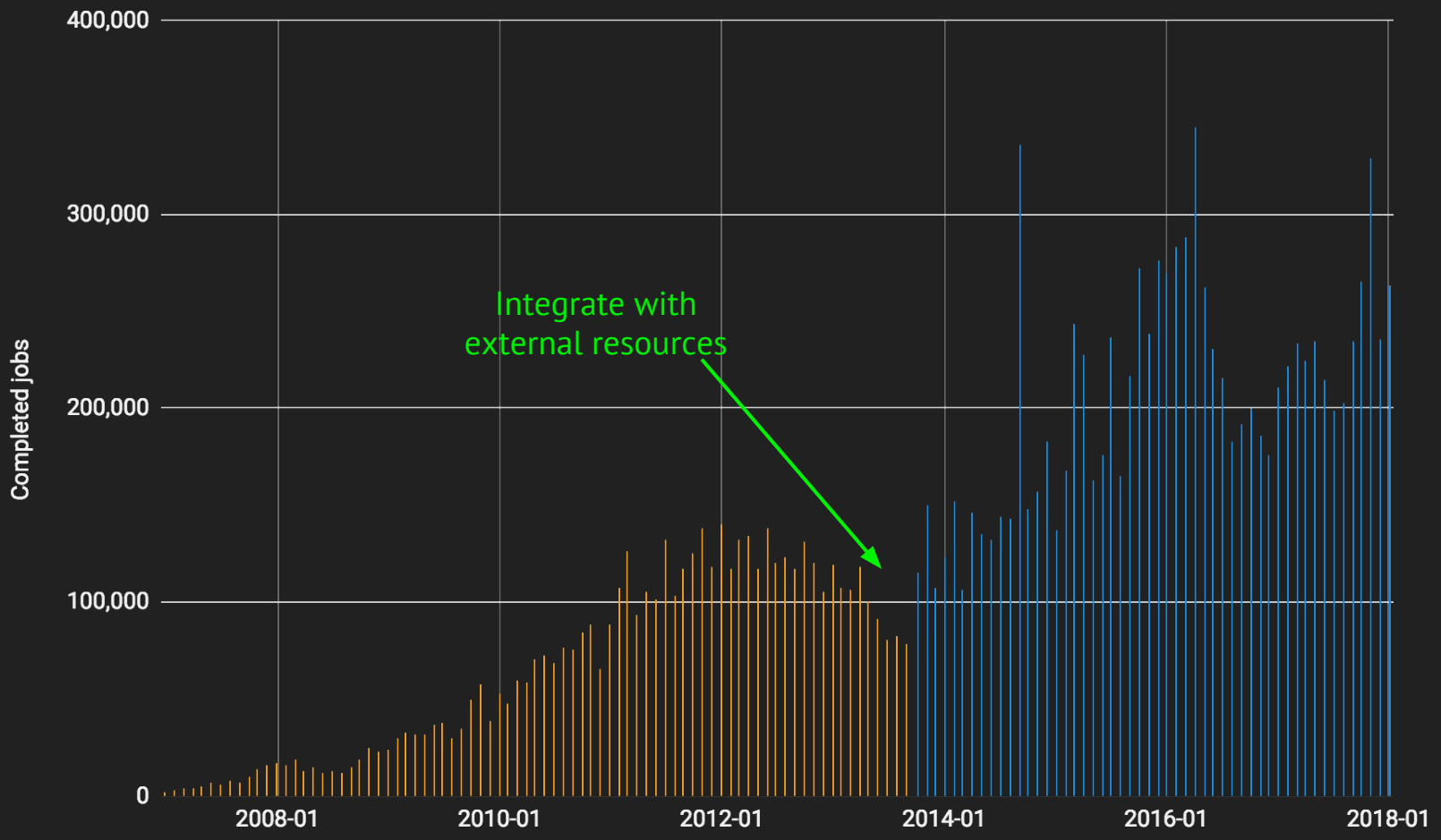
Expanding compute capacity

Leveraging National Cyberinfrastructure: Galaxy/XSEDE Gateway

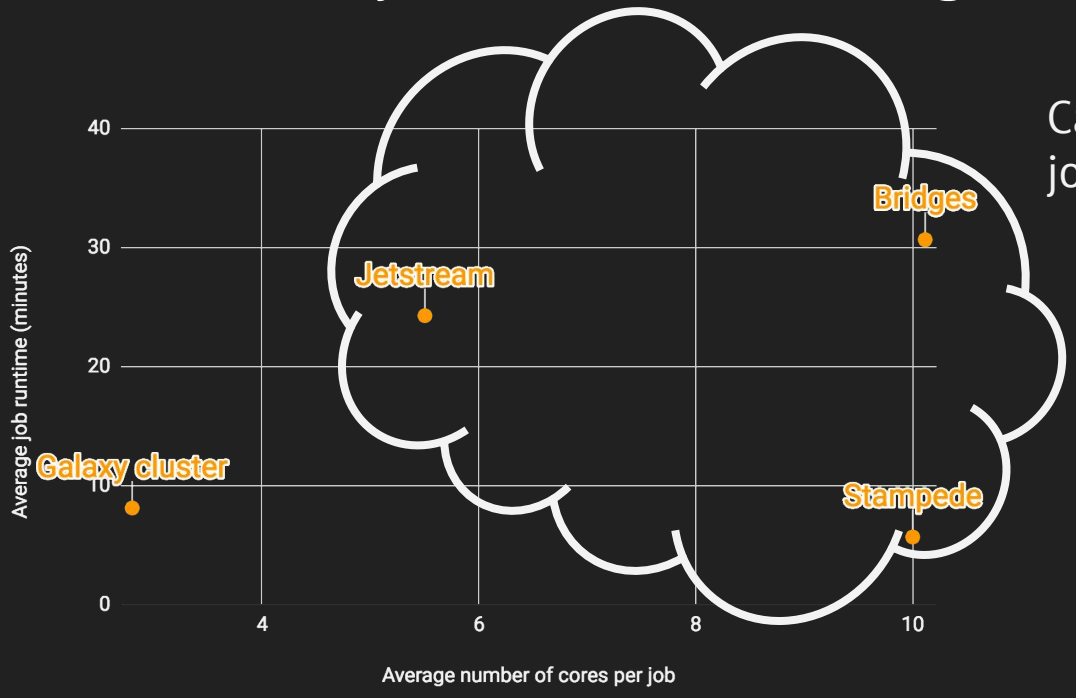
Shared XSEDE resources
Dedicated resources

Main Compute Architecture




400,000
300,000
200,000
100,000
0
Completed jobs
Integrate with
external resources
2008-01
2010-01
2012-01
2014-01
2016-01
2018-01

More than job counts, scaling moved the horizon

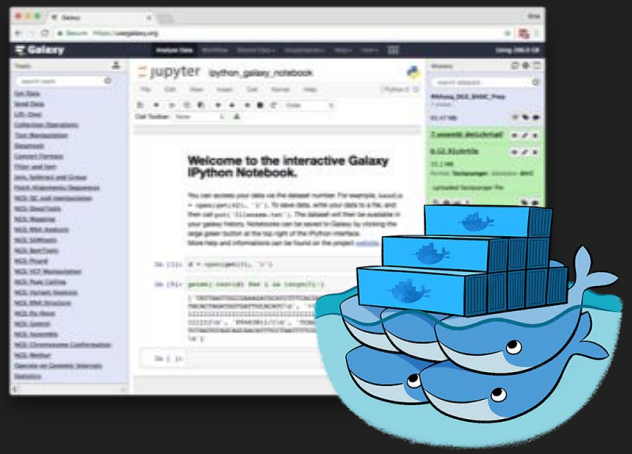


Can now run larger jobs, and more jobs:

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

Flexibility

Usability

100+ public Galaxy Servers

bit.ly/gxyServers

ARGalaxy

Biologie Fonctionnelle, Insectes et Interactions

CENTRUM MEDISCHE GENETICA

biomina

MIMODD

VectorBase
Bioinformatics Resource for Invertebrate Vectors of Human Pathogens

Human Communication Science Virtual Laboratory

Above and beyond speech, language and music: a Virtual Laboratory for Human Communication Science

SAINT Galaxy APOSTL

AspenDB

Tsai Lab at UGA

ballaxy

Powered by the Biochemical Algorithms Library Project

BioMaS

Bioinformatic analysis of Metagenomic AmpliconS

ABiMS

Analyses and Bioinformatics for Marine Science

EMBRC FRANCE ifb Station Biologique Roscoff cnrs UPMC

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

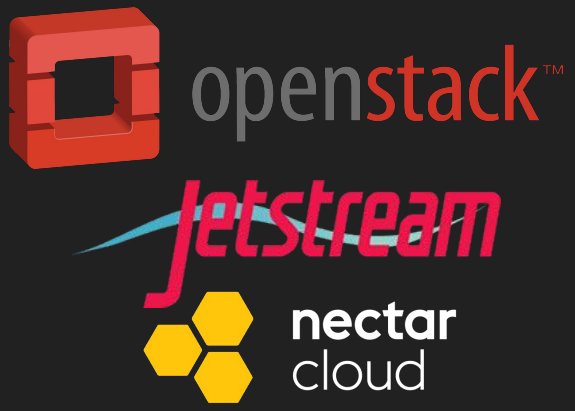
Flexibility

Usability

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

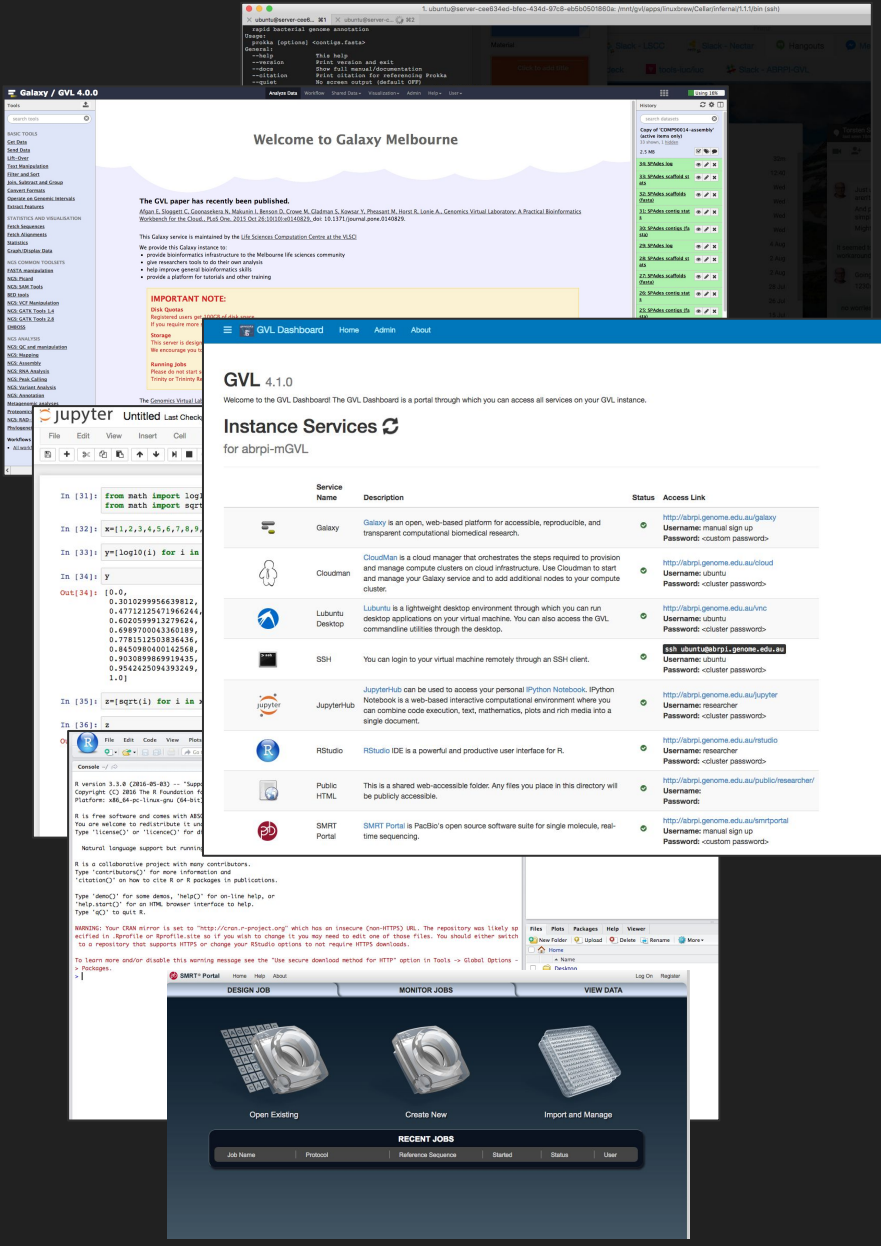
- Based on CloudMan: a cloud manager for deploying Galaxy on a variety of cloud providers



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.

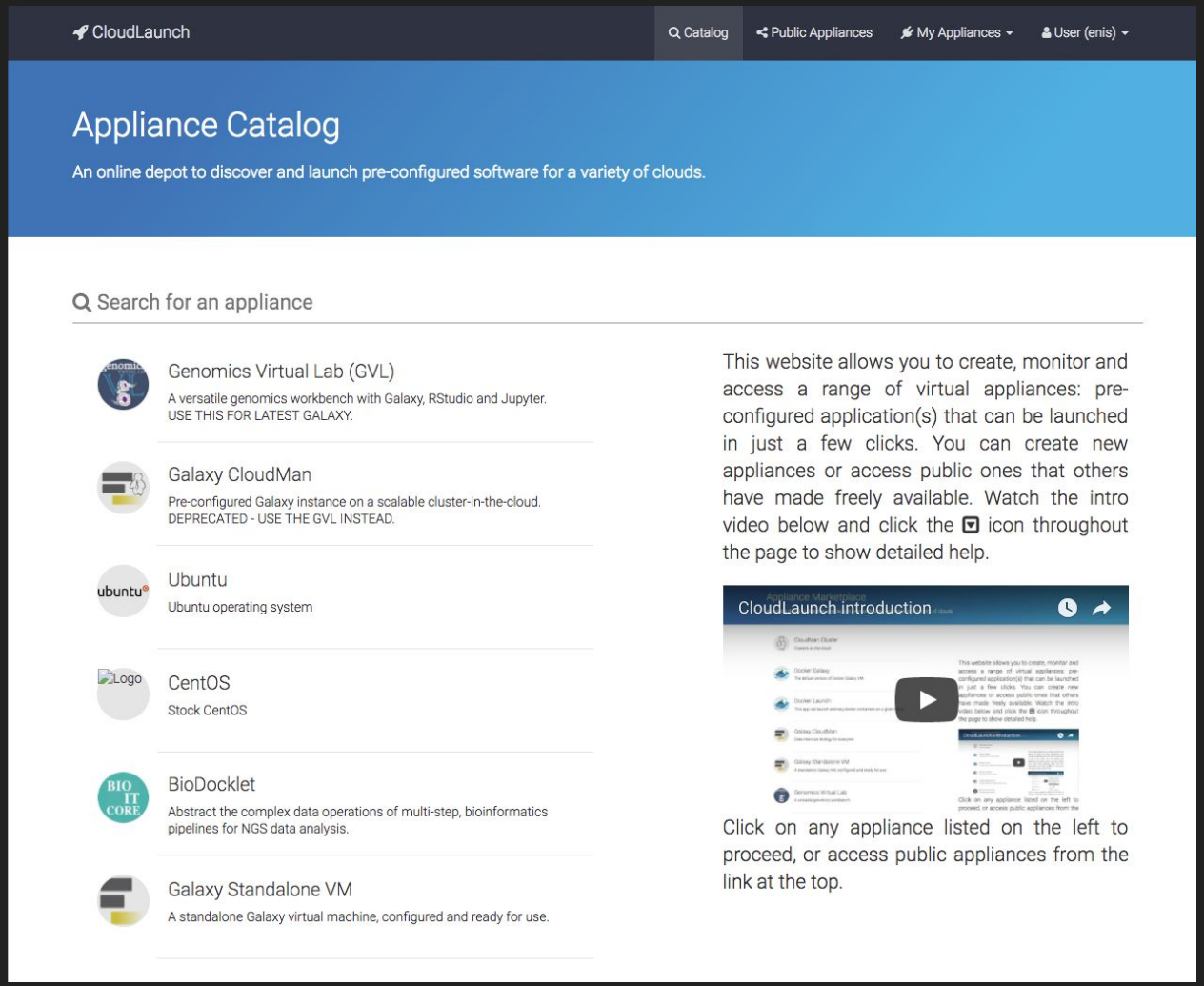


genomics VIRTUAL LAB

gvl.org.au

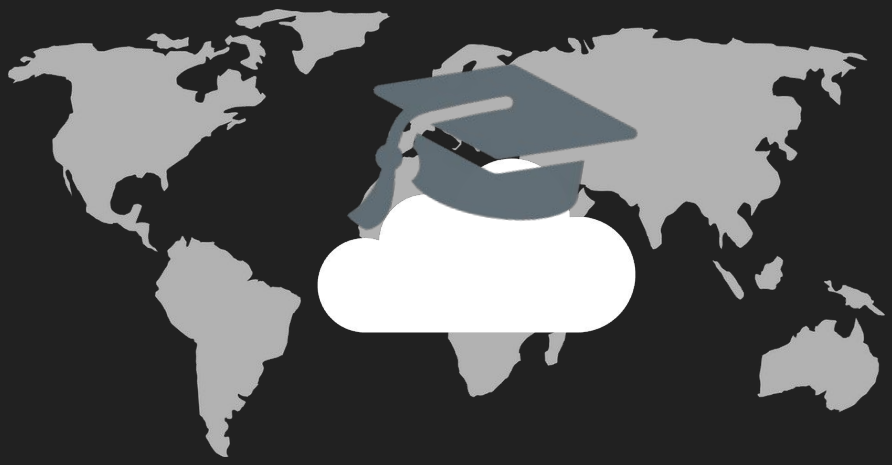
THE UNIVERSITY OF MELBOURNE

Launch-your-own, via CloudLaunch



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

Scaling across clouds



Towards Galaxy-on-the-cloud 2.0

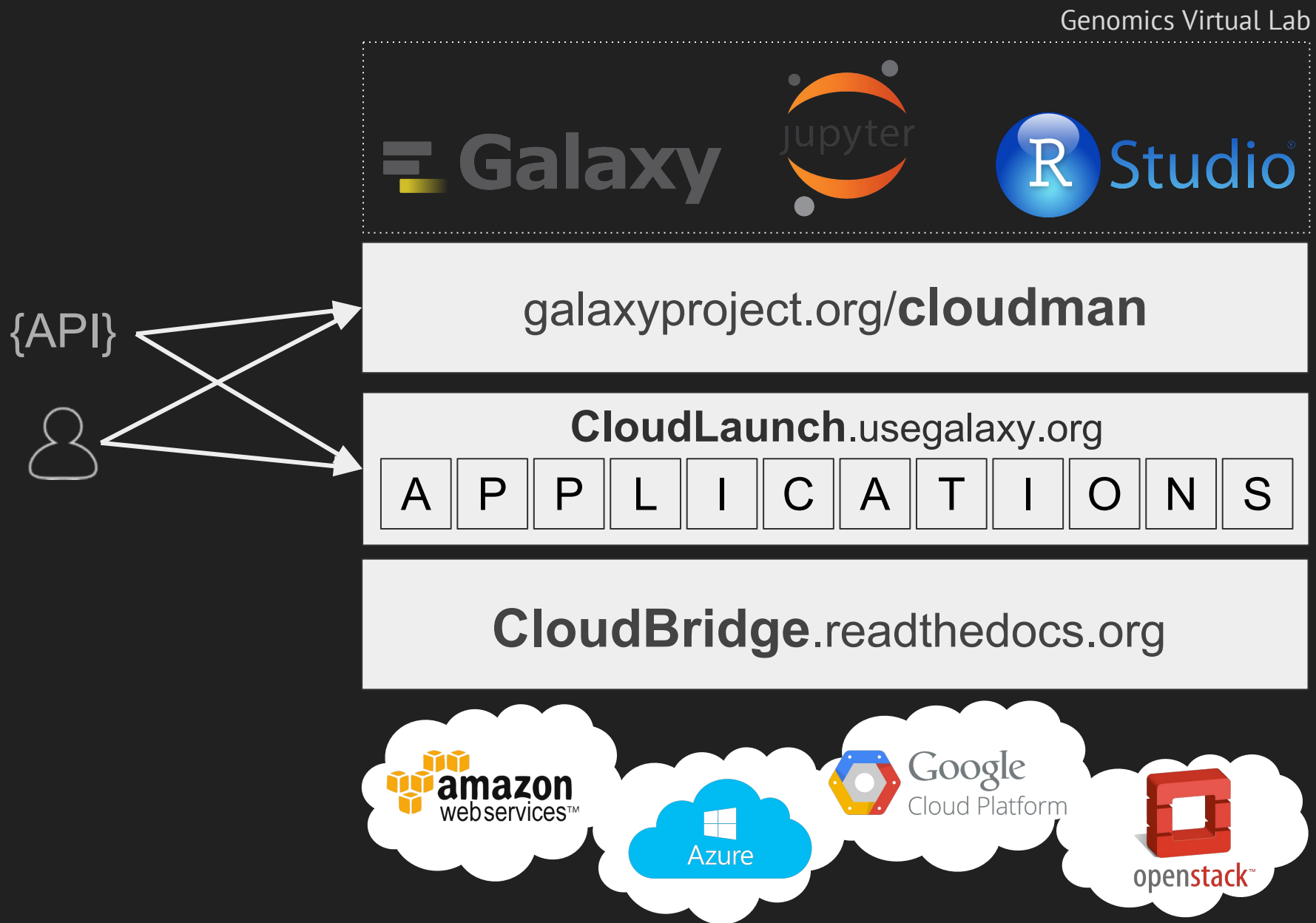
Goals:

- 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?



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

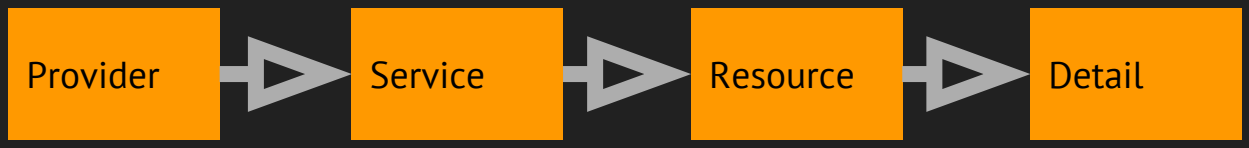
CloudBridge: a simple, open-source Python multi-cloud library.

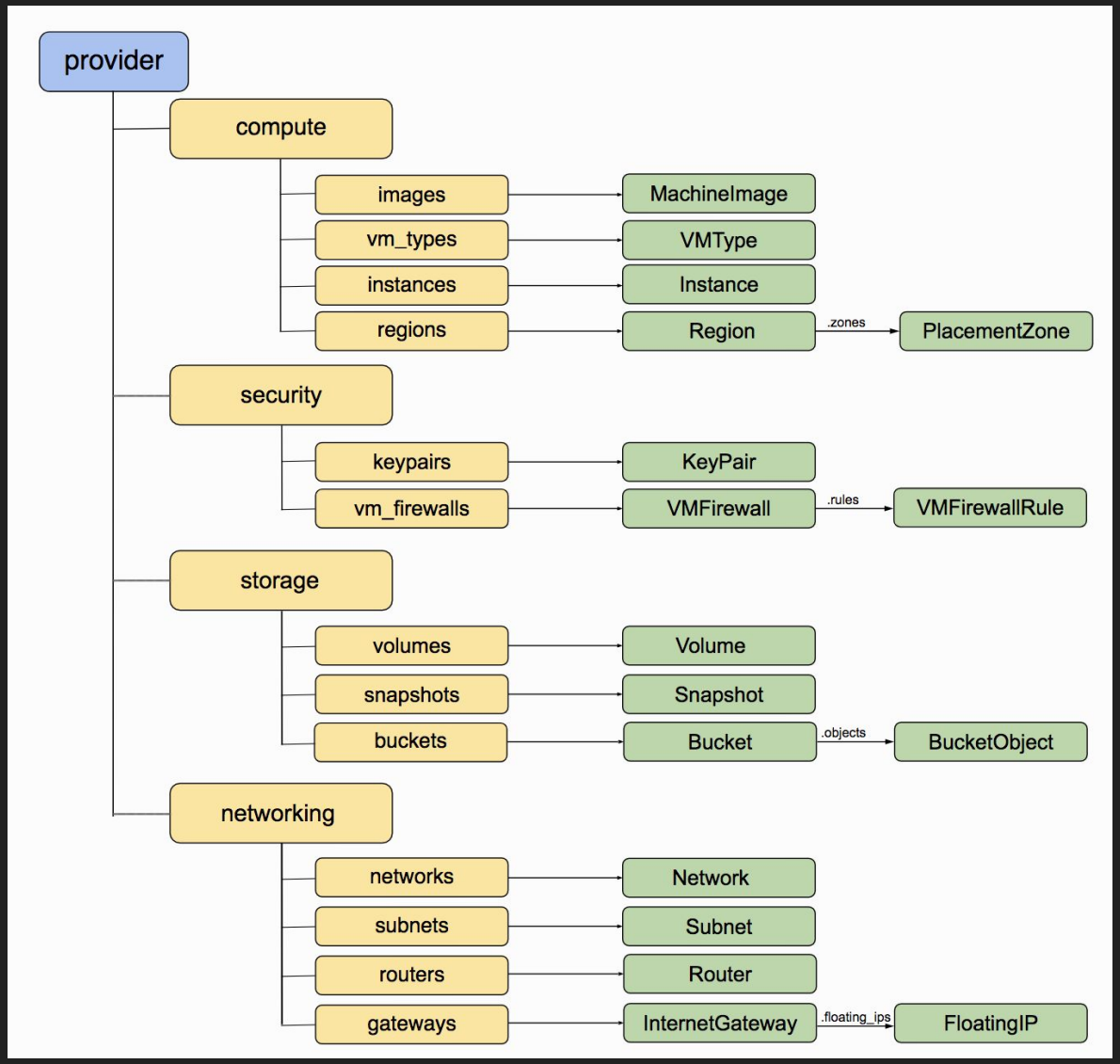
Uniform API irrespective of the underlying provider
No special casing of application code, unlike Apache Libcloud
Simpler code

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


Provider
Service
Resource
Detail

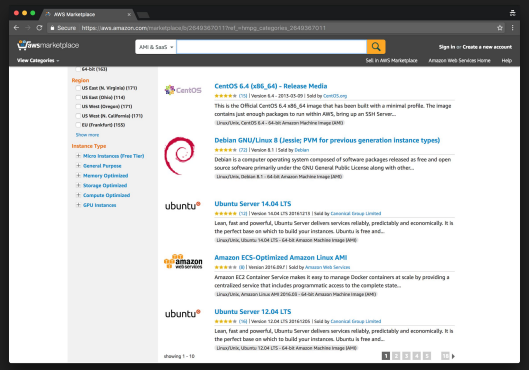

provider
compute
images
MachineImage
vm_types
VMType
instances
Instance
regions
Region
.zones
PlacementZone
security
keypairs
KeyPair
vm_firewalls
VMFirewall
.rules
VMFirewallRule
storage
volumes
Volume
snapshots
Snapshot
buckets
Bucket
.objects
BucketObject
networking
networks
Network
subnets
Subnet
routers
Router
gateways
InternetGateway
.floating_ips
FloatingIP

CloudLaunch

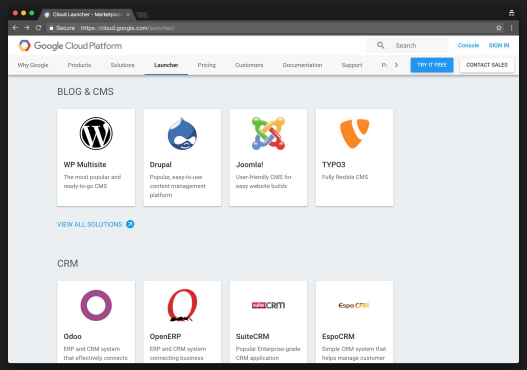
A gateway for discovering and launching applications on a variety of clouds.

Afgan, E., Lonie, A., Taylor, J., Goonasekera, N., "**CloudLaunch: Discover and Deploy Cloud Applications**", *(in press)*, Future Generation Computer Systems.

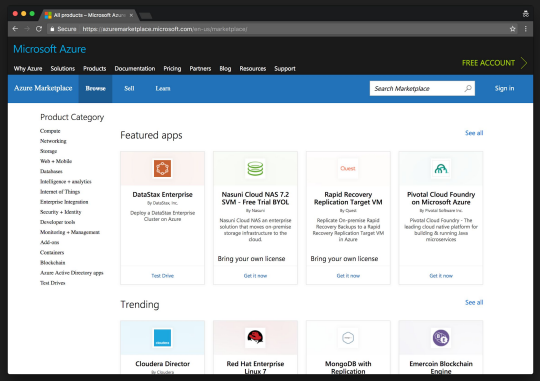
CloudLaunch-as-a-Service



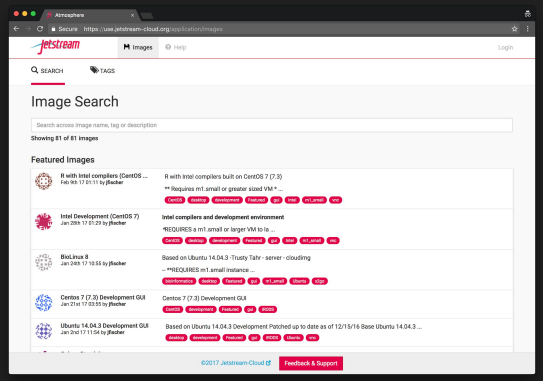
AWS Marketplace



GCE Solutions



Azure Marketplace



Jetstream Atmosphere VMs

CloudLaunch features

Cloud-agnostic

Backed by CloudBridge, use native cloud capabilities for infrastructure management

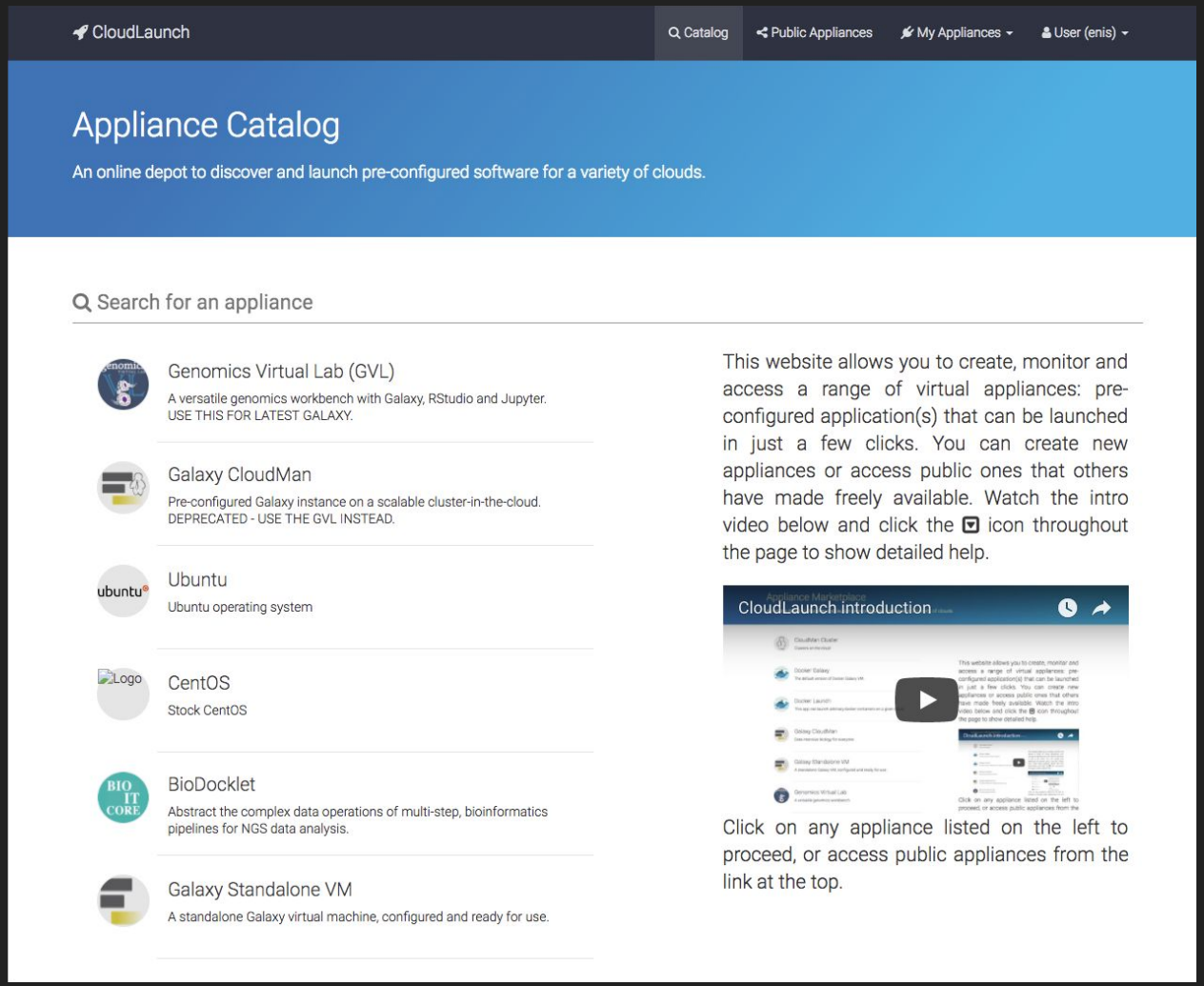
Pluggable and extensible

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



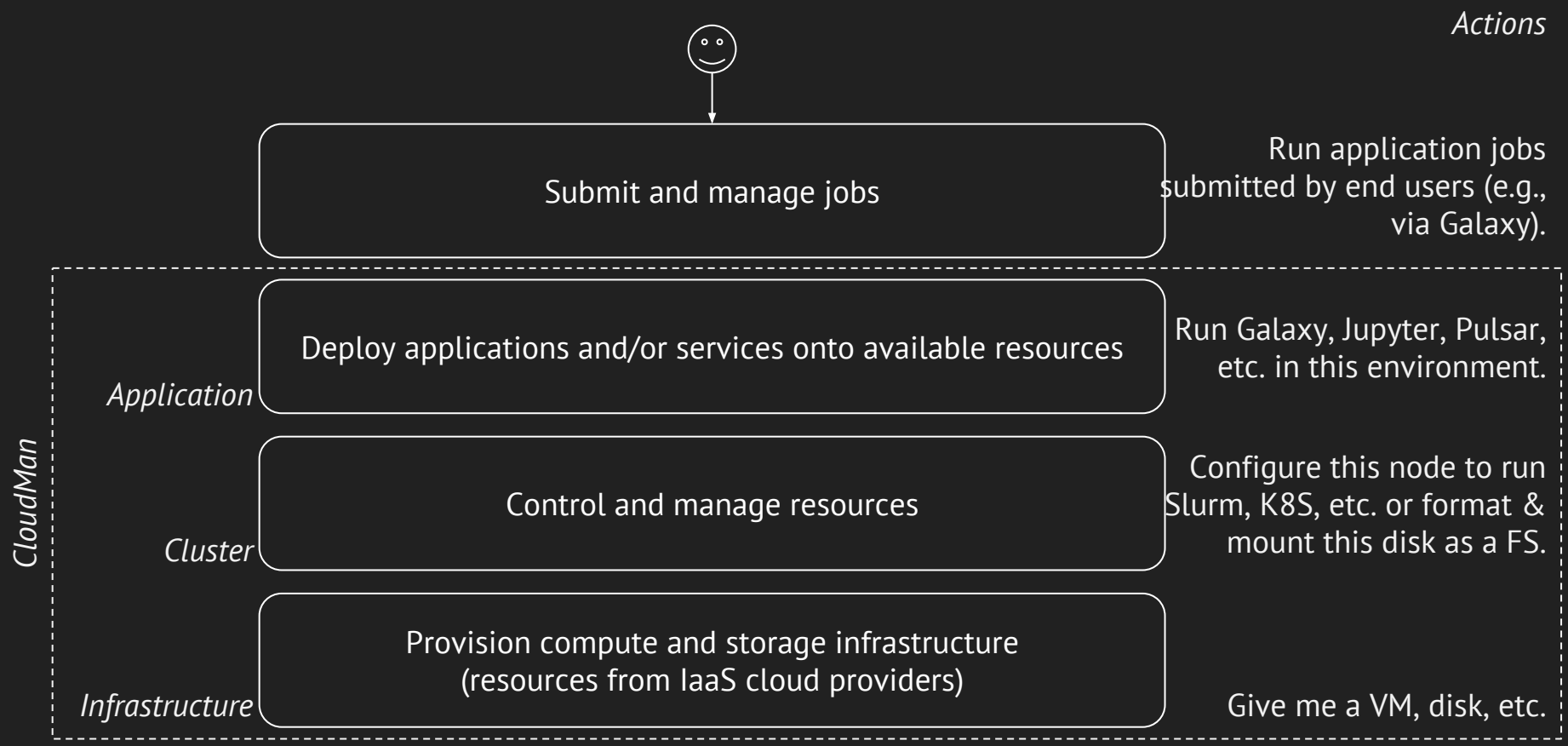
Demo using https://launch.usegalaxy.org

CloudMan and beyond

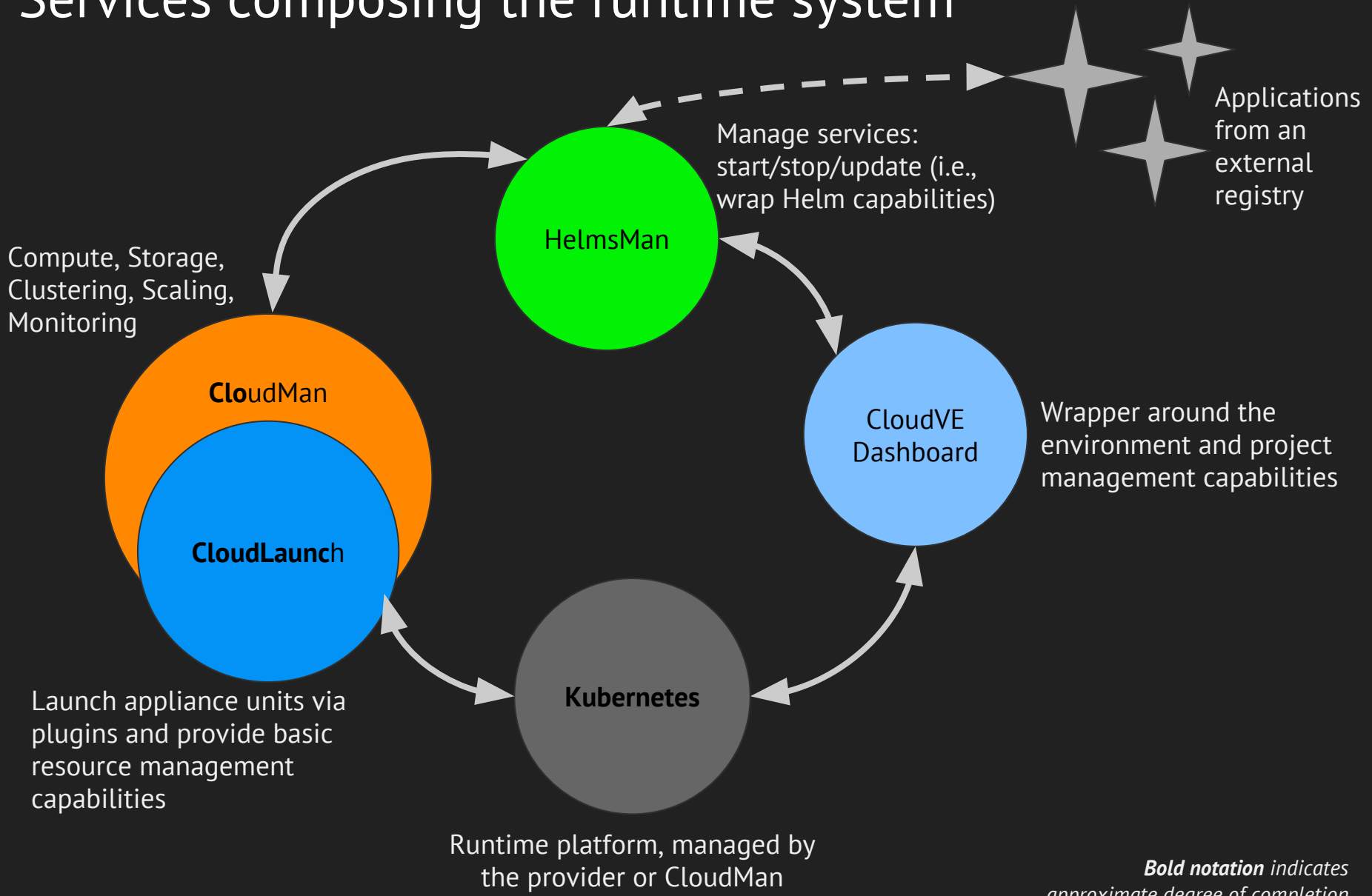
Manage deployed infrastructure and applications

Managing deployed infrastructure

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



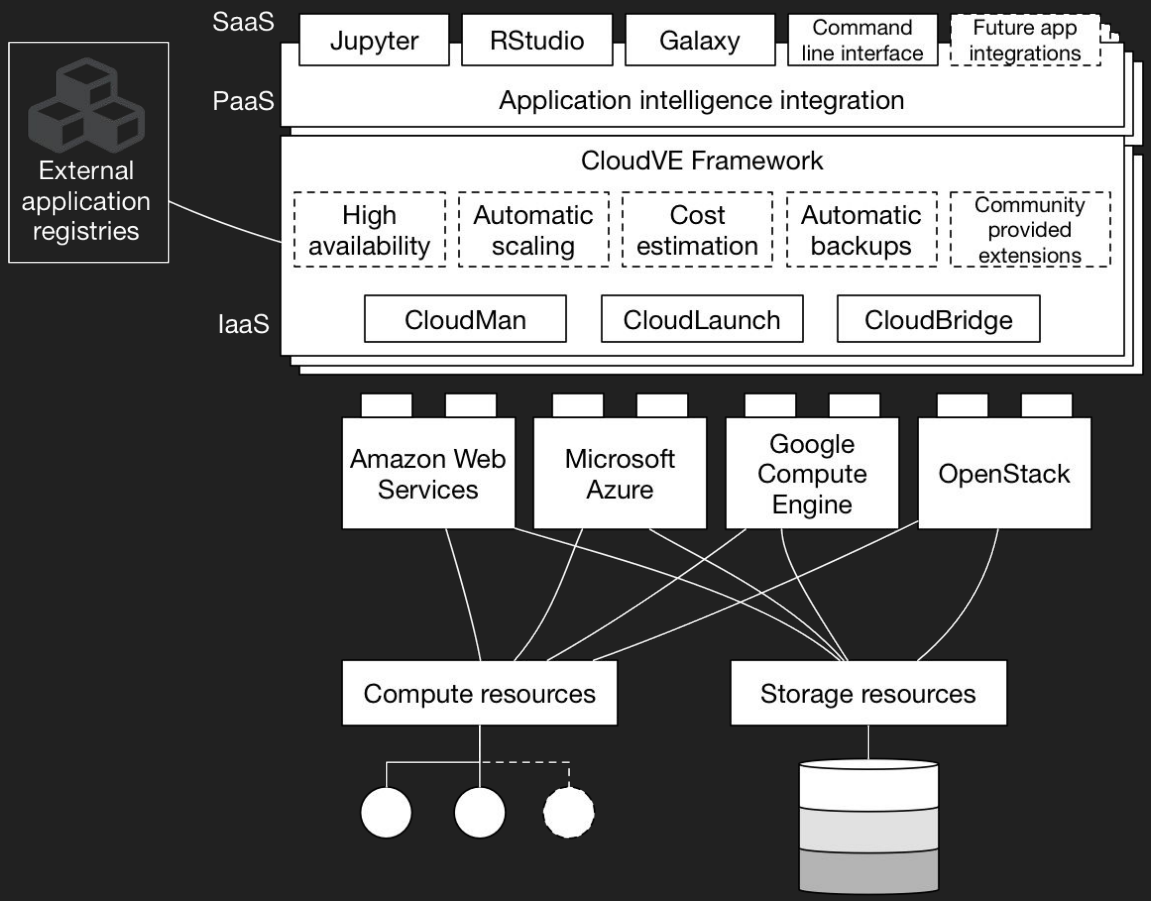
Services composing the runtime system



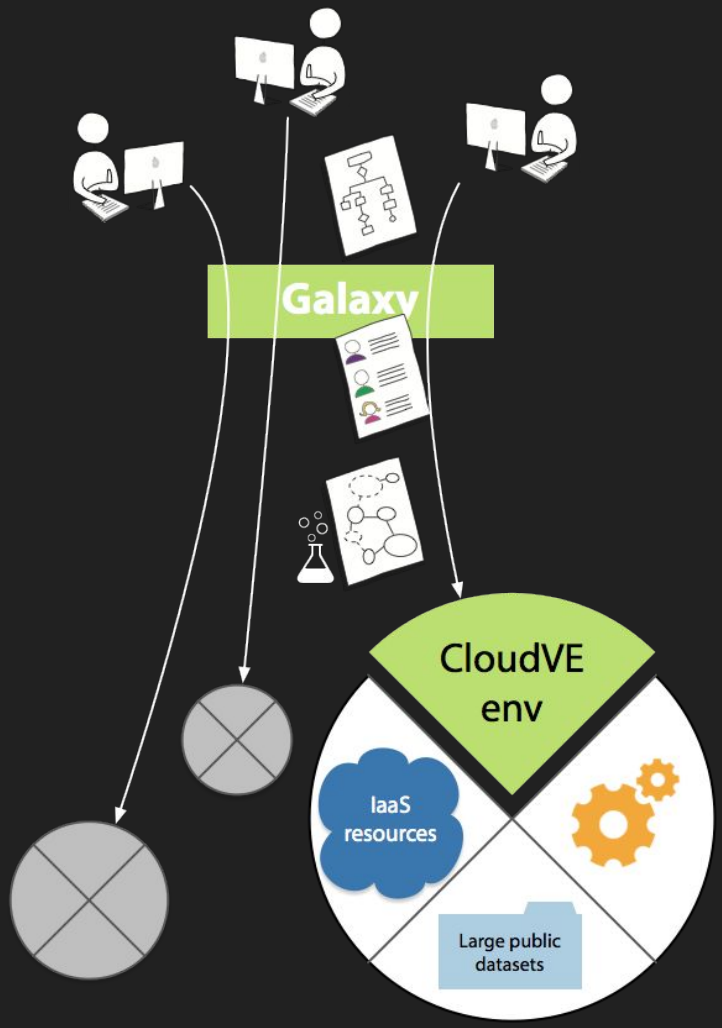
Bold notation *indicates approximate degree of completion*

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

- Cumulati
- Customiza
- Can grow ve

zable

nd

enance

arge

components

Acknowledgments

Institutions

Johns Hopkins University

Penn State

OHSU

The University of Melbourne

Projects

genomics virtual lab

Galaxy

SGCI Science Gateways Community Institute

Infrastructure

XSEDE

Jetstream

amazon web services™

Azure

nectar cloud