CloudCRV, Virtual Cluster Appliance and ATLAS

Deploy Your Cluster to the Cloud with 1-Click

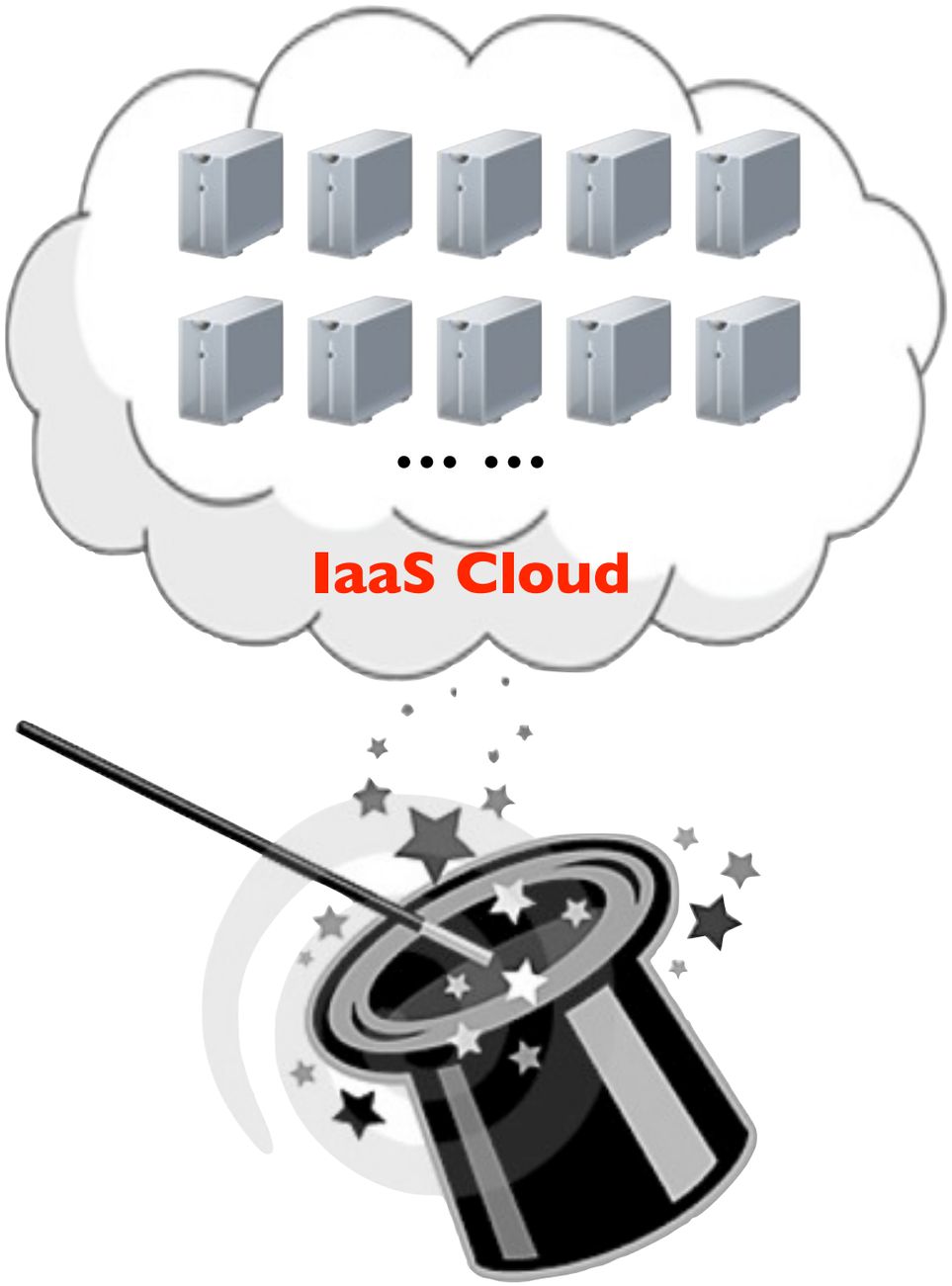
Yushu Yao
Lawrence Berkeley National Laboratory

Overview

- We are trying to investigate if cloud will help ATLAS.
- There have been a lot of debate on performance, cost, operation model of it.
- However, to do any of the above, we need an easy way to deploy an ATLAS Computing Cluster onto the cloud, and send our workload there.
- We developed a tool to do that, and sharing it here...

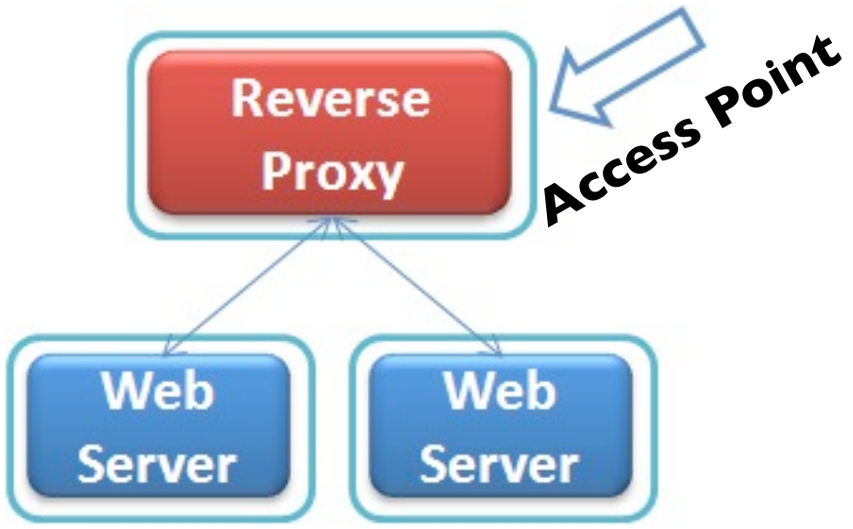
The Cloud is Good

- With the help of IaaS* cloud (Such as Amazon EC2), anyone can lease a large amount of Computing Resource
- E.g. get 100 linux computers
- The idea is: Any cloud user can setup their own cluster, and benefit from the scalability that the cloud provides.
- However, what do I do with the 100 computers?



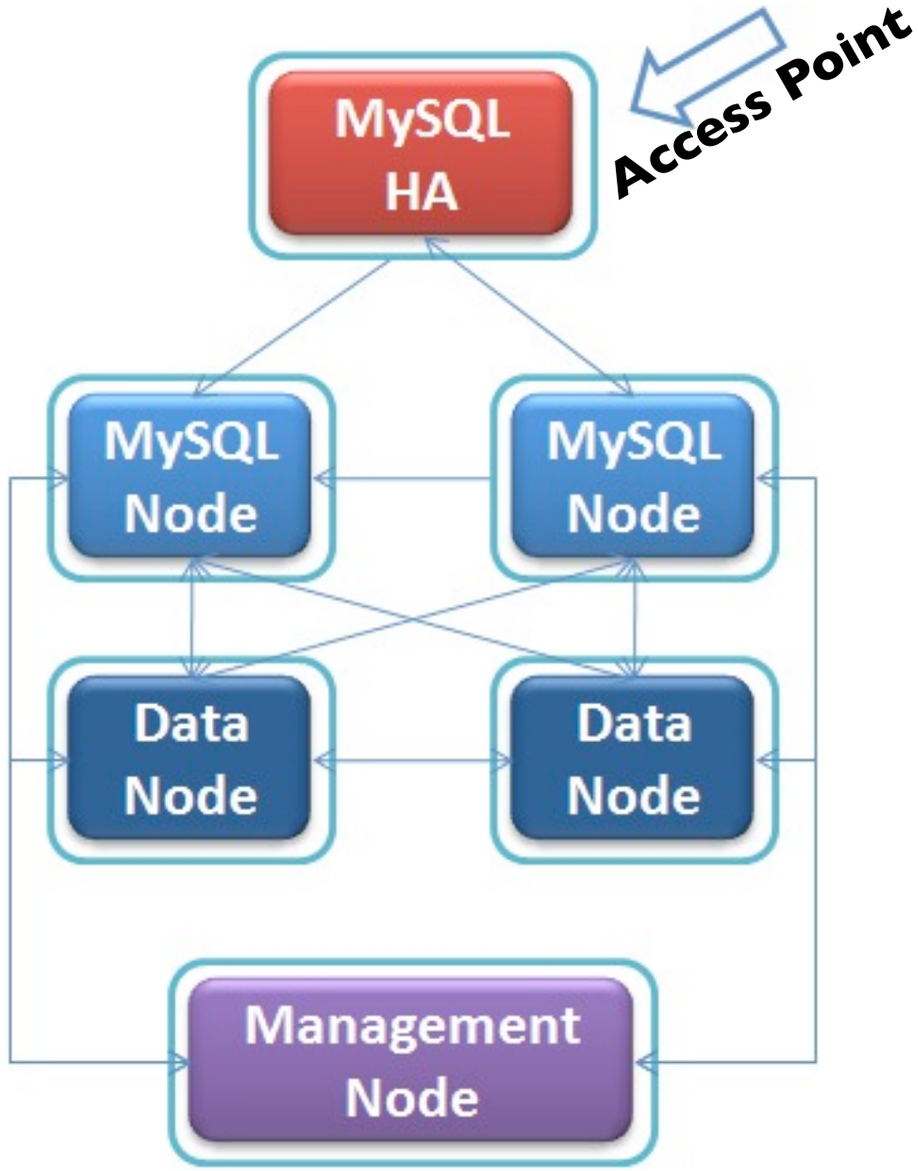
***IaaS**: Infrastructure as a Service, e.g EC2, Eucalyptus, Nimbus, Open Nebula?

Deploy This Cluster is NOT Easy



Load Balanced
Web Server Cluster

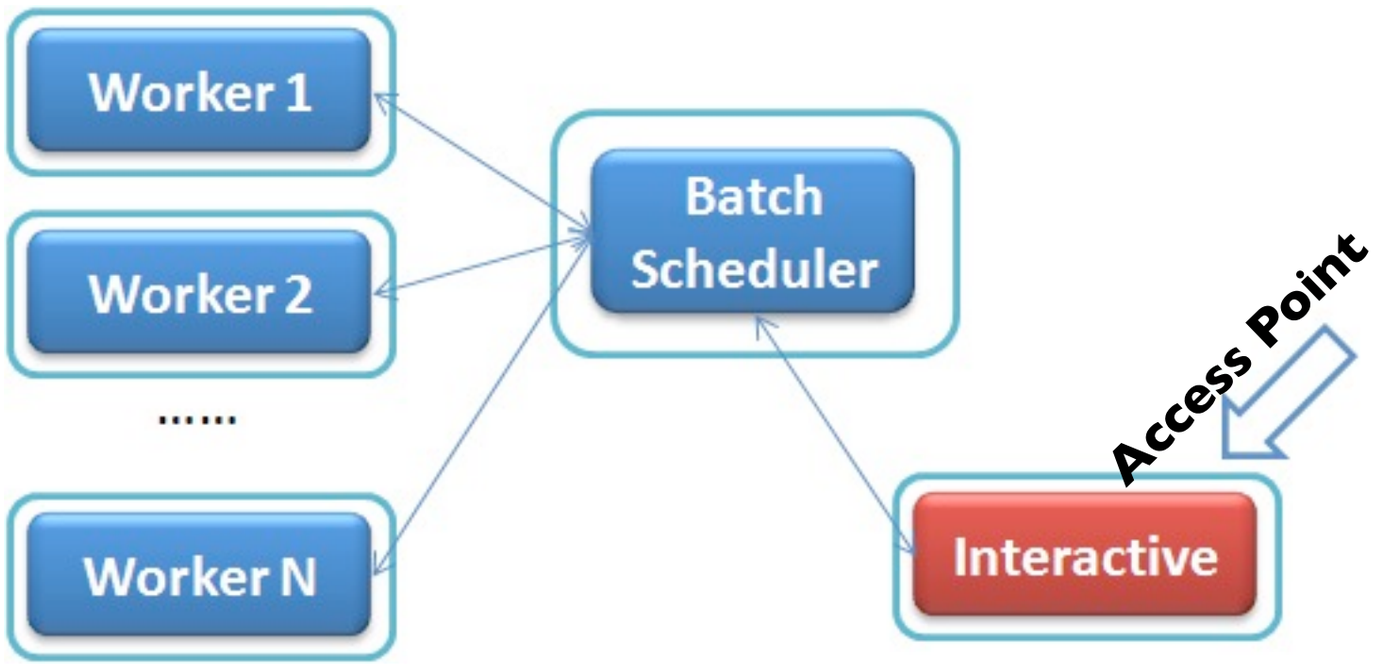
Deploy This Cluster is NOT Easy



MySQL Cluster

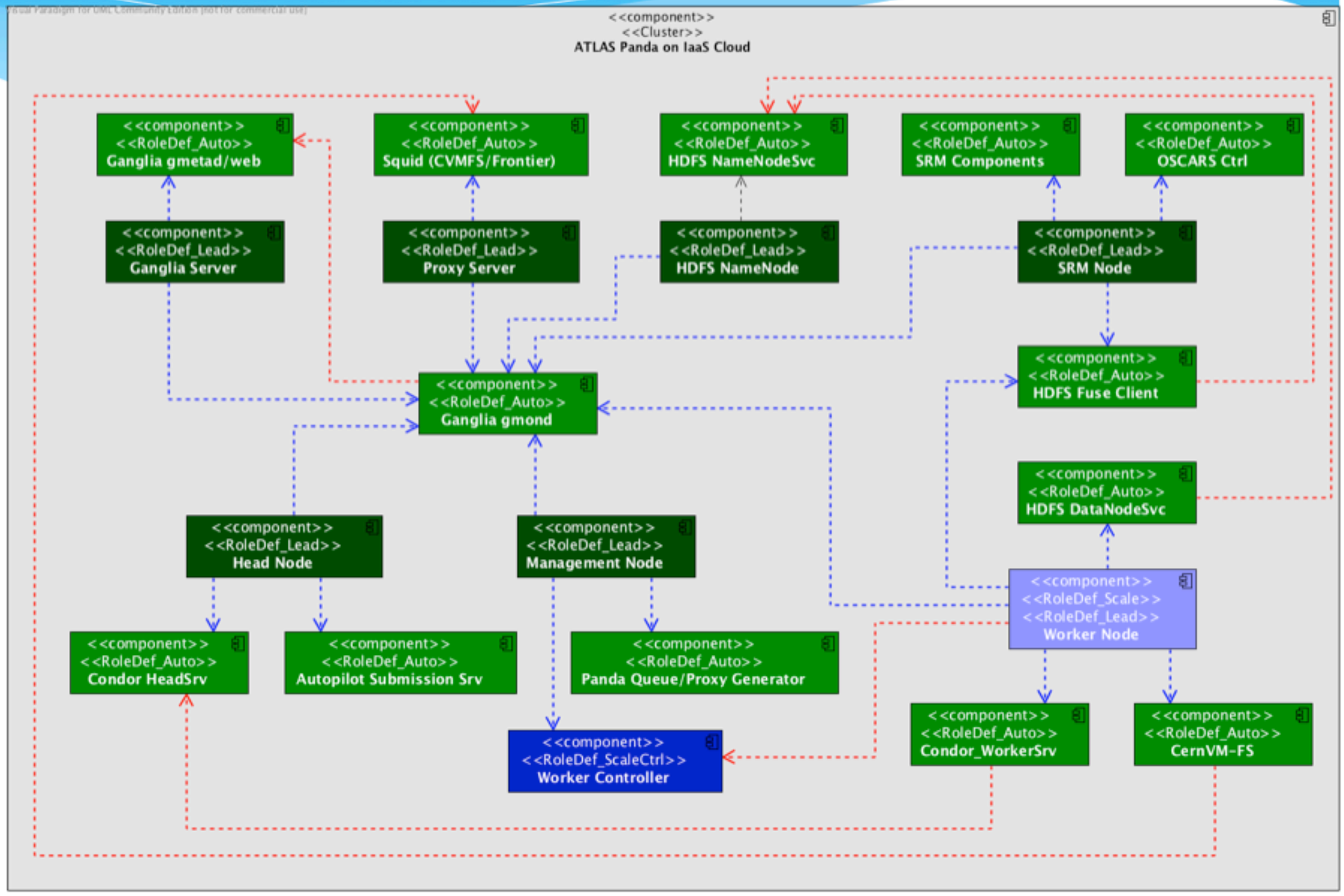
Deploy This Cluster is NOT Easy

Deploy This Cluster is NOT Easy



Auto-Scaling Batch Cluster (Condor Pool, etc)

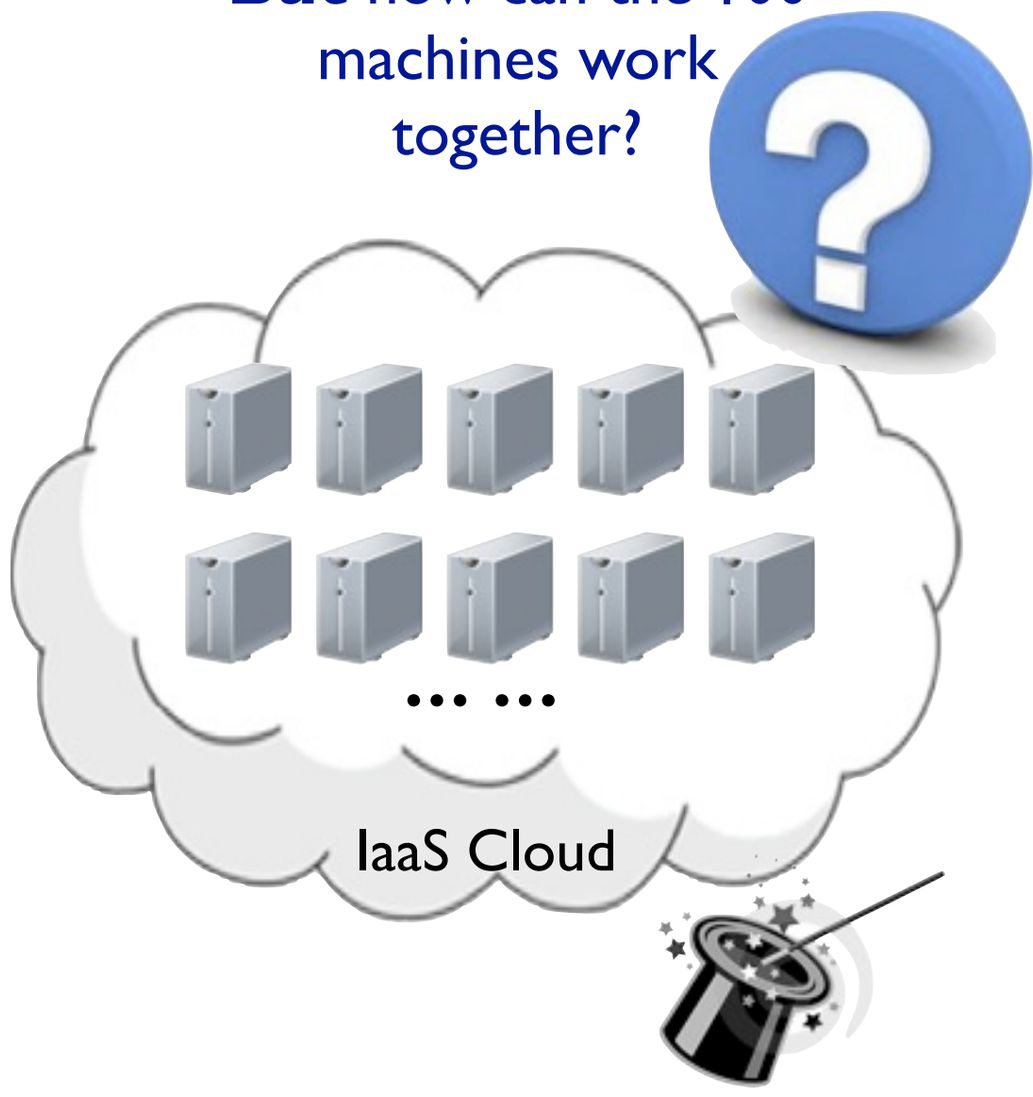
Try This? (The ATLAS Cluster)



Really Good?

- I would like the machines to work as a complex cluster. How to configure them?
- I'm not an IT professional, I don't know how to do it
- Even if I know how to configure, it's too much work to login to 100 machines and do it by hand
- How can I scale the size of the cluster on-demand?

Yes, I can easily get 100 machines.
But how can the 100 machines work together?

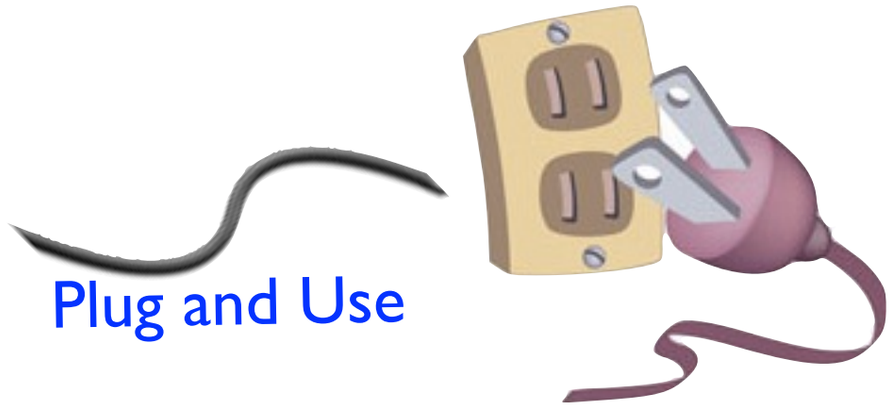


An Analogy

Parts in a Fridge

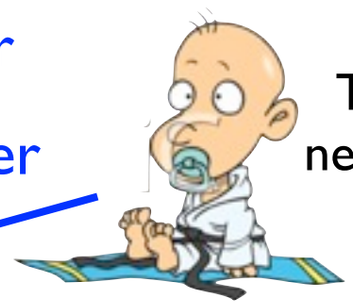
Manufacture

Fridge



Plug and Use

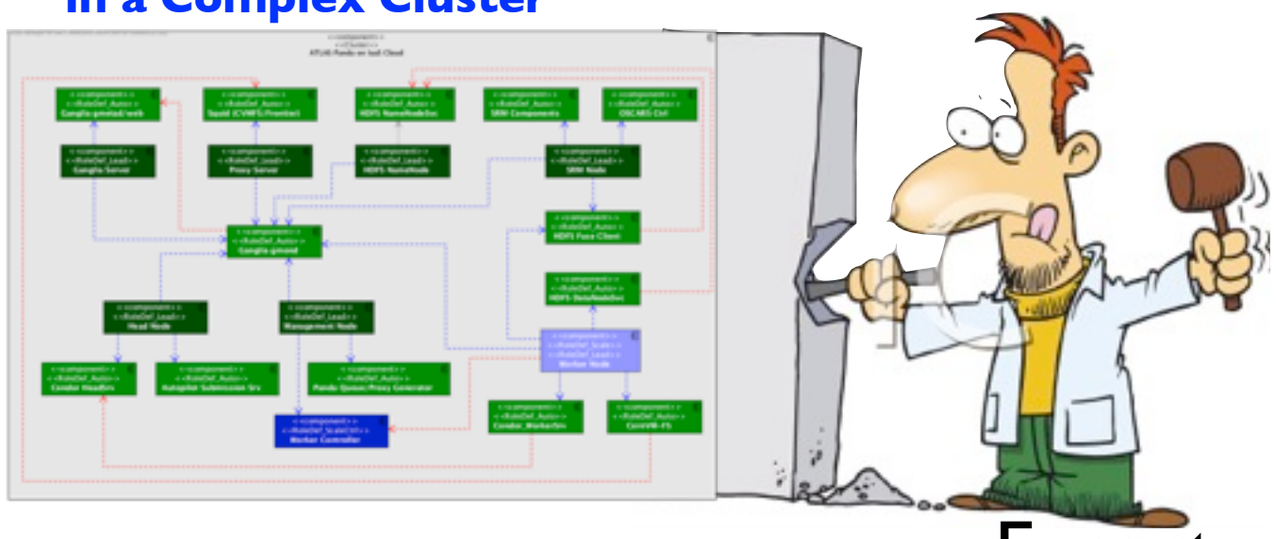
Substances	A Cluster	A Fridge
	Various of Apps and Services	Parts in a Fridge
	Computing Power from Cloud	Electricity
Actors	Cloud Provider	Electricity Provider
	Cluster Designer	Fridge Manufacturer
	Cluster Manager/Deployer	Fridge User
Actions	Design a Cluster	Manufacture a Fridge (Need expert)
	Deploy a Cluster	Plug the Fridge to the Wall (Simple)



The guy who needs to deploy a Cluster

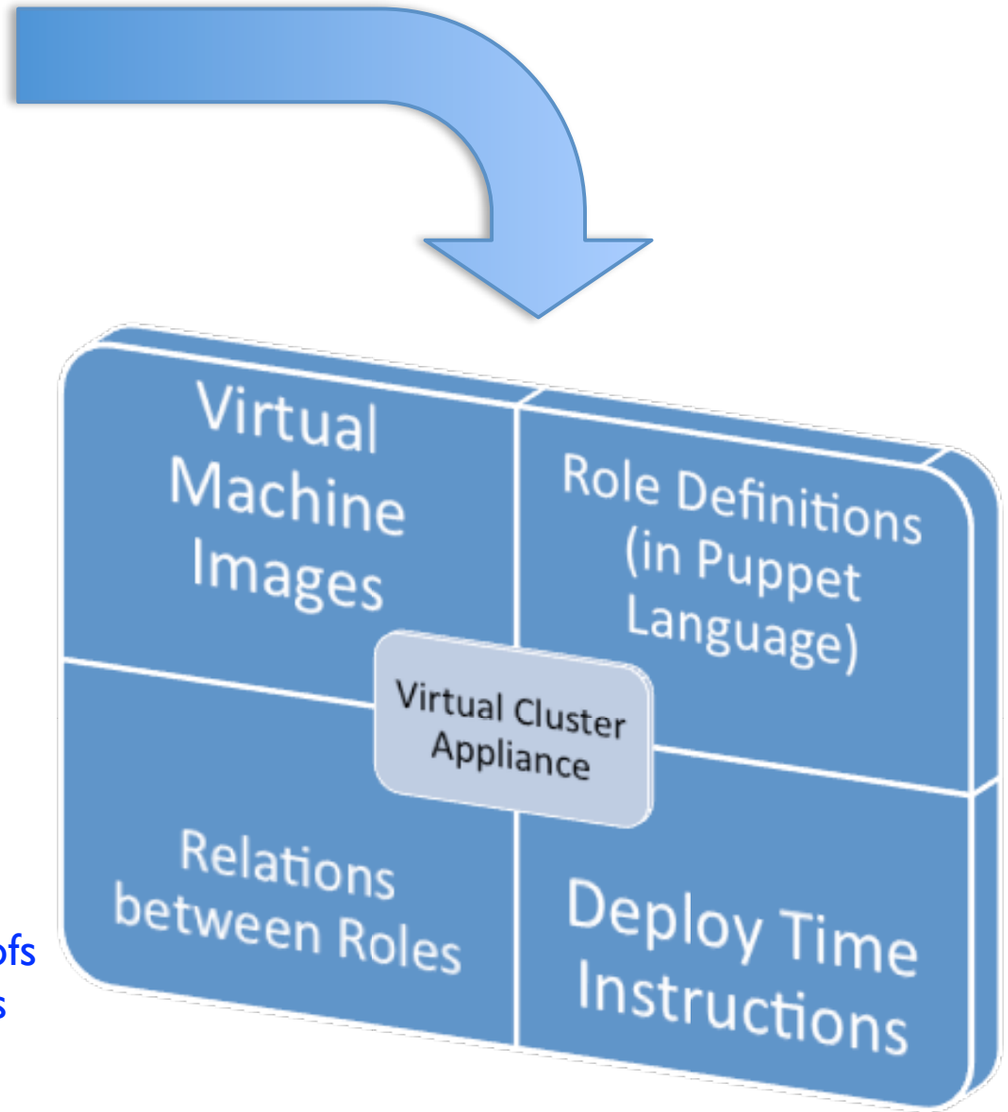
Pack and Ship a Cluster like a Fridge

The Components (roles) in a Complex Cluster



Expert

Create a Virtual Cluster Appliance



Ship to cluster manager

Remember those univ. profs with credit cards? Or his postdoc?

Virtual Cluster Appliance (VCA)

BERKELEY LAB

Virtual Appliance

Pre-installed pre-configured OS and Application Stack That performs certain tasks

What if my tasks need more than one VM to perform?

Define Multiple Virtual Machines to hold these services/applications

+

How they work together

Virtual Cluster Appliance

Virtual Cluster Appliance = { **Multiple Virtual Machines** + **Multiple Services/Applications** + **Their Relationships**

Lifecycle of a VCA

- ✦ VCAs are Designed by Experts:
  - ✦ Use scripts ***(bash, puppet, cfengine, etc)*** to define:
    - ✦ How to configure each computer
    - ✦ How one computer work with the rest

- ✦ Virtual Cluster Appliances are Packaged into a set of Scripts and specifications
  - ✦ Easy to ship/copy

- ✦ To deploy a VCA, one need a tool to:
  - ✦ Control the lifetime of the VMs inside the cluster
  - ✦ Configure the VMs to work together
- ✦ **CloudCRV (Cluster-Roles-VMs) is a tool to do this**

Deploying a VCA like Plugging in a Fridge

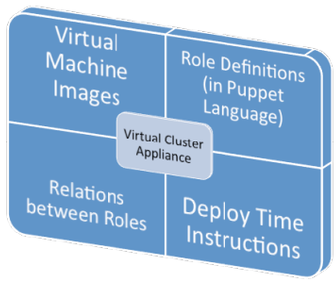
Cluster Manager

Deploy

Start Up CloudCRV

CloudCRV

Specify a Virtual Cluster Appliance



Set basic Parameters Eg. EC2 Account?

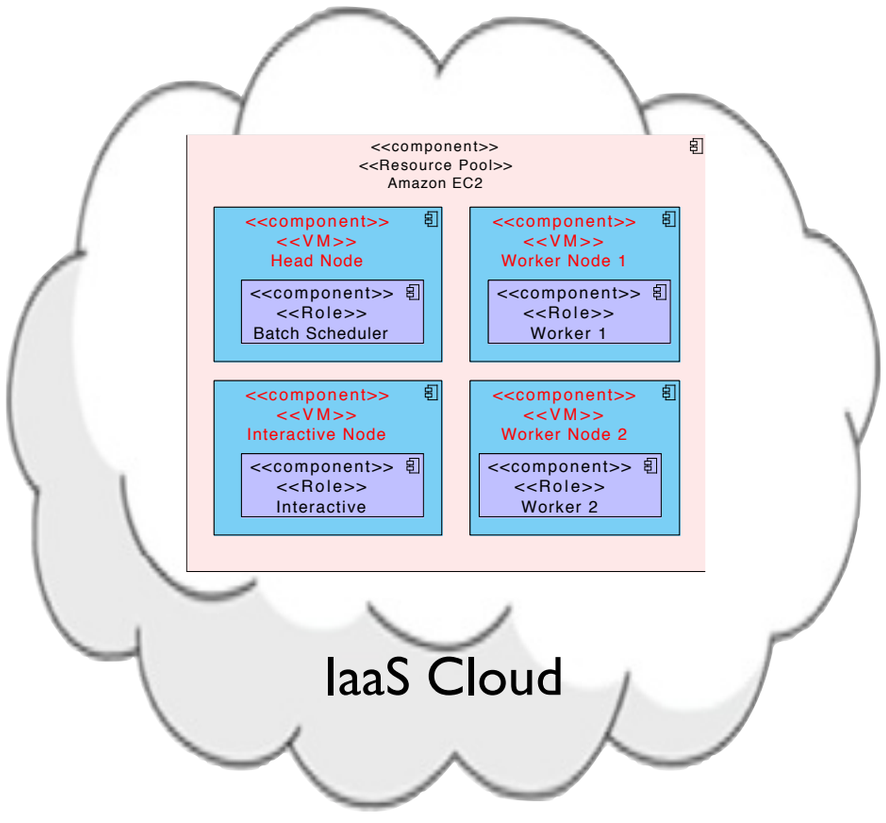
Click the **Deploy** Button

- Allocate Resources
- Configure Services

Including Contexualization

Cluster Users

Submit Jobs, etc



IaaS Cloud

2-min Demo

Why we need VCA

- For Cluster Managers (Users):
  - No need to know the details of how to configure a cluster, just launch it.
- For Cluster Designers (Software Manufacturers):
  - Easier user interface, help promoting the product.
- For Cloud Providers:
  - Easier launching clusters meaning more clusters launched, meaning more revenue.

Not only on IaaS Cloud

- We showed an example on IaaS
- CloudCRV can not only deploy to EC2 like clouds. It also support other resource pools, e.g. libvirt based VMs, raw hardware via gPXE
- To add support for a new "resource pool", simply extend the interface and implement 3 functions.

Heterogeneous cluster

- CloudCRV allows "pre-existed" roles that are not deployed/configured by CloudCRV.
- I.e. if you already have a running HPC cluster, you can easily use CloudCRV to scale it up into the cloud, without changing the existing one.

Check Point

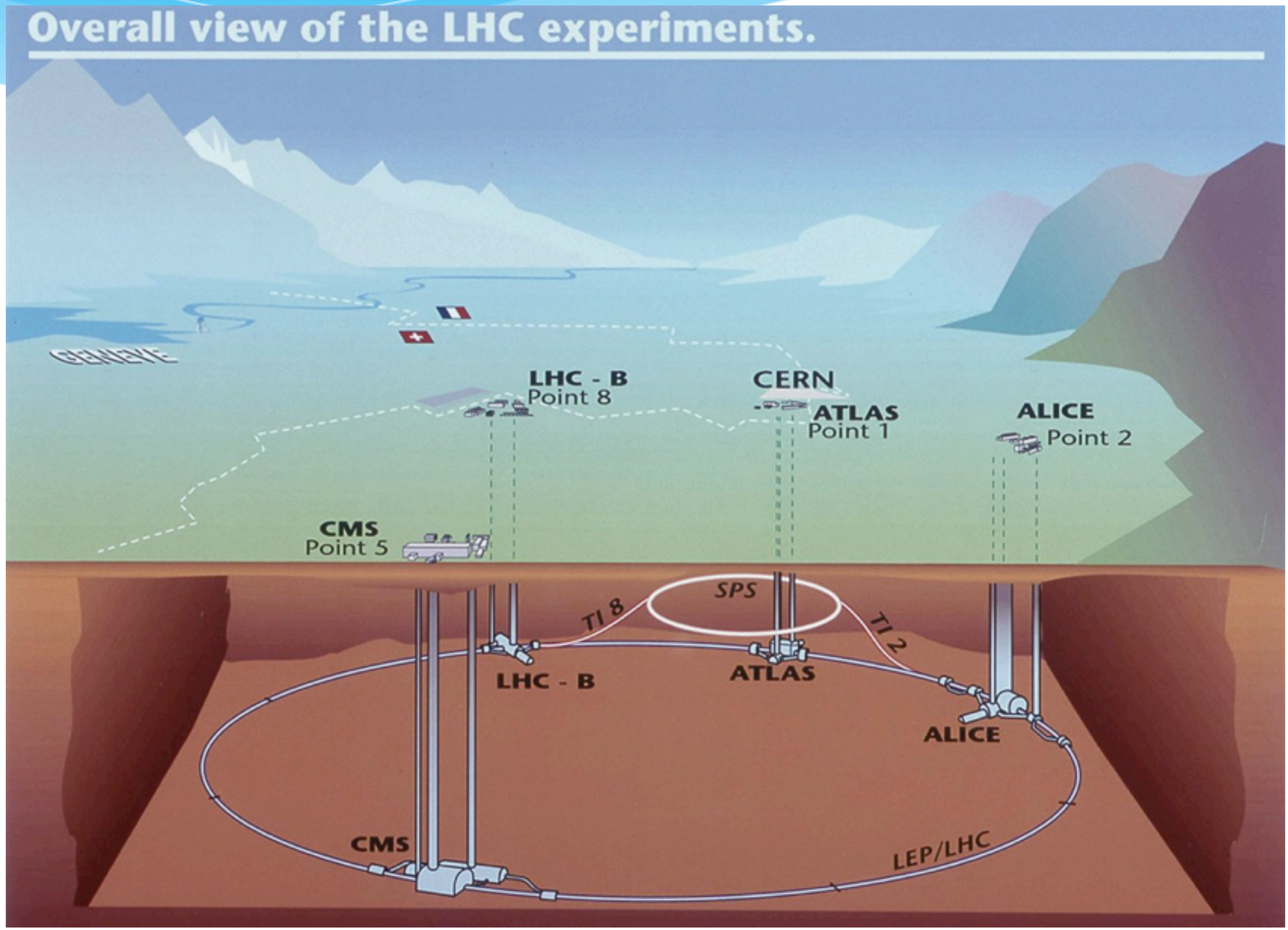
- IaaS Cloud only gives you computers, but not functional clusters.
  - A Virtual Cluster Appliance is an expert-designed cluster that can be easily deployed by non-expert users
- CloudCRV deploys a Virtual Cluster Appliance to IaaS Cloud (and other platforms)
- More Info:
  - Project: http://code.google.com/p/cloudcrv/
  - Poster: http://indico.cern.ch/getFile.py/access?contribId=26&sessionId=5&resId=0&materialId=poster&confId=92498

Want to try it?

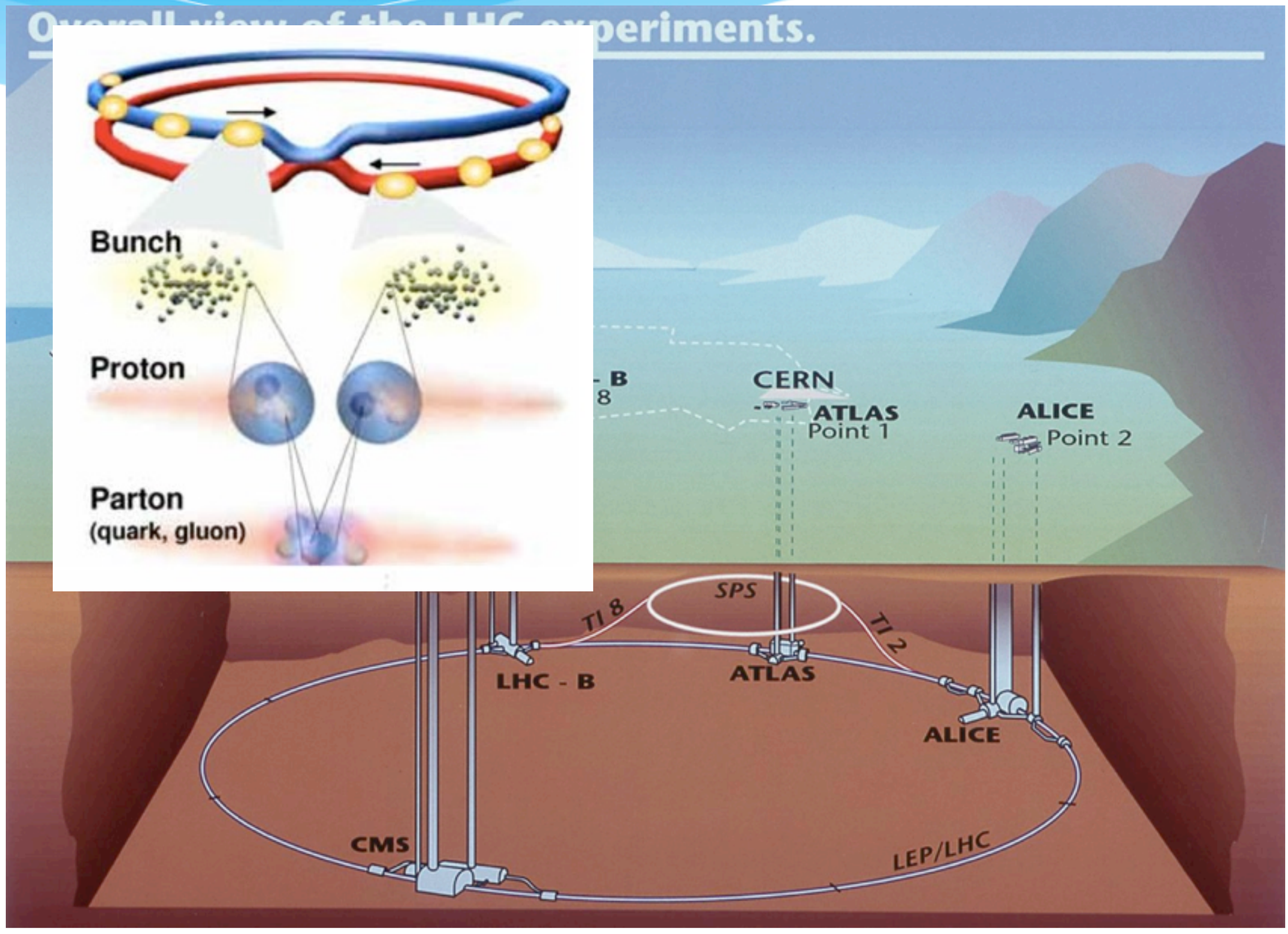
- For early adopters, give us your requirements, we can provide you a VCA to start with.

- yyao@lbl.gov

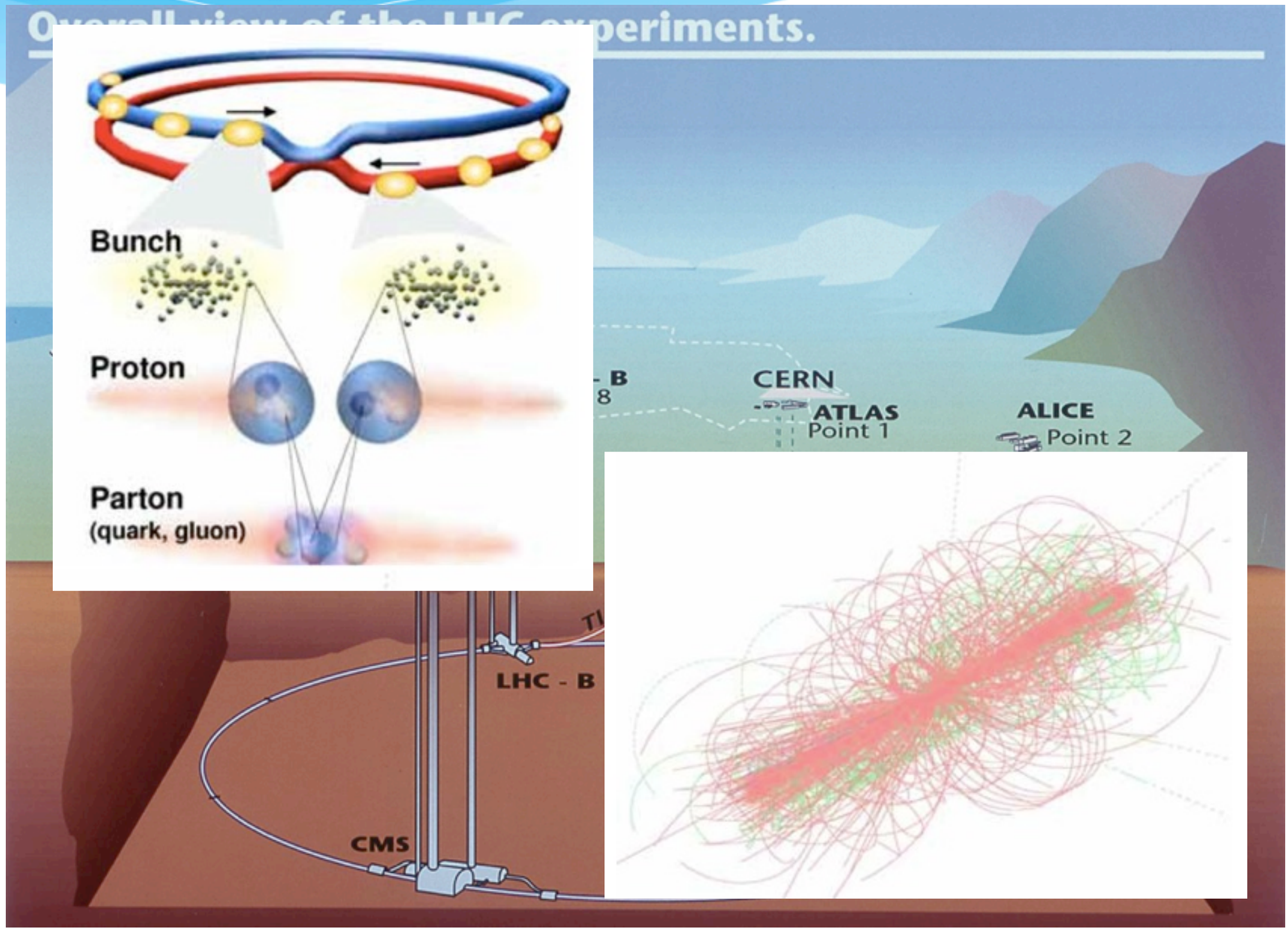
ATLAS & LHC


Overall view of the LHC experiments.
GENEVE
LHC - B
Point 8
CERN
ATLAS
Point 1
ALICE
Point 2
CMS
Point 5
SPS
TI 8
TI 2
LHC - B
ATLAS
ALICE
CMS
LEP/LHC

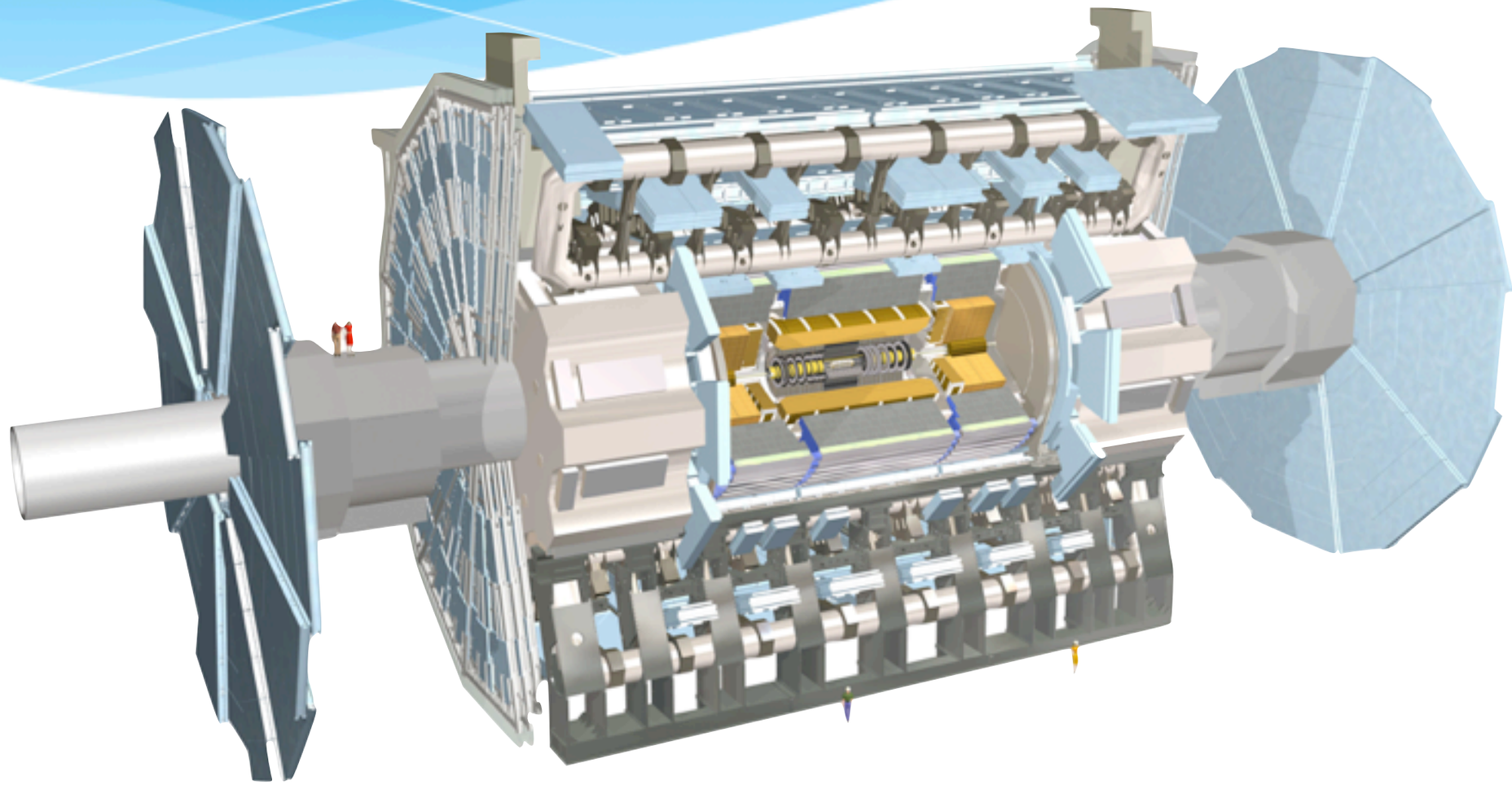
ATLAS & LHC


Overall view of the LHC experiments.
Bunch
Proton
Parton
(quark, gluon)
CERN
ATLAS
Point 1
ALICE
Point 2
SPS
TI 8
TI 2
LHC - B
ATLAS
ALICE
CMS
LEP/LHC

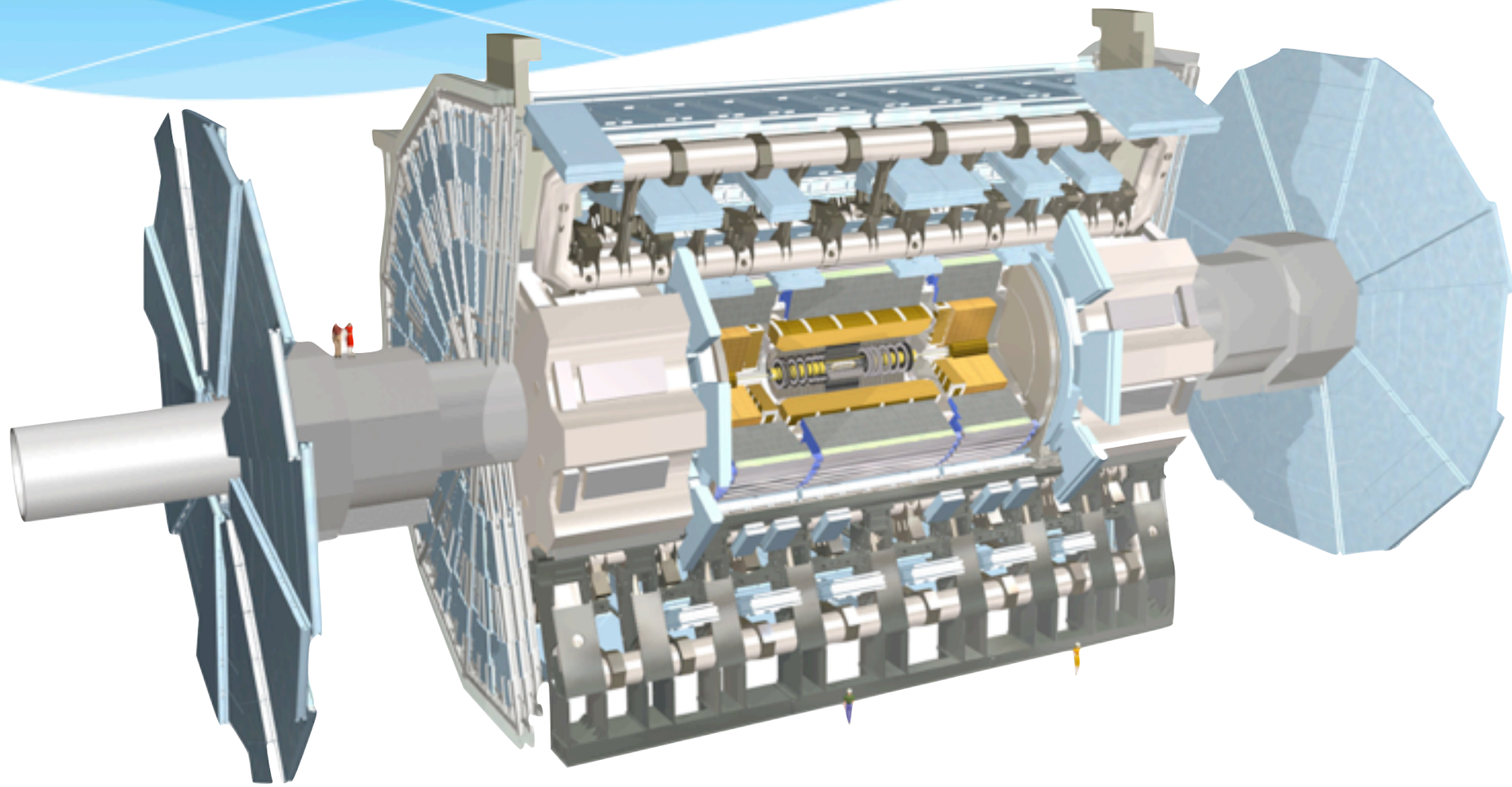
ATLAS & LHC


Overall view of the LHC experiments.
Bunch
Proton
Parton
(quark, gluon)
CERN
ATLAS
Point 1
ALICE
Point 2
LHC - B
CMS

ATLAS Detector

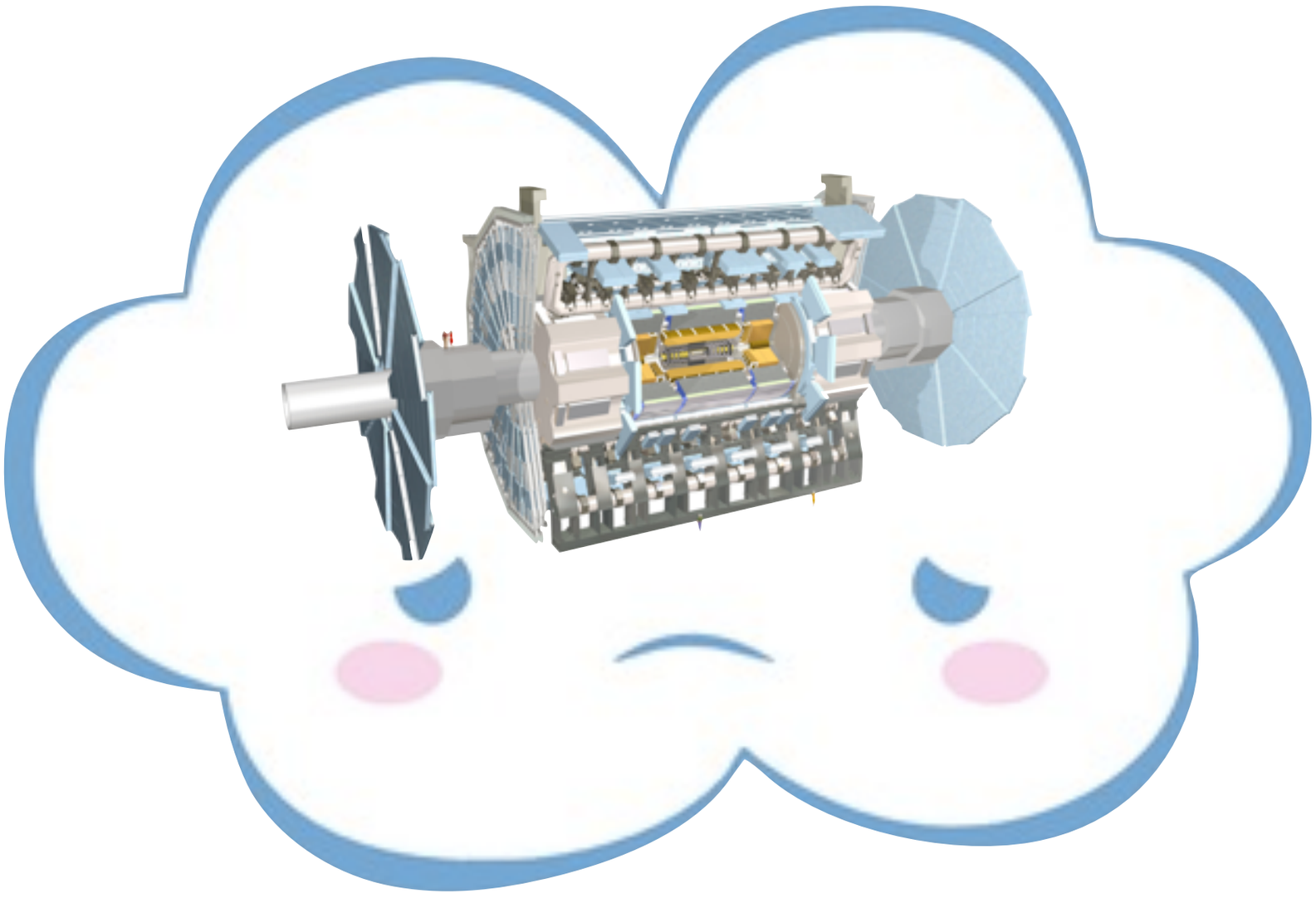


The ATLAS Project



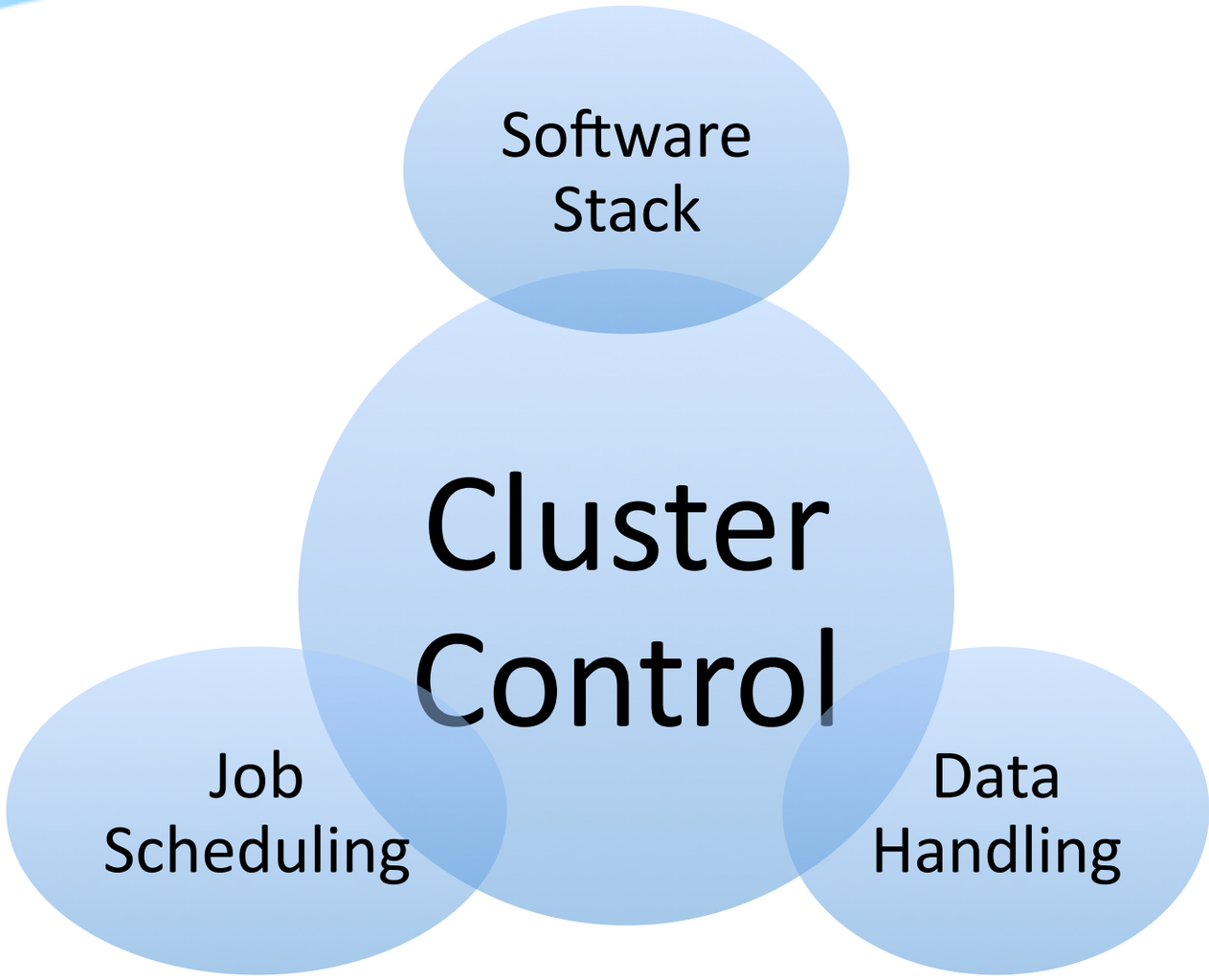
~3000 Scientists around the world (with very different funding situation)
Very Large Software Stack
Huge Computing Needs (Both CPU and Storage)

ATLAS In the Cloud?



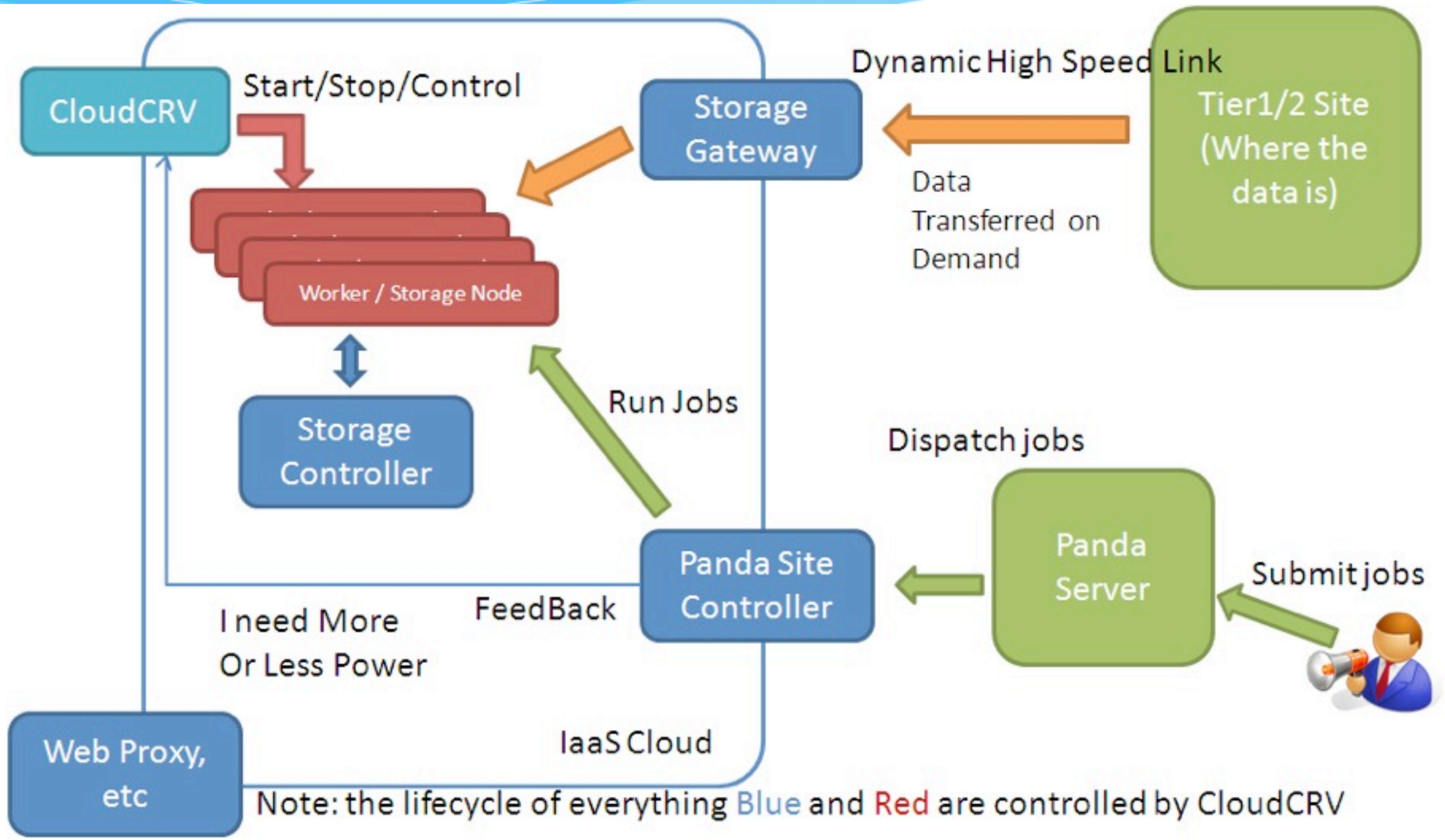
Key Components
Needed for us to use the Cloud

- Software Stack
- Data Handling
- Job Scheduling
- Cluster Control/ Management



We need them to be scalable, efficient, and user friendly

Auto-Scale ATLAS Cluster on Magellan

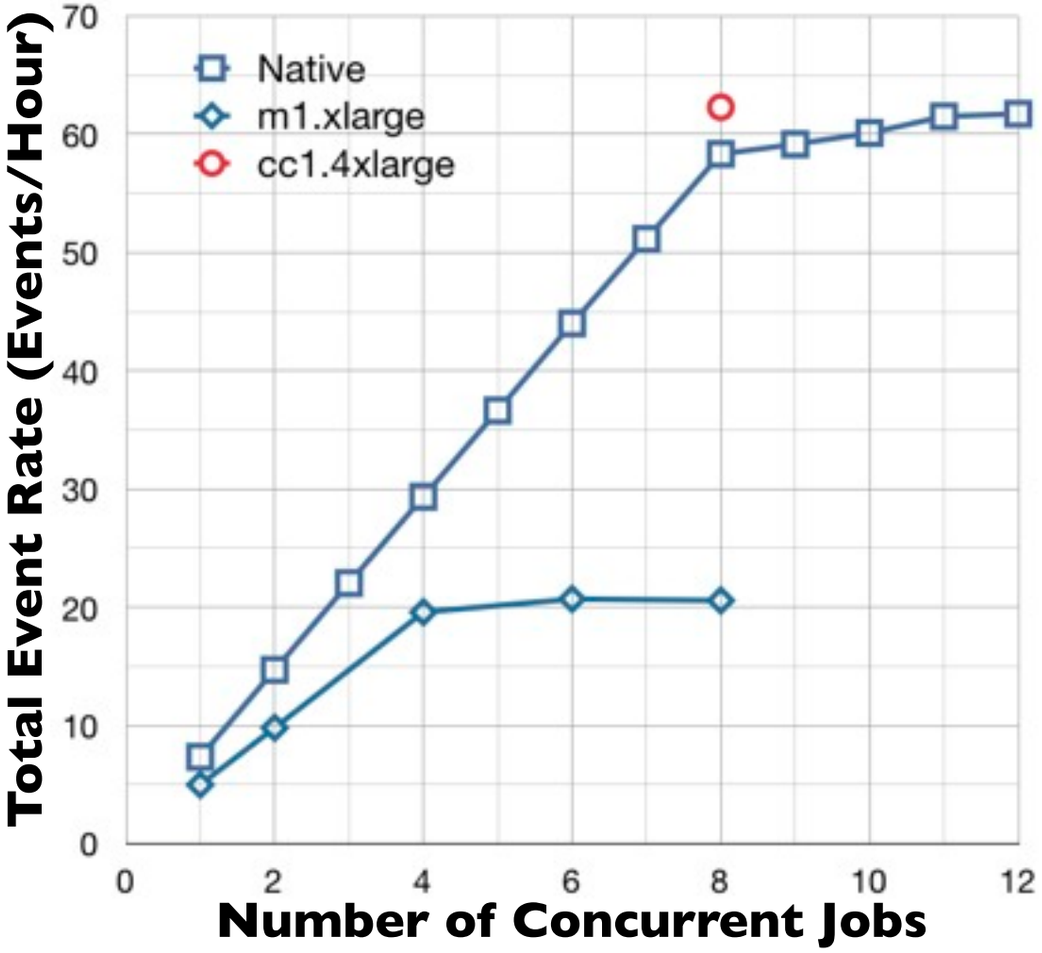


- Scale on-demand
- Storage on Worker (HDFS)
- Panda Based
- High Speed Data Link

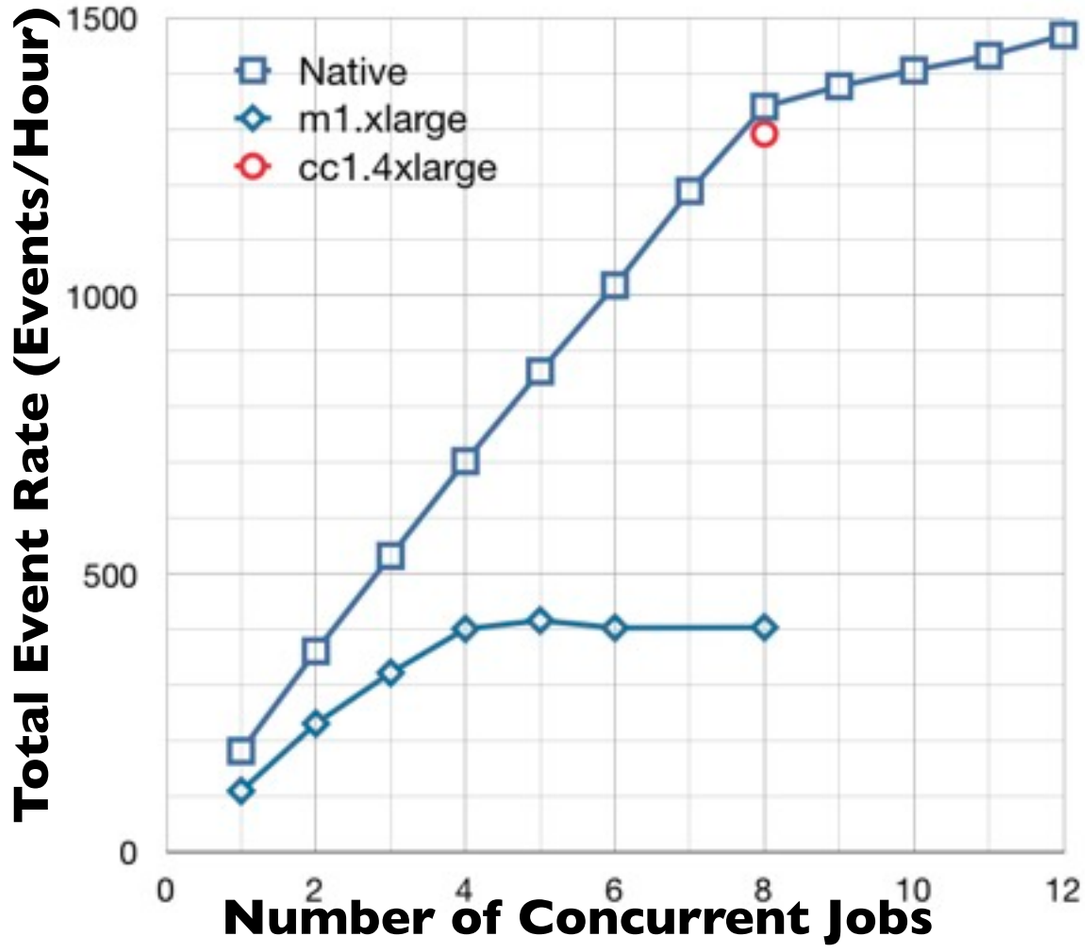
Performance and Cost (on EC2)

Measured Throughput

Simulation



Reconstruction

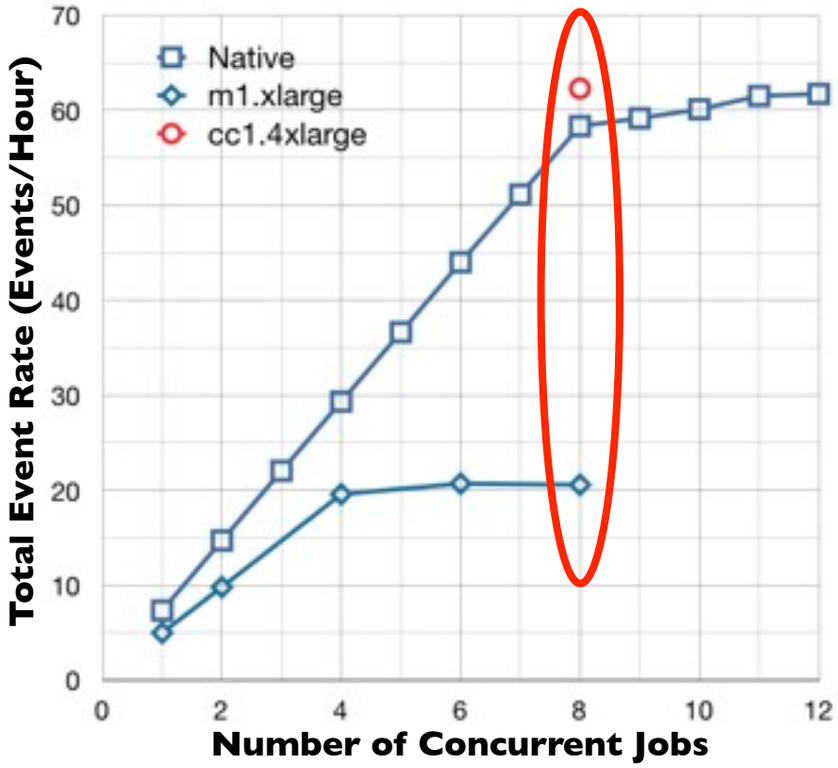


- m1.xlarge is very slow
- cc1.4xlarge is about the same speed as local/native
- Tested 1 Point for cc1.4xlarge, others in progress

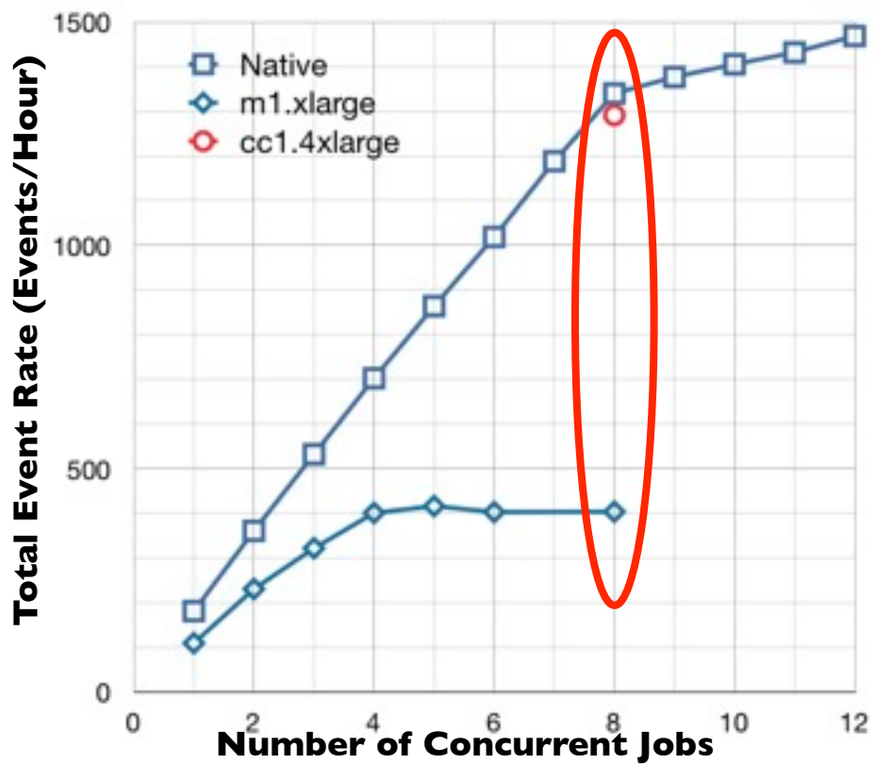
How about cost?

Cost Estimate (Cost per 1K Evt)

Simulation



Reconstruction



Assume we run 8 concurrent jobs for all cases, the cost per 1k event is calculated.

Cost / 1K Event (USD)	EC2		Small ATLAS Center (<400 cores)	Large Center (hundreds K cores)
	m1.xlarge	cc1.4xlarge		
Simulation	37	26	11	5
Reconstruction	1.88+Storage	1.24+Storage	0.48	0.24

EC2 cost doesn't include storage (in/out/store), which is very significant as well!!!!

Cost?

Cost / 1K Event (USD)	EC2		Tier3 Size Center	Large Center (hundreds K cores)
	m1.xlarge	cc1.4xlarge		
Simulation	37	26	11	5
Reconstruction	1.88+Storage	1.24+Storage	0.48	0.24

- Who said Cloud is cheaper?
- EC2 is making too much money!!!
- No Doubt, the IaaS model is very attractive. But to make it cost-effective, we need to be our own cloud-provider. (Or a non-profit cloud provider might also work :-)

Summary

- Still remember what VCA and CloudCRV are?
  - Want to try it? contact me: yyao@lbl.gov
- With the help of CloudCRV, we are building an auto-scaling ATLAS compute cluster on IaaS cloud (e.g. Magellan)
- EC2 is costly for us, it might be cheaper with spot pricing. Best for backup computing power.

BERKELEY LAB

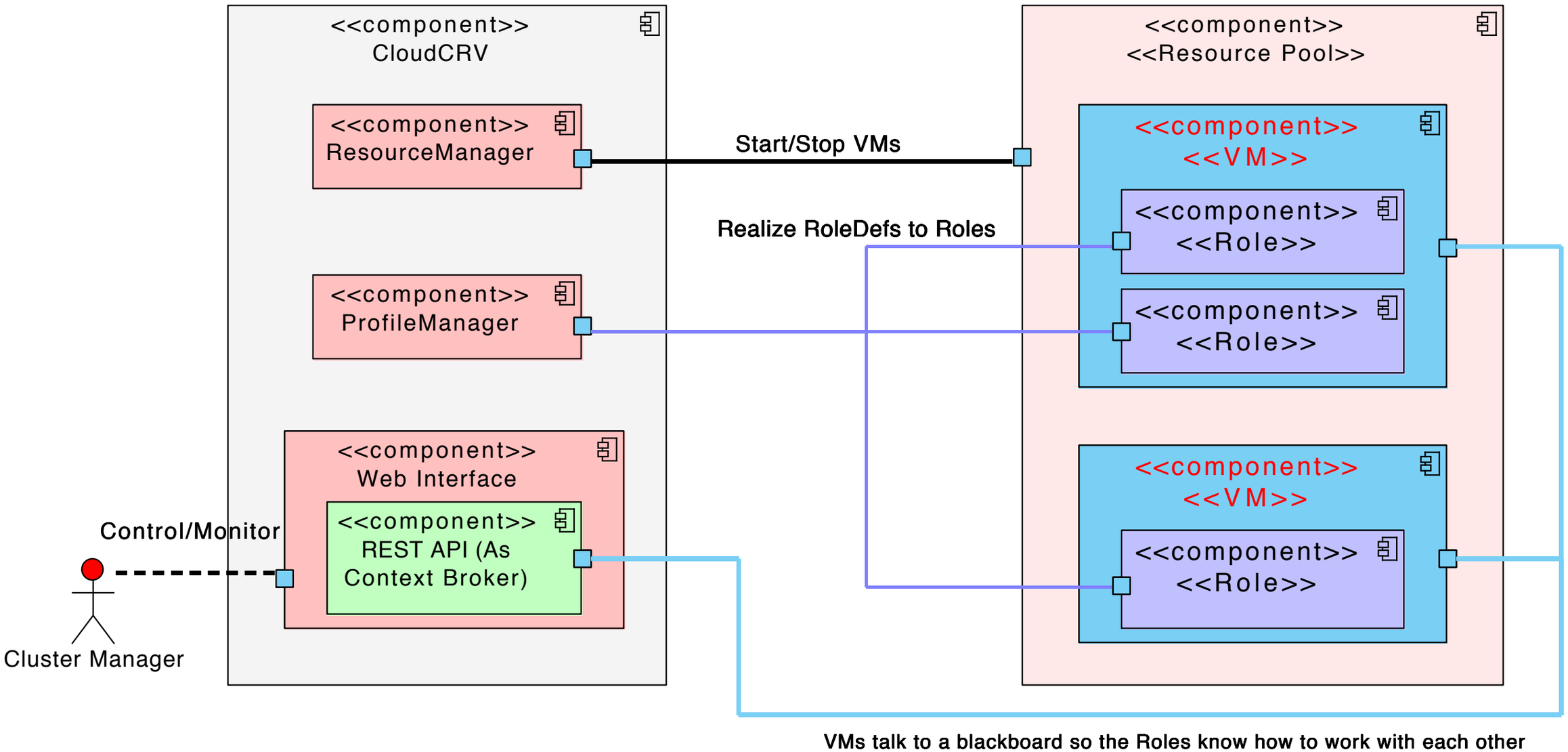
More details in the backup slides.

Questions?

Backup

CloudCRV Details

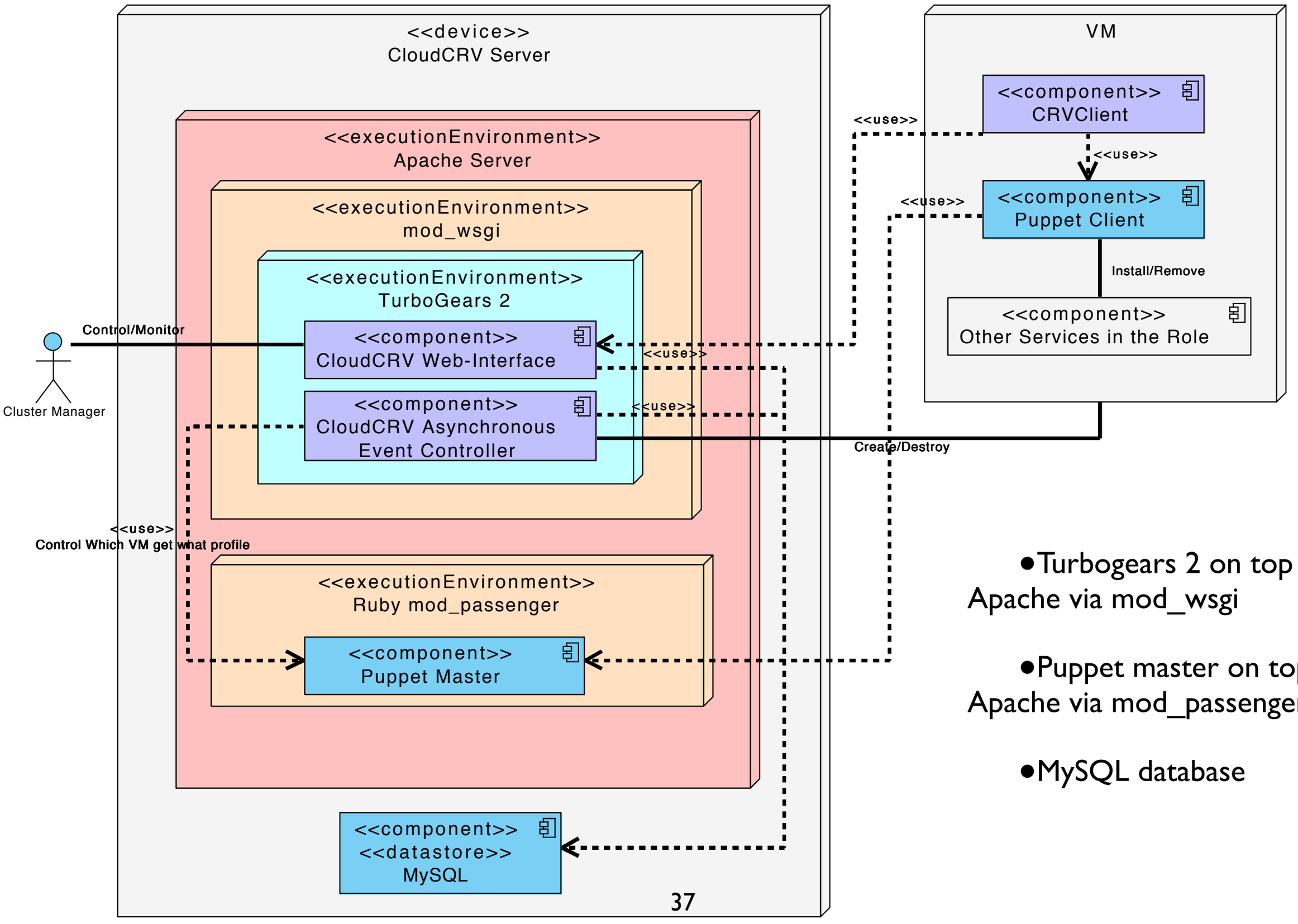
Components of CloudCRV



Roles Interact with each other via broker

- **ResourceManager**: Allocate/Deallocate resource in the resource pools (Turn on/off VMs)
- **ProfileManager**: Realize a RoleDef to a Role that's sitting on a VM
- **Web Interface**:
  - Let the Cluster Manager Control the deployment of the Cluster and Monitor the status
  - Provide a media for VMs (Roles) to talk with each other so that the all VMs/Roles can work together as a cluster

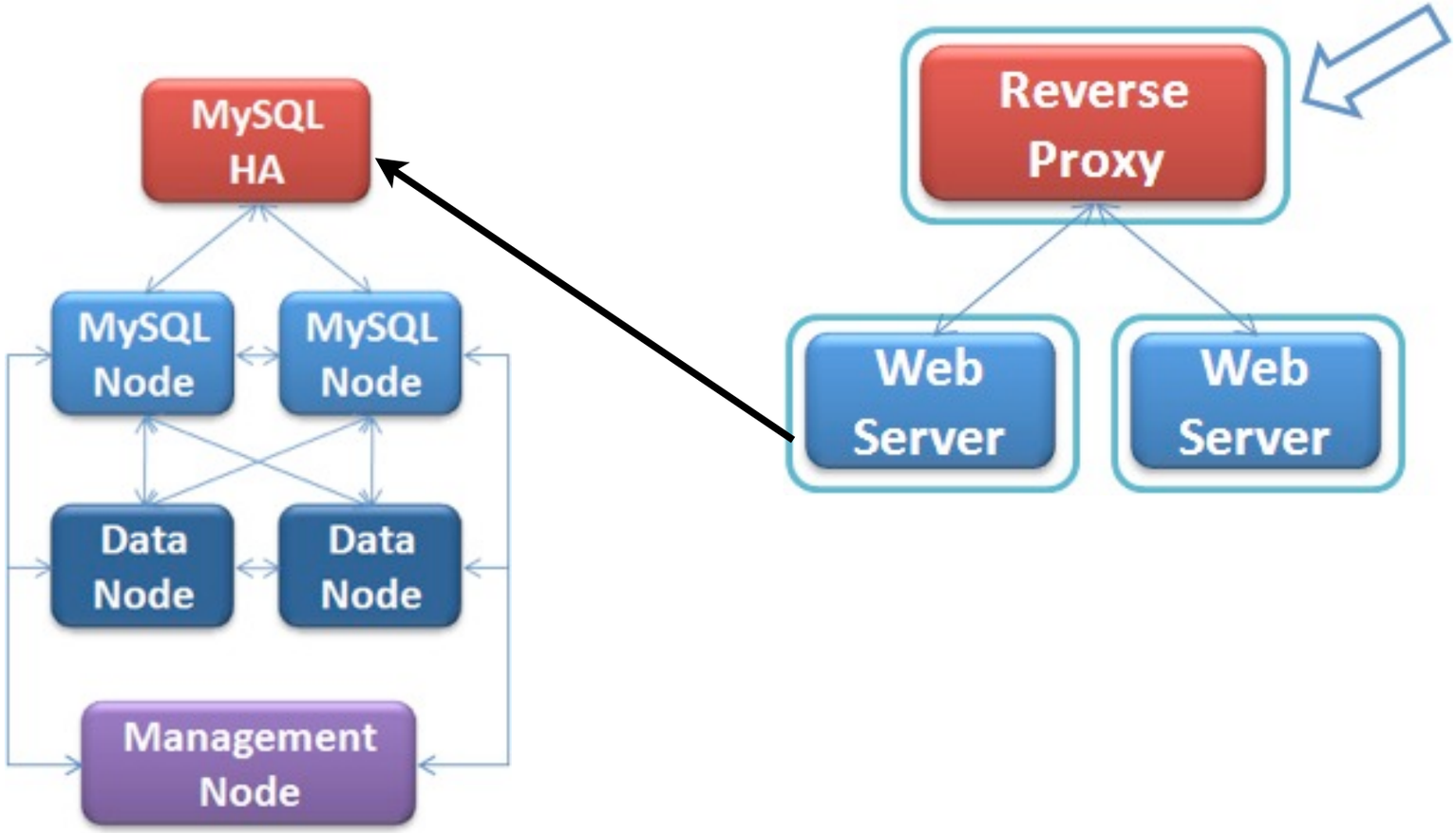
The CloudCRV Server and Clients



- Turbogears 2 on top of Apache via mod_wsgi
- Puppet master on top of Apache via mod_passenger
- MySQL database

Reusability

- VCAs are defined as module-based scripts
  - Designers can easily reuse components in other VCAs or combine several VCA to a new one
  - Much faster development cycle
- E.g. Adding High Availability MySQL to Web Server cluster is simple



Security

- SVN Repository/Scripts can prevent unauthorized use
- CloudCRV manages its own CA, all web service are HTTPS based
  - Manger UI enforce user/pass authentication
  - REST API enforce PKI based authentication - each VM has its own Cert.
- Puppet network traffic is encrypted and authenticated with PKI, using the Certs provided by CloudCRV

How to Design a Cluster

- 1. Know how to setup individual services
- 2. Write a script for each role (preferable Puppet scripts)
- 3. One Python script to define the relations of the roles/vms in a cluster
- 4. Put everything to an SVN repository

Python Part

BERKELEY LAB

```
#Create Resource Pool
magellan_euca=EucaVMMaker("magellan_euca")
cpc=CentOS5_Puppet_ClientMaker(name="RHEL5_Puppet",doc="CRVClient for RHEL5 Clients using
    Puppet",templatedir="/usr/local/cloudcrv/cloudcrv/crvclient")
ppm=Puppet_ProfileMaker(name="Puppet",doc="Profile Maker using Puppet")
rp=RP("publicIP","Resource Pool with Publicly Addressed VMs, Puppet Profiles and RHEL5
Clients",profilemaker=ppm,vmmaker=acs_euca,clientmaker=cpc)

#Role Definitions
rdGangliaSrv=RoleDef('GangliaSrv','Ganglia Server (MetaDaemon and WebFront)','at3_ganglia_srv',"puppet")
rdGangliaClient=RoleDef('GangliaClient','Ganglia Client (Monitoring Daemon)','at3_ganglia_daemon',"puppet")

#remote dependency
rdGangliaClient.addDependOn(rdGangliaSrv)

rdProxy=RoleDef('Proxy',"Proxy Server","at3_proxy","puppet")

#Local dependency
rdProxy.addDependOn(rdGangliaClient,local=True)

#the cluster
cl=Cluster("tier3")

#roles
rGangliaSrv=cl.addRole(name="rGangliaSrv", roledef=rdGangliaSrv, vm=rp.newVM("vmGangliaSrv"), enabled=True)
rProxy=cl.addRole(name="rProxy", roledef=rdProxy, vm=rpprivate.newVM("vmProxy"), enabled=True)
... ...
```

Define a Role with Puppet Script

Ganglia Server
(gmetad, web-front)

```
define at3_ganglia_srv($webaddr="", $roleid="") {
   #Get some attributes from its dependents
   $clientlist=get_dependent_attrlist($webaddr,$roleid,"GangliaClientAddr")
   #Install necessary RPMs
   package { ["ganglia", "ganglia-web", "ganglia-gmetad","httpd"]:
     ensure => installed,    }
   #Create configuration file
   file { "/etc/gmetad.conf": ... ...  notify => Service["gmetad"],    }
   #Start Service
   service { ["gmetad", "httpd"]:  ensure => running,  enable => true,  ... ...
     require => [Package ["httpd", "ganglia", "ganglia-web", "ganglia-gmetad"],File["/etc/
gmetad.conf"]]   } }
   #Publish its IP so that its dependents know where the server is
   set_role_attr($webaddr,$roleid,"GangliaSrvAddr","$ipaddress")
}
```

Ganglia Client
(gmond)

```
define at3_ganglia_daemon($webaddr="", $roleid="") {
   #Figure out where the Ganglia Server is
   $gangliasrvaddr=get_provider_attr($webaddr,$roleid,"GangliaSrvAddr")
   #Install RPMs
   package { ["ganglia", "ganglia-gmond"]:     ensure => installed  }
   #Config files
   file { "/etc/gmond.conf":... ...      notify => Service["gmond"],   }
   #Start Service
   service { [ "gmond"]: ensure => running, enable => true, ... ...
     require => [Package["ganglia", "ganglia-gmond"], File["/etc/gmond.conf"]] } }
   #Publish its IP so that server can authenticate it
   set_role_attr($webaddr,$roleid,"GangliaClientAddr","$ipaddress")
}
```

Puppet

- Configuration Management system
- Use a language to describe a service:

```
class ssh {
    package { ssh: ensure => installed }
    file { sshd_config:
        name => "/etc/ssh/sshd_config",
        owner => root,
        group => root,
        source => "puppet://server/apps/ssh/sshd_config",
        after => Package[ssh]
    }
    service { sshd:
        ensure => running,
        subscribe => [Package[ssh], File[sshd_config]]
    }
}
```

Demo in Pictures

Login

CloudCRV

Drive on the Cloud

Clusters | Roles | Resource Pools/VMs | Login

Now Viewing: login

Only for logged in users

Login

Username:

Password:

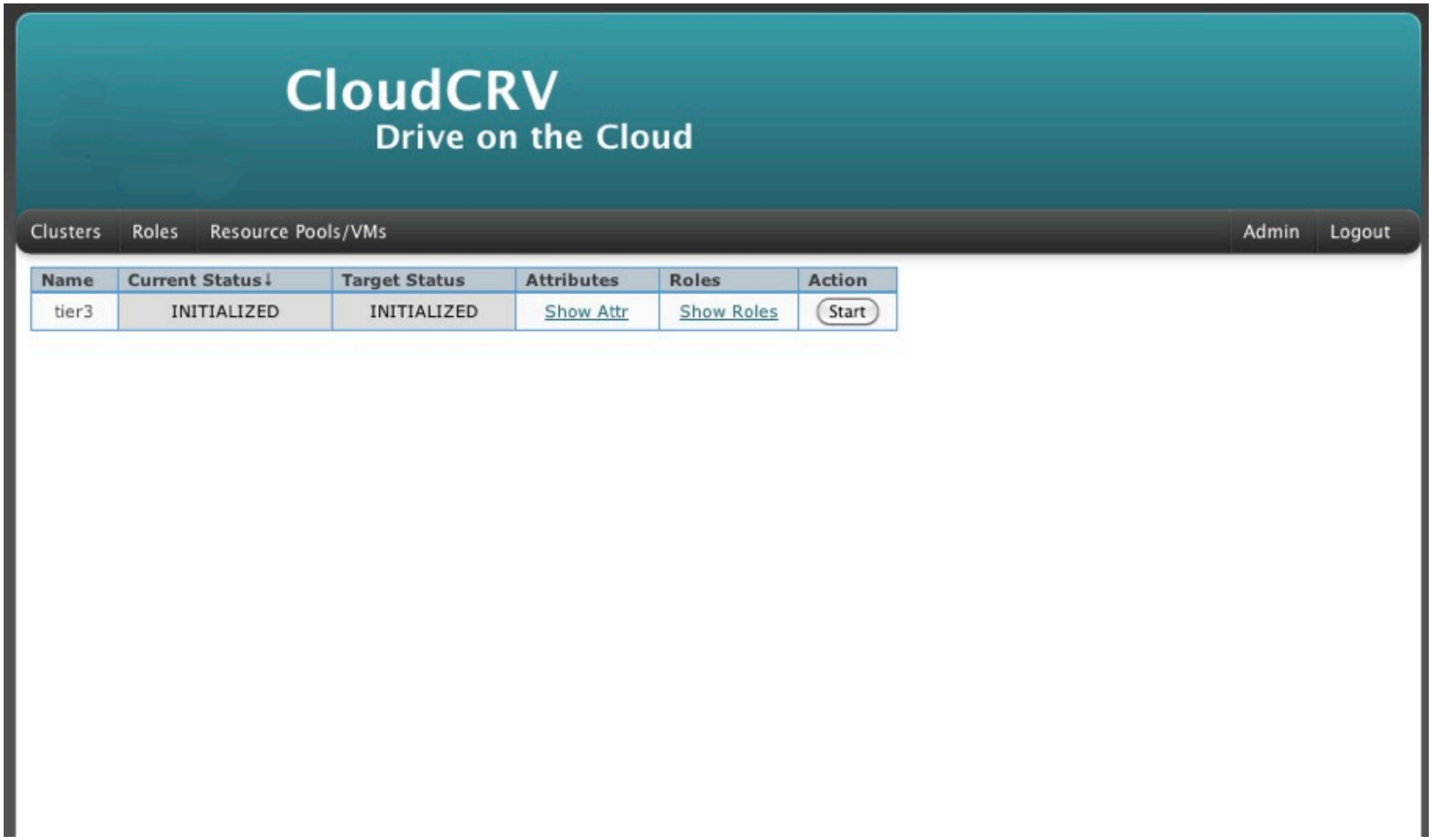
Login

Please refer to https://code.google.com/p/cloudcrv/

TURBOGEARS under the hood

Powered by TurboGears 2

List of Clusters



List of Roles

CloudCRV
Drive on the Cloud

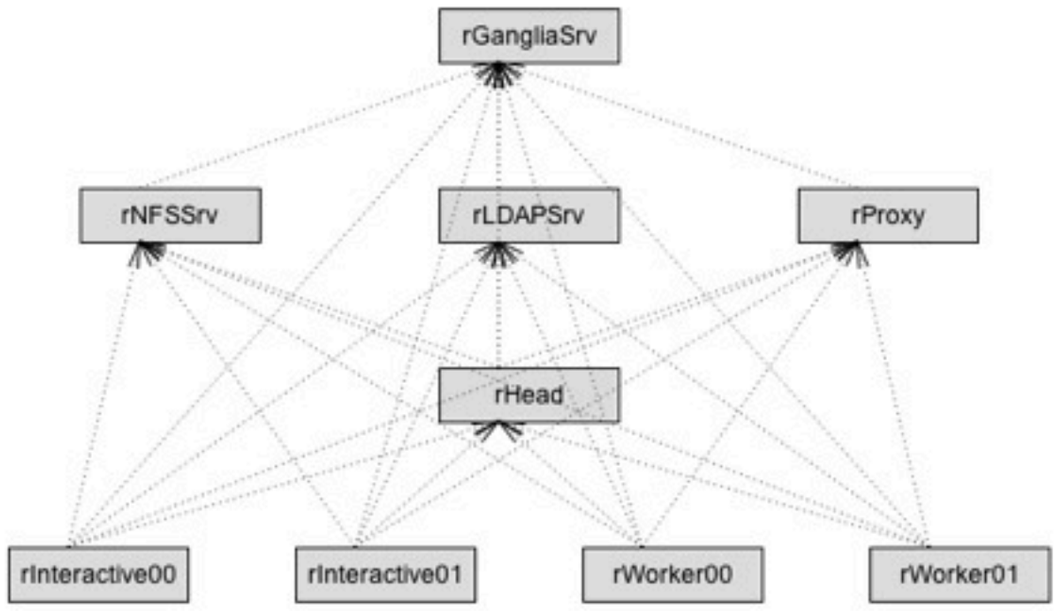
Clusters | Roles | Resource Pools/VMs | Admin | Logout

INITIALIZED
STARTINGVM
APPLYING
RUNNING
REMOVING
STOPPINGVM
FIXING
EXCEPTION

Note:

STOPPED = INITIALIZED

STARTED = RUNNING



Hide Autoroles

VM	RoleID↓	Name	Current Status	Target Status	Attributes	dependOn	Local depOn	dependBy	Local depBy	roleDef
vmGangliaSrv	1	rGangliaSrv	INITIALIZED	INITIALIZED	Show	Show	Show	Show	Show	GangliaSrv
vmNFSSrv	2	rNFSSrv	INITIALIZED	INITIALIZED	Show	Show	Show	Show	Show	NFSSrv
vmNFSSrv	3	autorole_rNFSSrv_GangliaClient	INITIALIZED	INITIALIZED	Show	Show	Show	Show	Show	GangliaClient
vmLDAPSrv	4	rLDAPSrv	INITIALIZED	INITIALIZED	Show	Show	Show	Show	Show	LDAPSrv
vmLDAPSrv	5	autorole_rLDAPSrv_GangliaClient	INITIALIZED	INITIALIZED	Show	Show	Show	Show	Show	GangliaClient
vmProxy	6	rProxy	INITIALIZED	INITIALIZED	Show	Show	Show	Show	Show	Proxy
vmProxy	7	autorole_rProxy_GangliaClient	INITIALIZED	INITIALIZED	Show	Show	Show	Show	Show	GangliaClient

List of ResourcePool/VM

CloudCRV

Drive on the Cloud

Clusters | Roles | Resource Pools/VMs | Admin | Logout

List of Resource Pools and VMs

Resource Pool #1: publicIP

Resource Pool with Publicly Addressed VMs, Puppet Profiles and RHEL5 Clients

› Show Details

List of VMs

VM_ID	Name↓	Current Status	Target Status	Identifier	PublicIP	PrivateIP	Attributes	Roles
1	vmGangliaSrv	INITIALIZED	INITIALIZED				Show	Show
6	vmInteractive00	INITIALIZED	INITIALIZED				Show	Show
7	vmInteractive01	INITIALIZED	INITIALIZED				Show	Show

Resource Pool #2: privateIP

Resource Pool with Privatly Addressed VMs, Puppet Profiles and RHEL5 Clients

› Show Details

List of VMs

VM_ID	Name↓	Current Status	Target Status	Identifier	PublicIP	PrivateIP	Attributes	Roles
5	vmHead	INITIALIZED	INITIALIZED				Show	Show
3	vmLDAPSrv	INITIALIZED	INITIALIZED				Show	Show
2	vmNFSSrv	INITIALIZED	INITIALIZED				Show	Show
4	vmProxy	INITIALIZED	INITIALIZED				Show	Show
8	vmWorker00	INITIALIZED	INITIALIZED				Show	Show
9	vmWorker01	INITIALIZED	INITIALIZED				Show	Show

TURBOGEARS under the hood

Please refer to https://code.google.com/p/cloudcrv/

Powered by TurboGears 2

Click the Start Button

Resources Allocated

CloudCRV

Drive on the Cloud

Clusters | Roles | Resource Pools/VMs | Admin | Logout

List of Resource Pools and VMs

Resource Pool #1: publicIP

Resource Pool with Publicly Addressed VMs, Puppet Profiles and RHEL5 Clients

▸ Show Details

List of VMs

VM_ID	Name	Current Status	Target Status	Identifier↑	PublicIP	PrivateIP	Attributes	Roles
1	vmGangliaSrv	RUNNING	RUNNING	i-510D08F3	131.243.2.18	192.168.2.4	Show	Show
6	vmInteractive00	RUNNING	RUNNING	i-4C6E09C4	131.243.2.26	192.168.2.10	Show	Show
7	vmInteractive01	RUNNING	RUNNING	i-49FF09D8	131.243.2.29	192.168.2.11	Show	Show

Resource Pool #2: privateIP

Resource Pool with Privatly Addressed VMs, Puppet Profiles and RHEL5 Clients

▸ Show Details

List of VMs

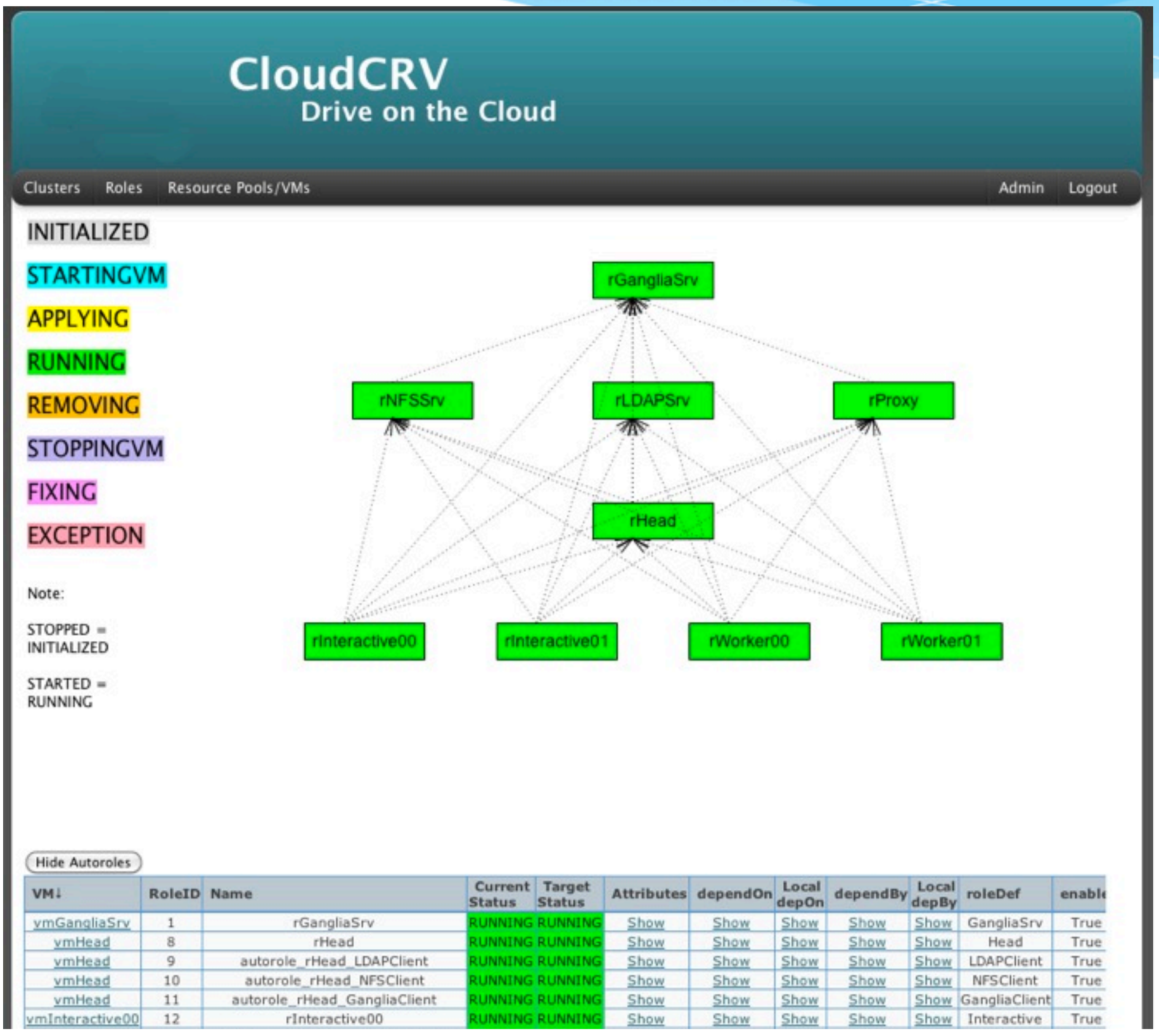
VM_ID	Name	Current Status	Target Status	Identifier↑	PublicIP	PrivateIP	Attributes	Roles
9	vmWorker01	RUNNING	RUNNING	i-579508C1	0.0.0.0	192.168.2.13	Show	Show
8	vmWorker00	RUNNING	RUNNING	i-57770A55	0.0.0.0	192.168.2.12	Show	Show
3	vmLDAPSrv	RUNNING	RUNNING	i-4FAD0946	0.0.0.0	192.168.2.6	Show	Show
2	vmNFSSrv	RUNNING	RUNNING	i-42060899	0.0.0.0	192.168.2.5	Show	Show
4	vmProxy	RUNNING	RUNNING	i-3AB60680	0.0.0.0	192.168.2.7	Show	Show
5	vmHead	RUNNING	RUNNING	i-32D205BE	0.0.0.0	192.168.2.9	Show	Show

TURBOGEARS under the hood

Please refer to https://code.google.com/p/cloudcrv/

Powered by TurboGears 2

Roles Defined



VM↓	RoleID	Name	Current Status	Target Status	Attributes	dependOn	Local depOn	dependBy	Local depBy	roleDef	enable
vmGangliaSrv	1	rGangliaSrv	RUNNING	RUNNING	Show	Show	Show	Show	Show	GangliaSrv	True
vmHead	8	rHead	RUNNING	RUNNING	Show	Show	Show	Show	Show	Head	True
vmHead	9	autorole_rHead_LDAPClient	RUNNING	RUNNING	Show	Show	Show	Show	Show	LDAPClient	True
vmHead	10	autorole_rHead_NFSClient	RUNNING	RUNNING	Show	Show	Show	Show	Show	NFSClient	True
vmHead	11	autorole_rHead_GangliaClient	RUNNING	RUNNING	Show	Show	Show	Show	Show	GangliaClient	True
vmInteractive00	12	rInteractive00	RUNNING	RUNNING	Show	Show	Show	Show	Show	Interactive	True

Cluster Ready

CloudCRV

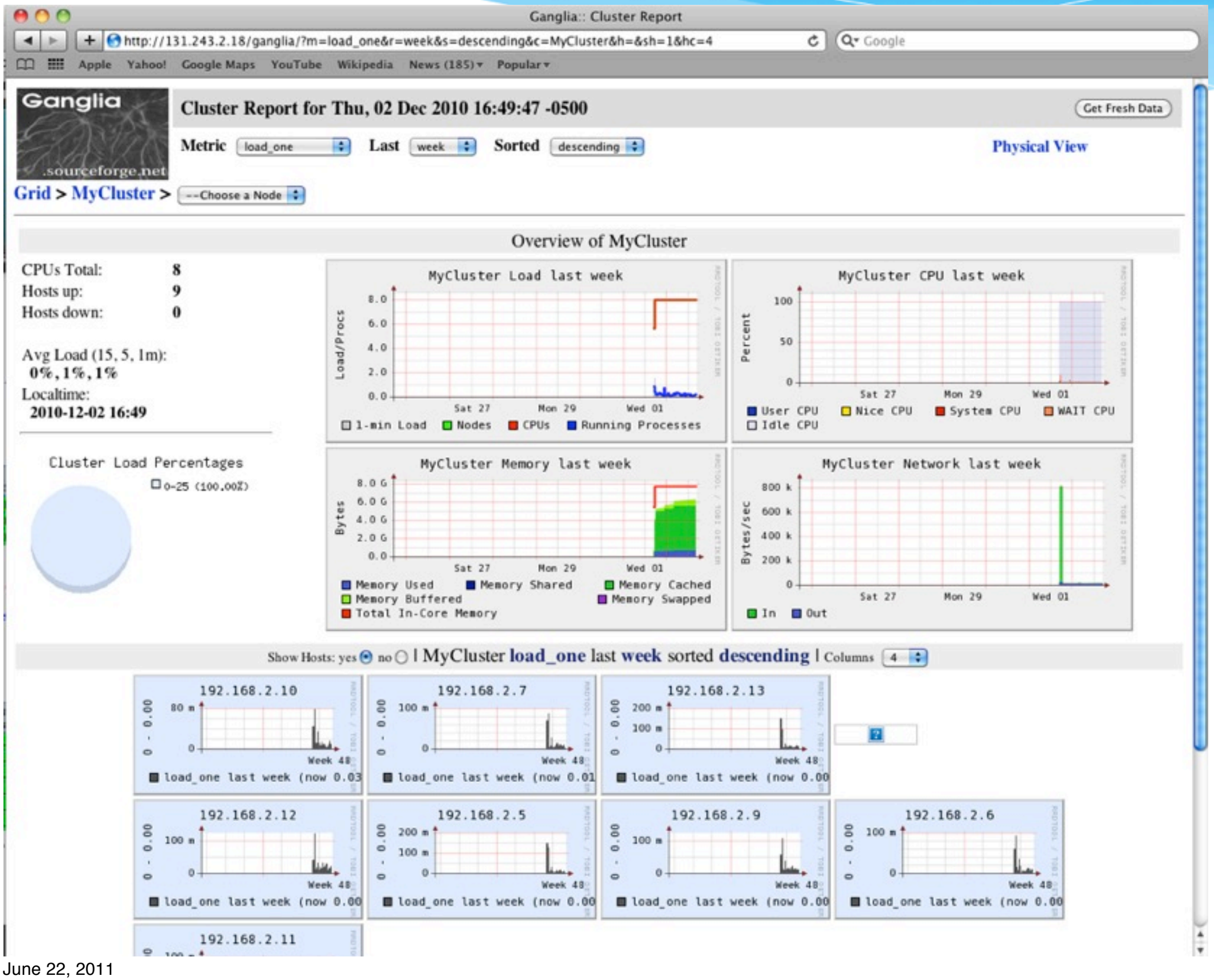
Drive on the Cloud

Clusters | Roles | Resource Pools/VMs | Admin | Logout

Name	Current Status↓	Target Status	Attributes	Roles	Action
tier3	RUNNING	RUNNING	Show Attr	Show Roles	Stop

BERKELEY LAB

Testing The Cluster



ATLAS Cloud Needs

Modes of Operation?

Centralized:

- Like a tier2 in the sky, deployed by one, run jobs for many.
- E.g. the cluster we are building on Magellan (details later in the talk)

De-centralized:

- Deployed (and paid) by one, run jobs for himself (e.g. a univ. prof with a credit card and a paper deadline)

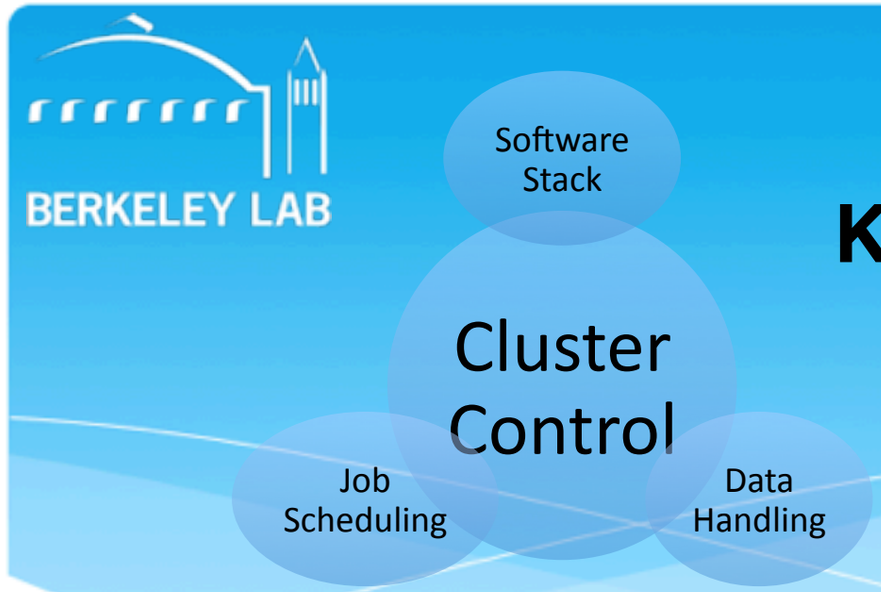
Both modes are possible. Independent of which mode, we face the same problems...

Cloud Problems

What do we need to solve?

- **Agility** (Super Scalable! Isn't it supposed to be a benefit?)
  - Yes, but there's no easy way to use it so far (how to setup the resource to run ATLAS, how to distribute jobs, etc)
- **Data** (Two aspects):
  - Getting data from/to cloud is expensive and inefficient
  - Storing the data in the cloud is tricky.

Several Key Components are need to solve these problems...



Software Stack

Key Components for ATLAS on the Cloud

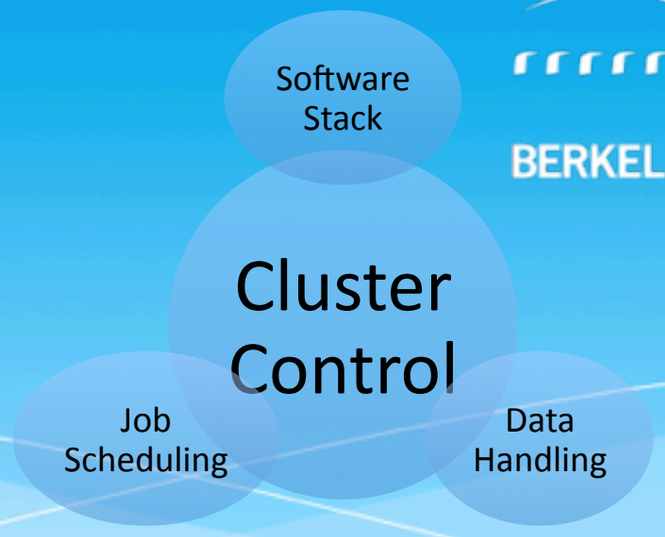
Thanks to CernVM(-FS)

- Web-based read-only FS ideal to distribute many small files to many clients.
- Use CernVM we also get an OS for free
- With proxy servers, it can scale as big as we need
- **Cloud Ready, Great!**

Details: Google CernVM

Data Handling

Key Components for ATLAS on the Cloud

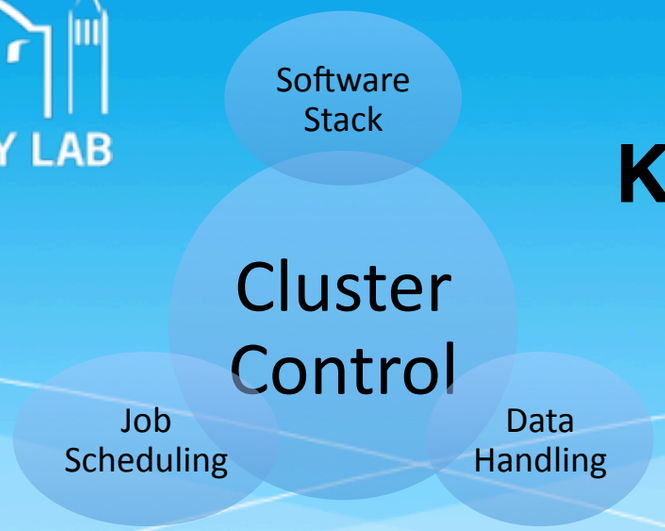


BERKELEY LAB

- **Storage on Worker:**
  - Very Important: we can't dedicate too many storage nodes, that's waste of money (when no worker is running)
- **Smarter Data Transfer**
  - Pre-staging dataset, reuse of data across jobs, etc
- **Possible Solutions:**
  - Mount HDFS across the scalable cluster
    - transparent add/remove node (Agility required)
    - Simplify data staging (1-step staging, no need to move from storage to worker)
  - Xrootd confederation (discover/transfer data better)
  - Reserved links (when possible, reduces transfer time)
  - ... ...

Job Scheduling

Key Components for ATLAS on the Cloud

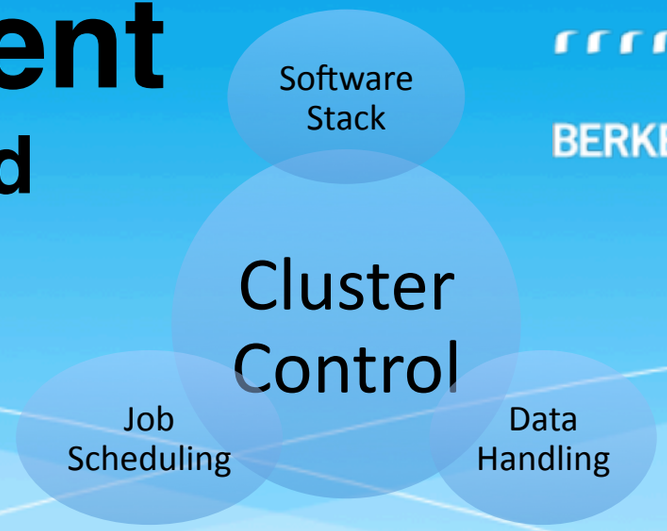


Panda

- Well tested, works well for ATLAS jobs
- Low overhead (management, etc)
- “Data Smart”, sort of
- Schedule whole node jobs with AthenaMP (much easier to handle when trying to take a node offline)

Cluster Control/Management

Key Components for ATLAS on the Cloud



BERKELEY LAB

We need a tool to:

- Allocate cloud resource when needed, release resource when done
- Configure the resources to do ATLAS work
  - tasks like: install CernVM-FS, configure HDFS, setup Panda, etc.
  - note that: each of the above task needs an expert to do
- Any one who need to setup such a cluster should be able to do this with one button click (especially for de-centralized modes)

And we have CloudCRV

Cost Calculation Assumptions

Not for accurate calculation, cost might be different for individual computer center.

- m1.xlarge: **$0.19** per core-hour (Storage excluded)
- c1.xlarge: **$0.20** per core-hour (Storage excluded)
- For US ATLAS Tier3 Center a rough estimate is around $0.05-0.10 per core-hour (including initial hardware and support), **we use the number $0.08** (Storage included)
- Large data centers (hundreds of K cores), $0.02-0.06 per core-hour, **we use the number $0.04** (Storage included)