VC3

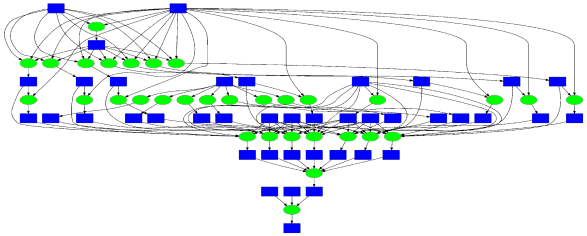
Virtual Clusters for Community Computation

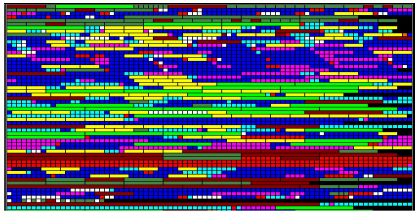
OSG All-Hands Meeting @ University of Utah
March 20, 2018

John Hover <jhover@bnl.gov> for the VC3 Team

University of Chicago University of Notre Dame Brookhaven National Lab

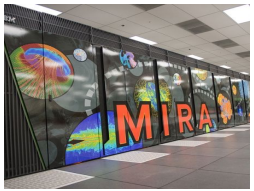
You have a complex scientific workflow/application that runs at one site, perhaps your home university.





Now, you want to migrate that workflow to national-scale infrastructure, and allow you and your collaborators to easily run similar workloads everywhere they have access…

HPCs

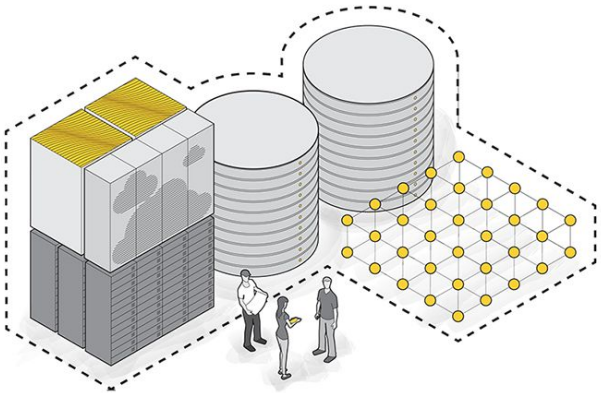


Distributed HTC

Open Science Grid

Campus Clusters

Commercial Clouds



VC3: A platform for provisioning customized clusters over heterogeneous resources for collaborative science teams

What is it?

VC3 aggregates allocation-based resources, dynamically constructing homogeneous virtual clusters (middleware) as a service. Key features:

- **Automated**: Clusters are requested, built, used, and torn down by the system, driven by a user-facing web portal.
- Utilizes **dynamic infrastructure**. Factories and other central services are spawned (and destroyed) as needed. Static components relatively lightweight.
- **Application (middleware) agnostic**: Cluster middleware can be HTCondor, WorkQueue. Extensible, e.g. Apache Spark, Kubernetes. (Jupyter?)
- **VC3 Builder:** Satisfies all dependencies specified in cluster definition, *as needed*.
- **User driven**: Oriented toward aggregating individual or small group allocations, e.g. campus clusters, academic clouds, university HPCs for **federated teams.**

What it's not

- VC3 is **not** a workload management system. It doesn't run **jobs**, it provisions a customized cluster with your chosen middleware and execution environment.
- It isn't oriented toward creating large scale, shared global clusters.
- Clusters are short-lived, for individuals or small groups, purpose-built for a workflow/task.
- Doesn't currently handle data as part of automation.
- Not a software product to be deployed by a VO; it will be a service. But all the code is open, packages and dependencies are published, so in theory someone *could.*
- Not developed from scratch. Integrates existing technologies and combines them into a fully automated, user-oriented service.

VC3 Concepts

Resource – A uniform site (campus cluster, HPC center, Cloud)

Allocation – A user account at a Resource that consumes some type of compute unit – (usually billed as Service Units, e.g., HPC centers, dollars (e.g., AWS, GCE), or priority (e.g., HTCondor).

Project – A self-identified set of users and allocations. All users in a project can use any allocation in the project, launching virtual clusters. This is where sharing of resources/allocations occurs.

Middleware – The cluster fabric. HTCondor, WorkQueue, Spark…

VC3 Concepts…

Nodeset – A set of equivalent nodes (e.g. workers, batch scheduler, login host)

Cluster Template – A description of the components (NodeSets) of a virtual cluster, including middleware, number and types of login/head/worker nodes.

Environment – A set of software/execution environment to be provided on a node (e.g. worker or login host) by the builder.

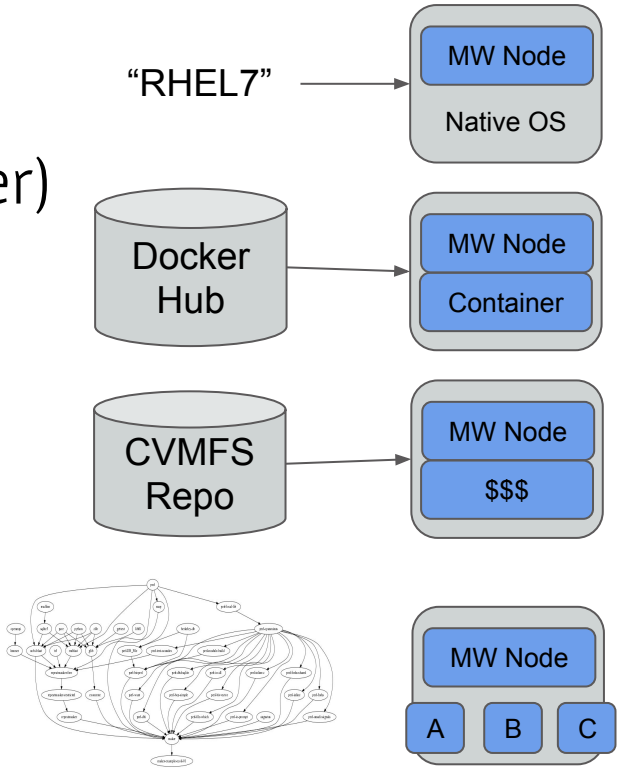
VC3 Concepts...

Virtual Cluster (VC) – A cluster running on behalf of a project on selected resources, using selected allocations, as defined by a cluster template and selected environment. I.e.

VC = Project + Cluster Template + Environment

Environments: More than just software

- Native Availability
  - Specify desired OS from list.
- Containers (Docker/Singularity/Shifter)
  - Specify image to pull from Docker Hub.
- On-Demand Deployment (CVMFS)
  - Specify CVMFS repo, system mounts it.
  - CVMFS via FUSE (kernel) or Parrot (user)
- Build on Site
  - Specify list of software packages needed.
  - VC3-Builder downloads and installs.

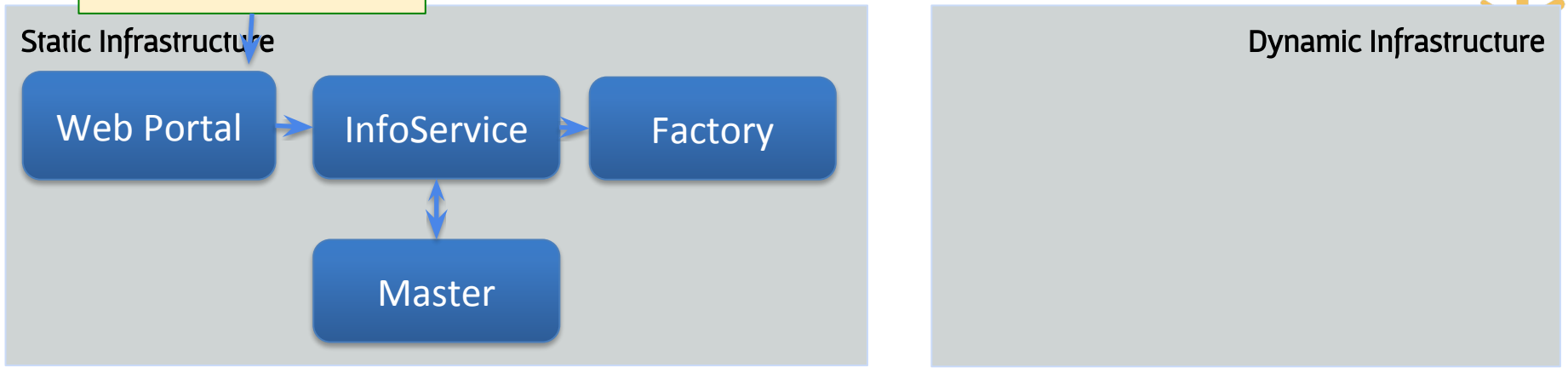


Representative Use Case (summary)

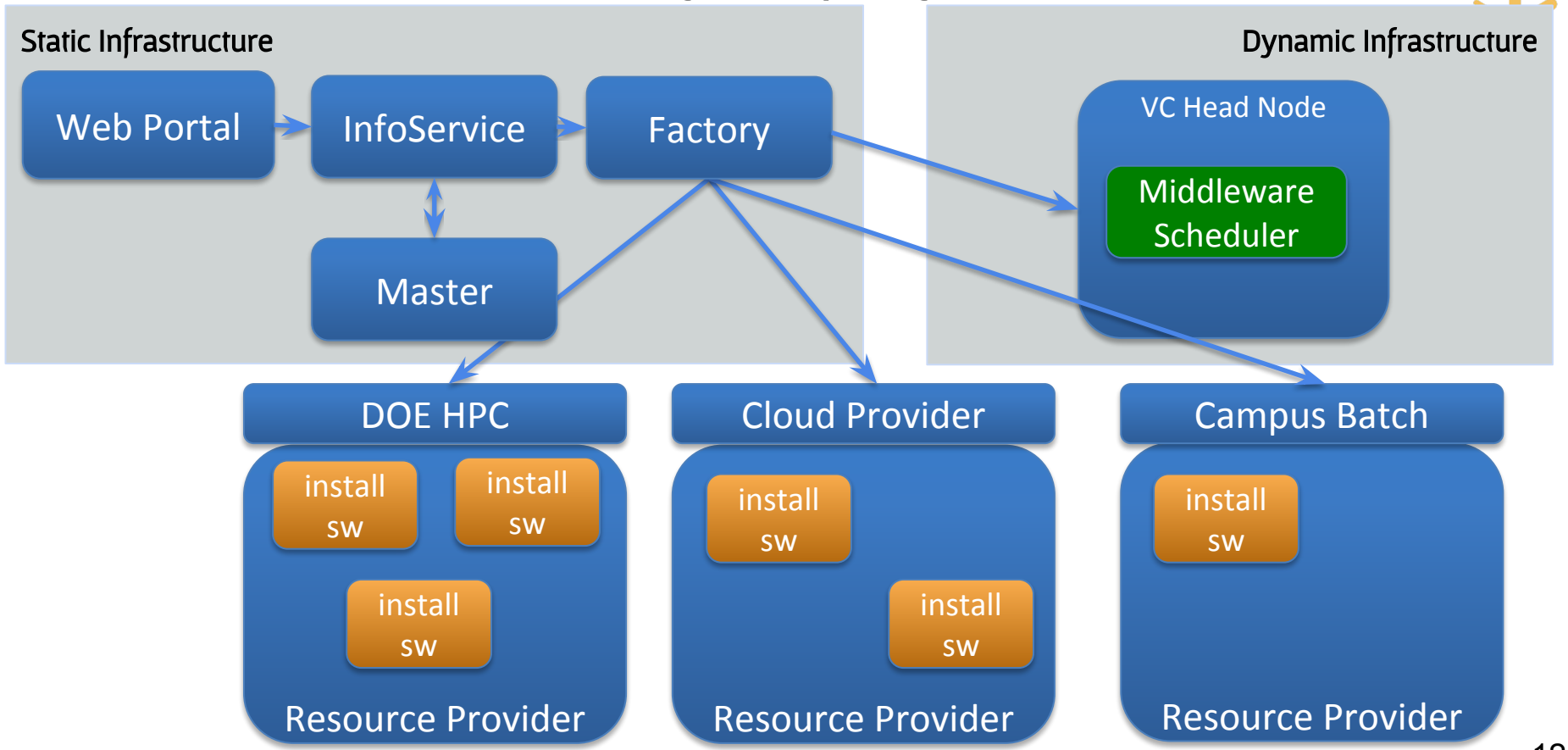
For a Virtual Cluster request the VC3 system…

- Sorts out which Project resources *can* service the request.
- Static factory launches/configures a provisioning factory and central middleware infrastructure (e.g. the vc3-factory + HTCondor central manager/schedd on login host.)
- The factory submits vc3-builders to target resources, re-submitting as needed.
- vc3-builder, for each worker, satisfies all dependencies *however needed on that node.* e.g,
  - If GCC 4.3 is not present, installs it.
  - If fuse not present for CVMFS, sets up Parrot.
- User then either loads workflow and data on managed infrastructure, or triggers remote submission into it.
- User triggers cluster teardown when done, (staging out data if not handled out of band).
- When all workers are gone, central infrastructure shut down. Request completed.

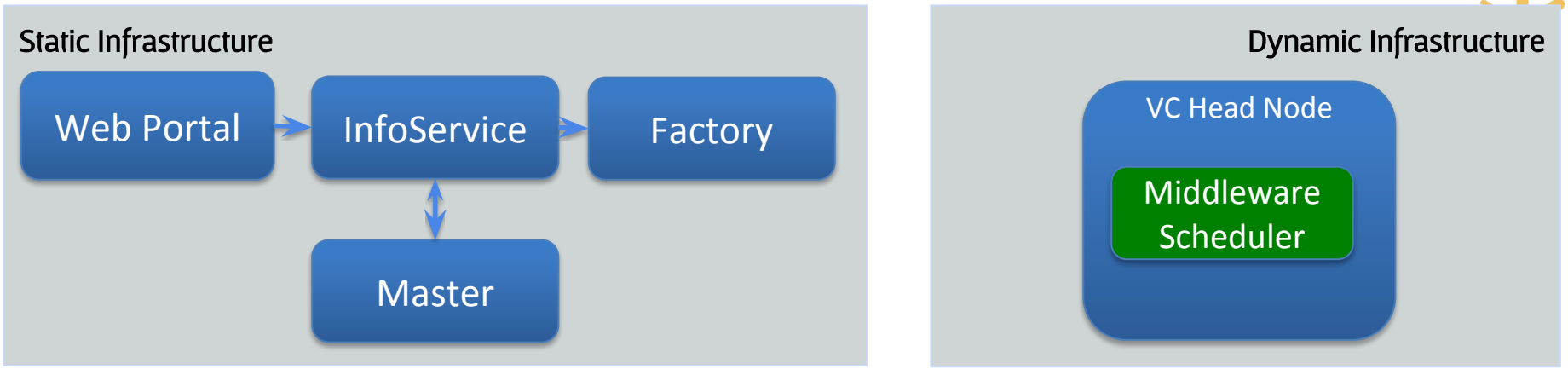
Initialization

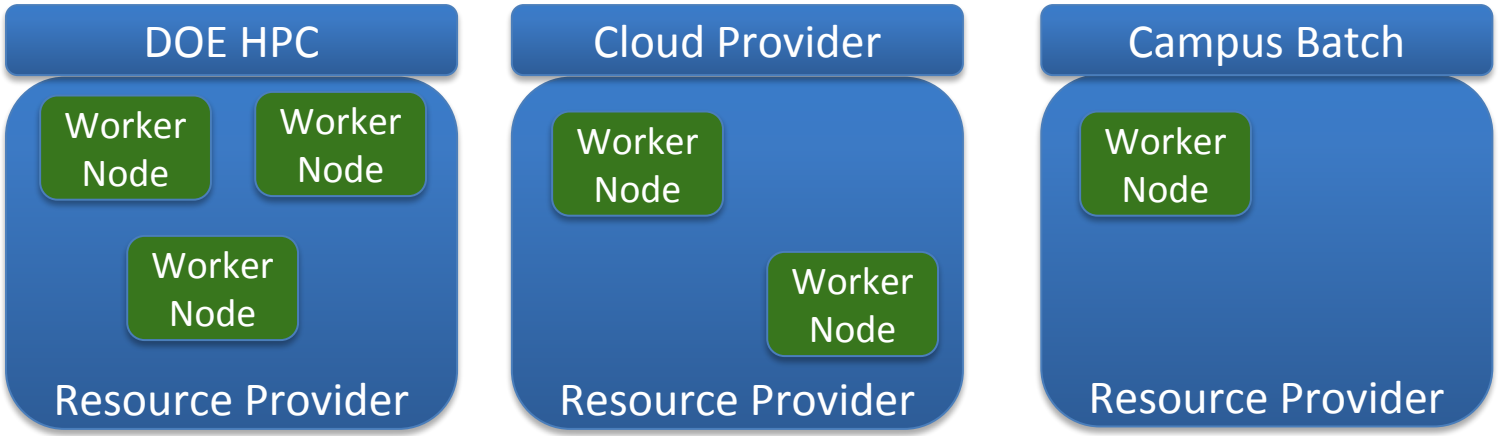


Overlay deployment

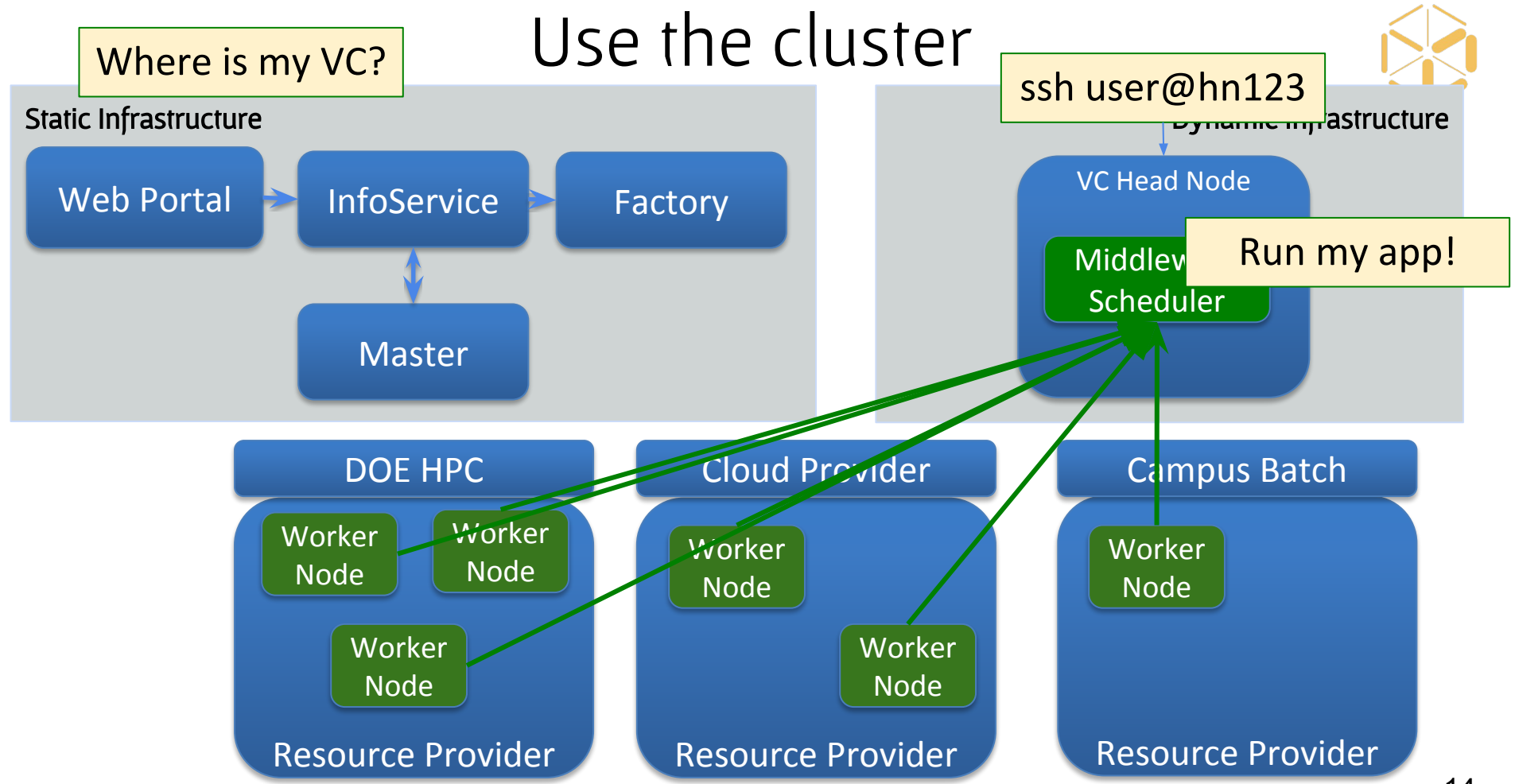


VC3

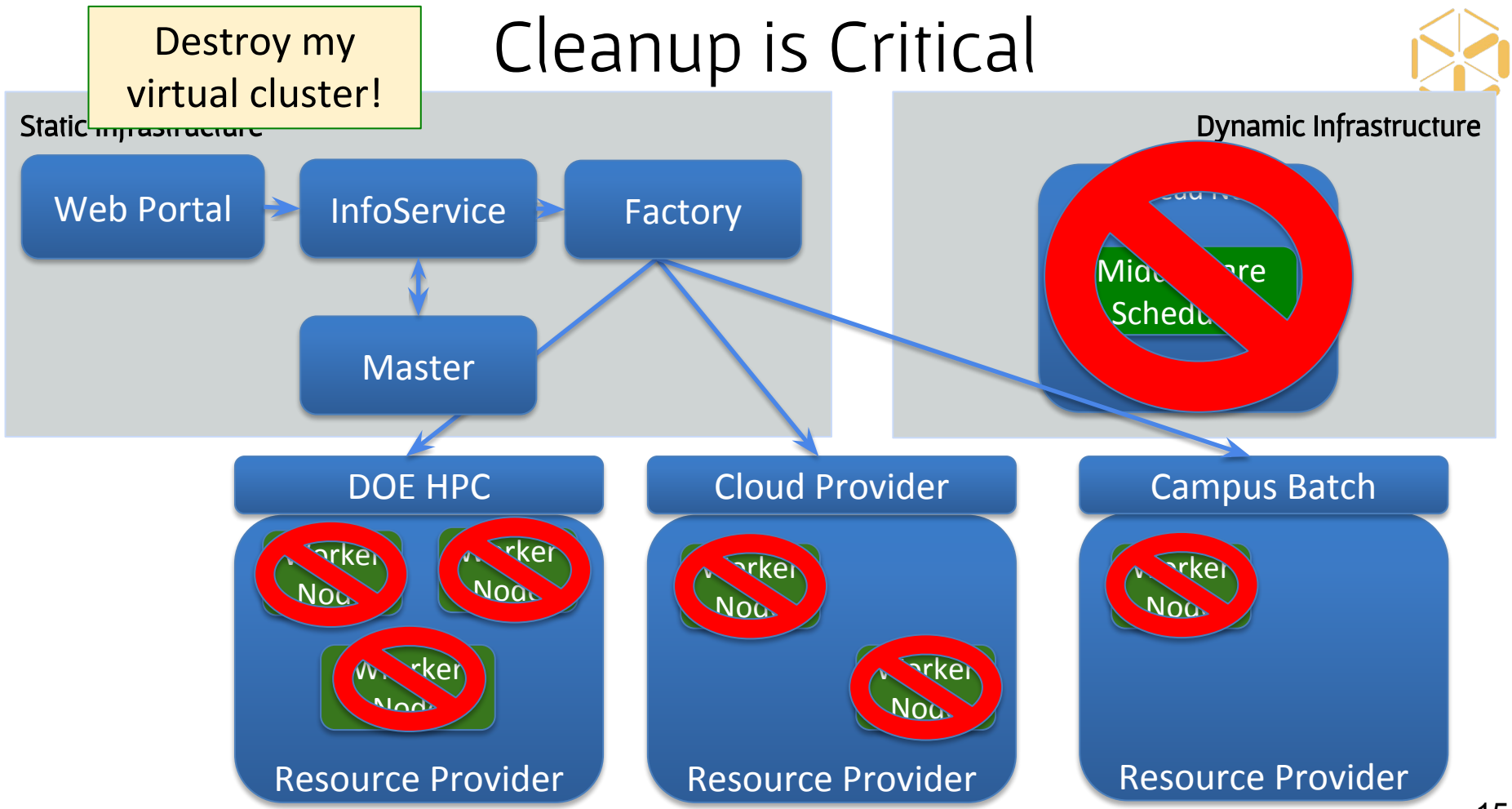

Static Infrastructure
Web Portal
InfoService
Factory
Master
Dynamic Infrastructure
VC Head Node
Middleware Scheduler


DOE HPC
Worker Node
Worker Node
Worker Node
Resource Provider
Cloud Provider
Worker Node
Worker Node
Resource Provider
Campus Batch
Worker Node
Resource Provider

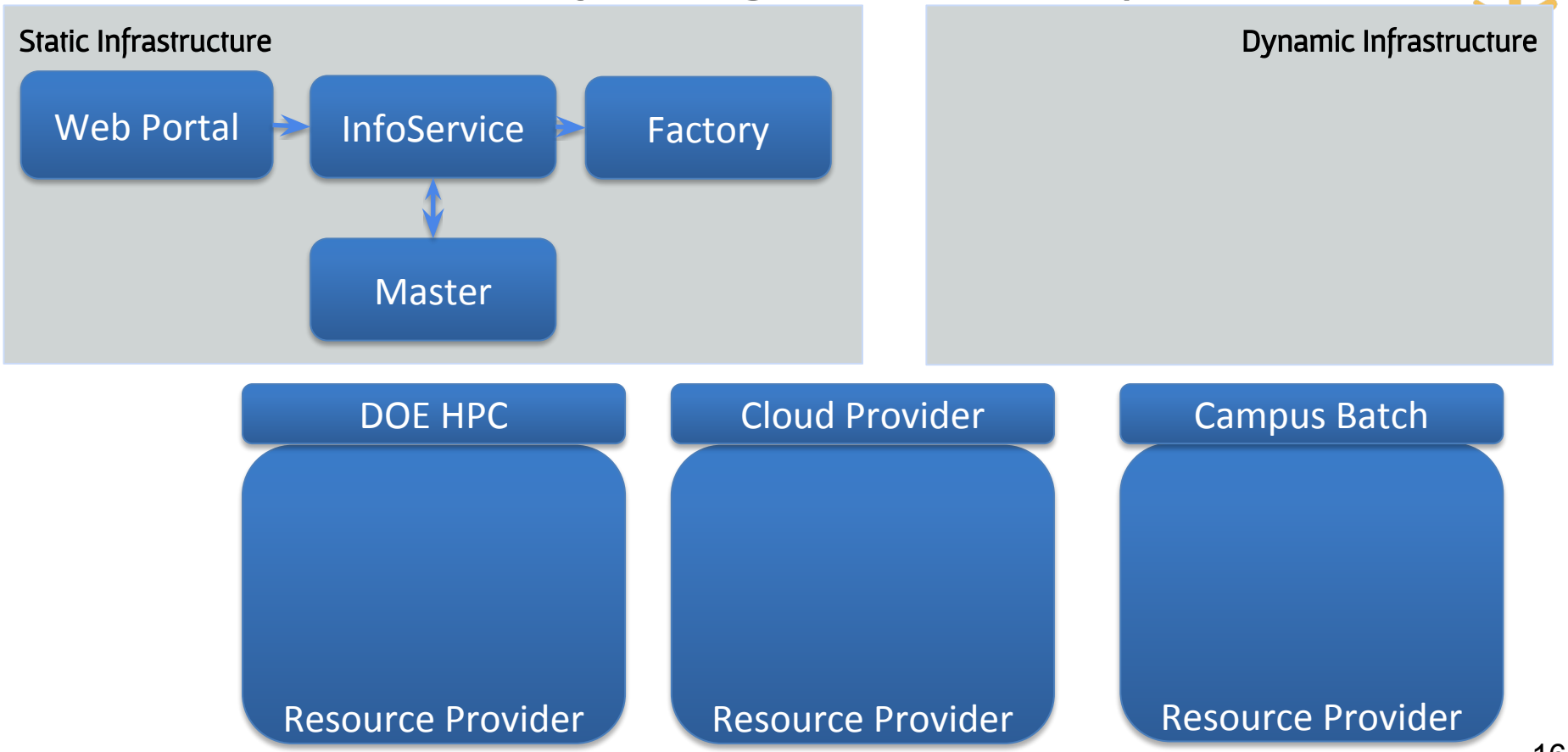
Use the cluster



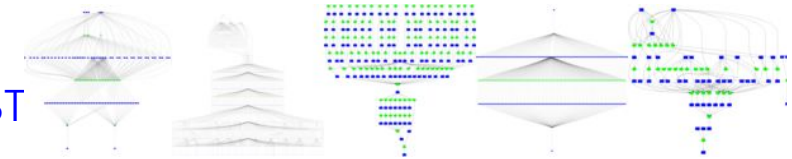
Cleanup is Critical



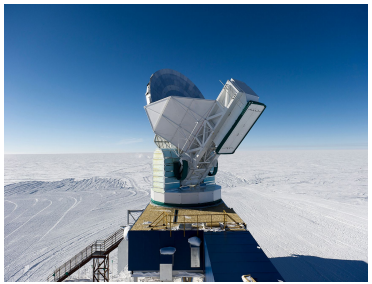
Everything Cleaned Up

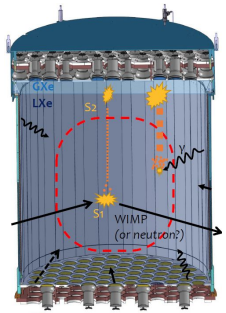


Working Middleware and Applications



- Various Bioinformatics Workflows
  - Makeflow + HTCondor + BWA, Shrimp, BLAST
- Lobster CMS Data Analysis
  - Work Queue + Builder + CVMFS
- South Pole Telescope (SPT-3G) Analysis Framework
  - HTCondor Jobs + Docker/Shifter + CVMFS
- XENON1T Analysis Framework
  - Pegasus + HTCondor + CVMFS
- MAKER Bioinformatics Pipeline
  - Work Queue + Builder
- IceCube Simulation Framework
  - HTCondor

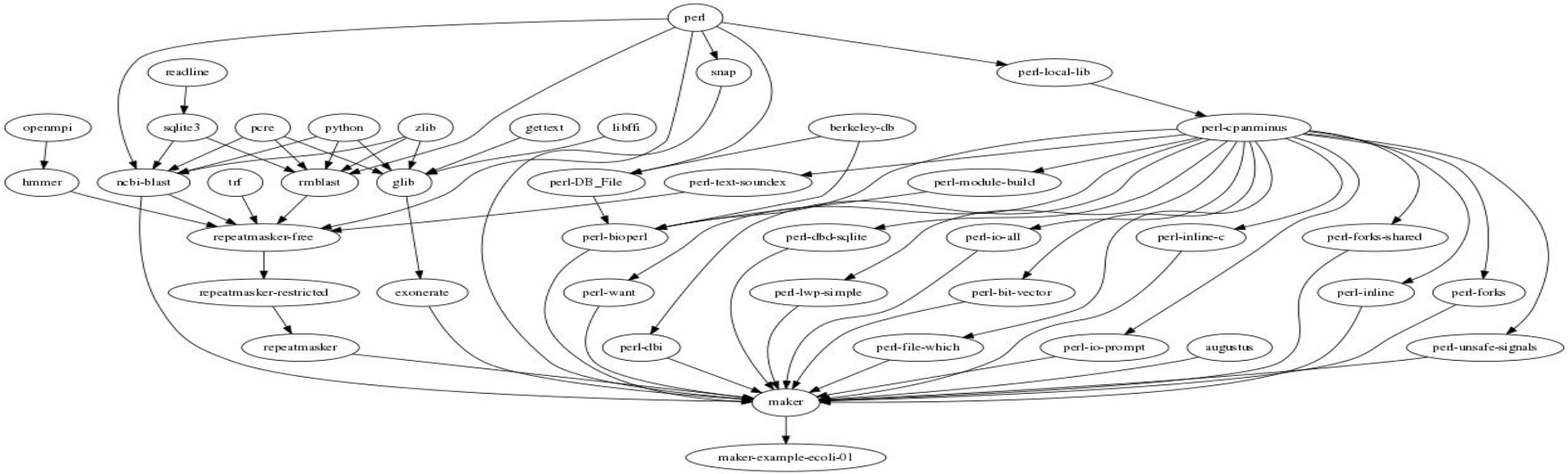




The MAKER Genomics Pipeline

http://www.yandell-lab.org/software/maker.html

vc3-builder **–require maker** maker -BIO



Custom docker container in Jetstream took weeks to install by hand. Converted to vc3-builder, successfully ported to Stampede in a single automated install.

Status, Activity, and Near Future

Tech Previews and Demos: Demos at NGNS meeting and OSG Blueprint. Closed testing with potential users at UChicago in Feb. More planned as features and scaling improve.

Expanding Target Resources: Adding sites, especially ones that require complicated tweaks to run properly; working on Cloud with user allocations, using AWS as dynamic infrastructure.

Expanding Supported Middleware: Apache Spark (supported by vc3-builder w/ S3), Jupyter

Expanding vc3-builder recipes: More software.

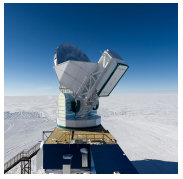
Refining cluster provisioning: E.g. slots are all single-core now. Working on multi-core. Ordered fill of sites, load balancing.

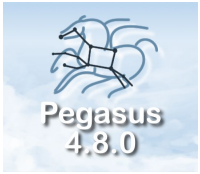
Refining UI: Making sure all internal features are usable cleanly from the portal.

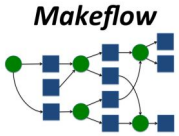
Alpha and Beta testing: Planning on open beta soon.

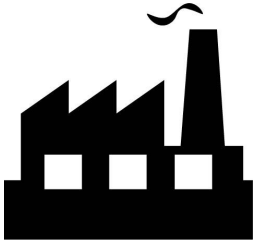
Collaborators and Connections

S2I2 Software Infrastructure









AutoPyFactory

VC3 Funding and Team

Funded for three years by DOE Office of Advanced Scientific Computing Research (ASCR) and NSF Next Generation Networking Services (NGNS)

Primary Investigators: Rob Gardner (UC), Douglas Thain (ND), and John Hover (BNL)

co-PIs: David Miller (UC), Paul Brenner (ND), Mike Hildreth (ND), Kevin Lannon (ND)

Development Team: Lincoln Bryant (UC), Benedikt Reidel (UC),Suchandra Thapa (UC), Jeremy Van (UC), Kenyi Hurtado Anampa (ND), Ben Tovar (ND), Jose Caballero (BNL).

VC3

Virtual Clusters for Community Computation

http://www.virtualclusters.org

https://github.com/vc3-project

Questions?

Extra Slides

Common Questions

How are users authenticated at the portal?

VC3 uses Globus/CILogin to allow users to use their local InCommon federated institutional accounts to enter VC3.

How are allocations accessed? What about user private keys?

As the user enters in their allocation information, they are given a command line to cut-and-paste to run on the resource that copies a VC3 public key into their .ssh/authorized_keys. No user private keys are held by VC3.

Is this only for end users?

No, this system would allow local power users to launch VCs on behalf of other users.

Is this only usable through the web application?

All functionality is available to us as power users/admins via a command line client. Currently access to the central components is restricted by firewall. But in the future we hope to allow full CLI functionality for power users.

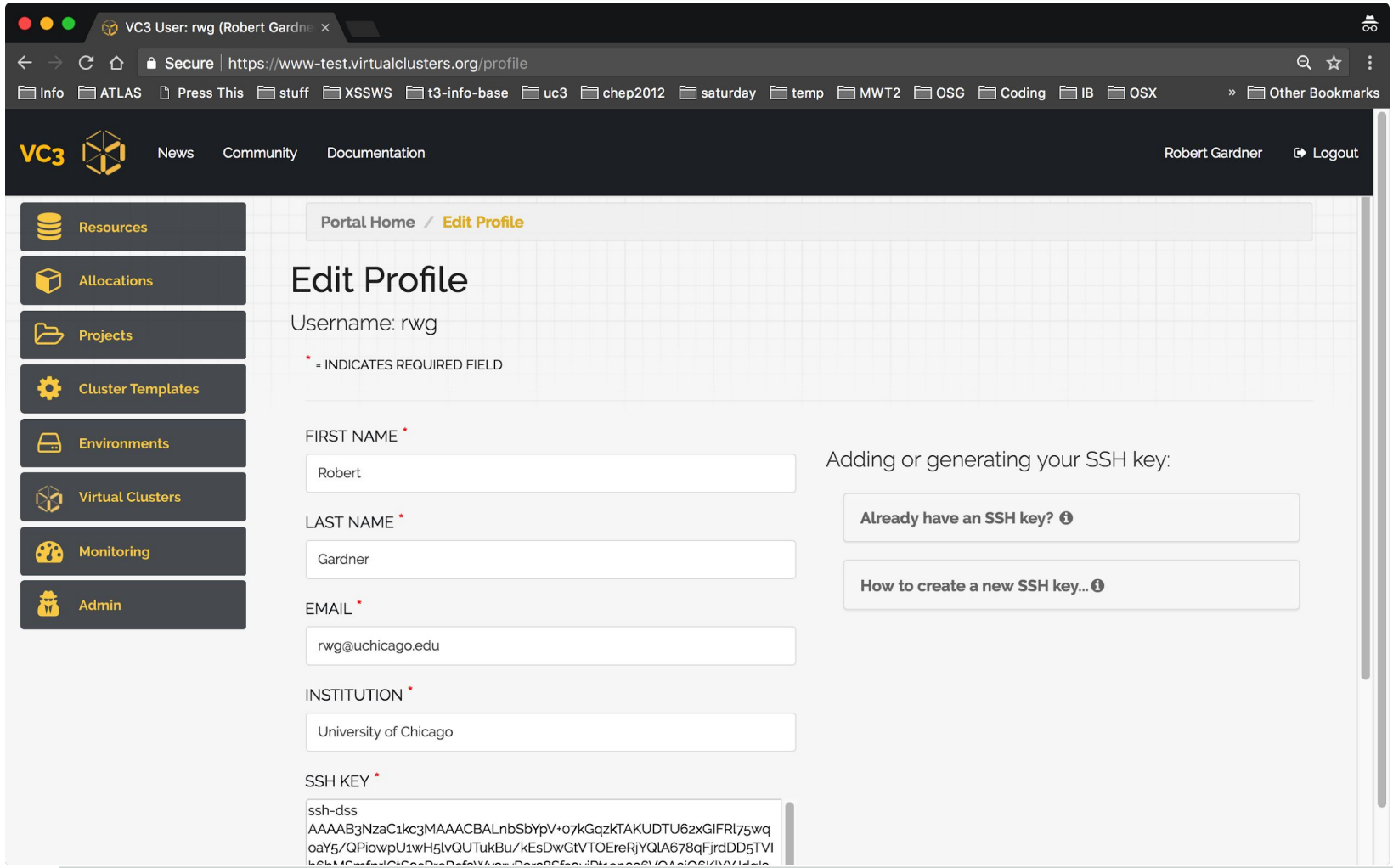
What about MFA?

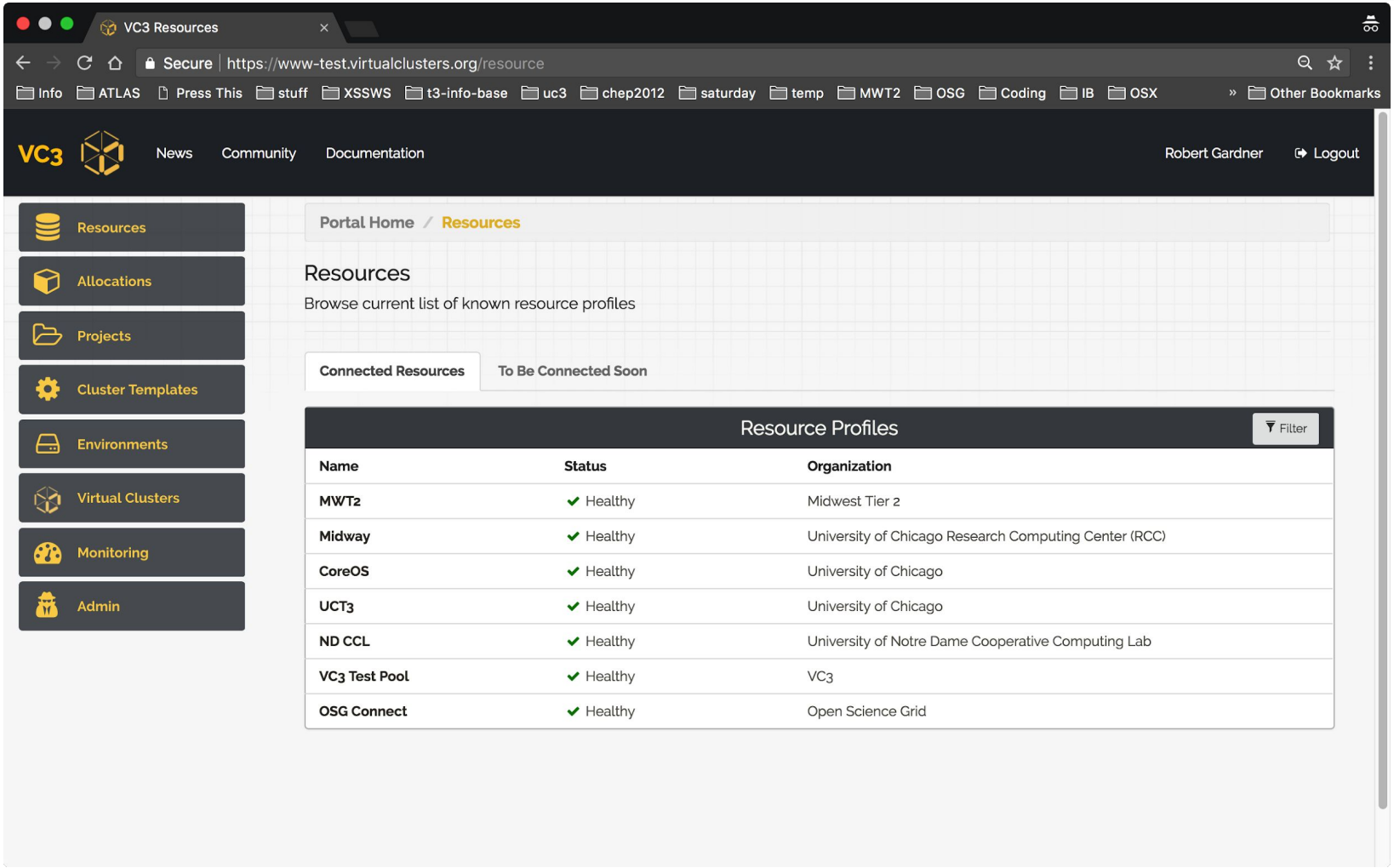
Not currently supported directly. Some suggestion that SSHkey + source IP may satisfy some sites!
We have a technical plan to allow a lot of automation, but with some limited user intervention needed.

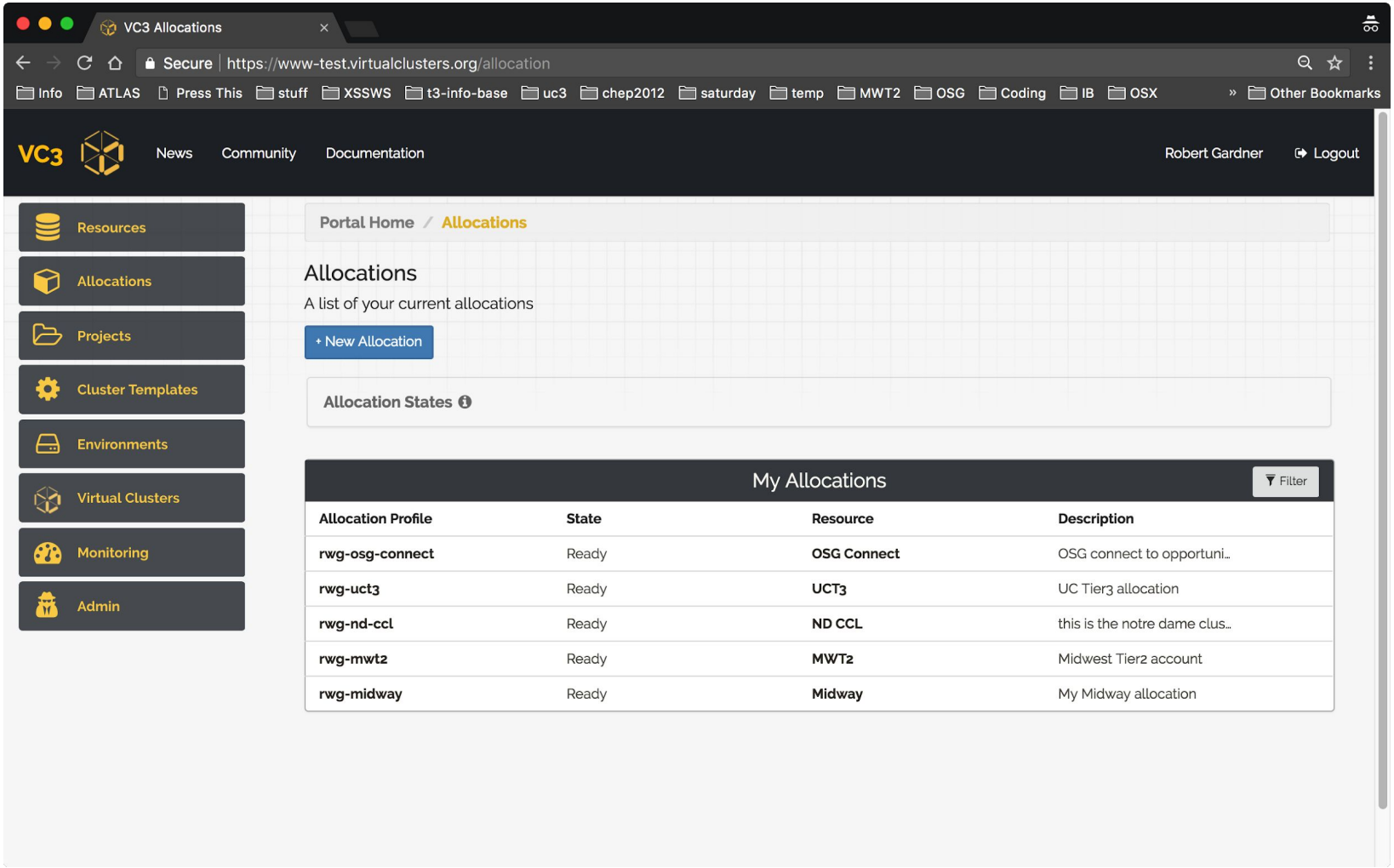
Common Questions...

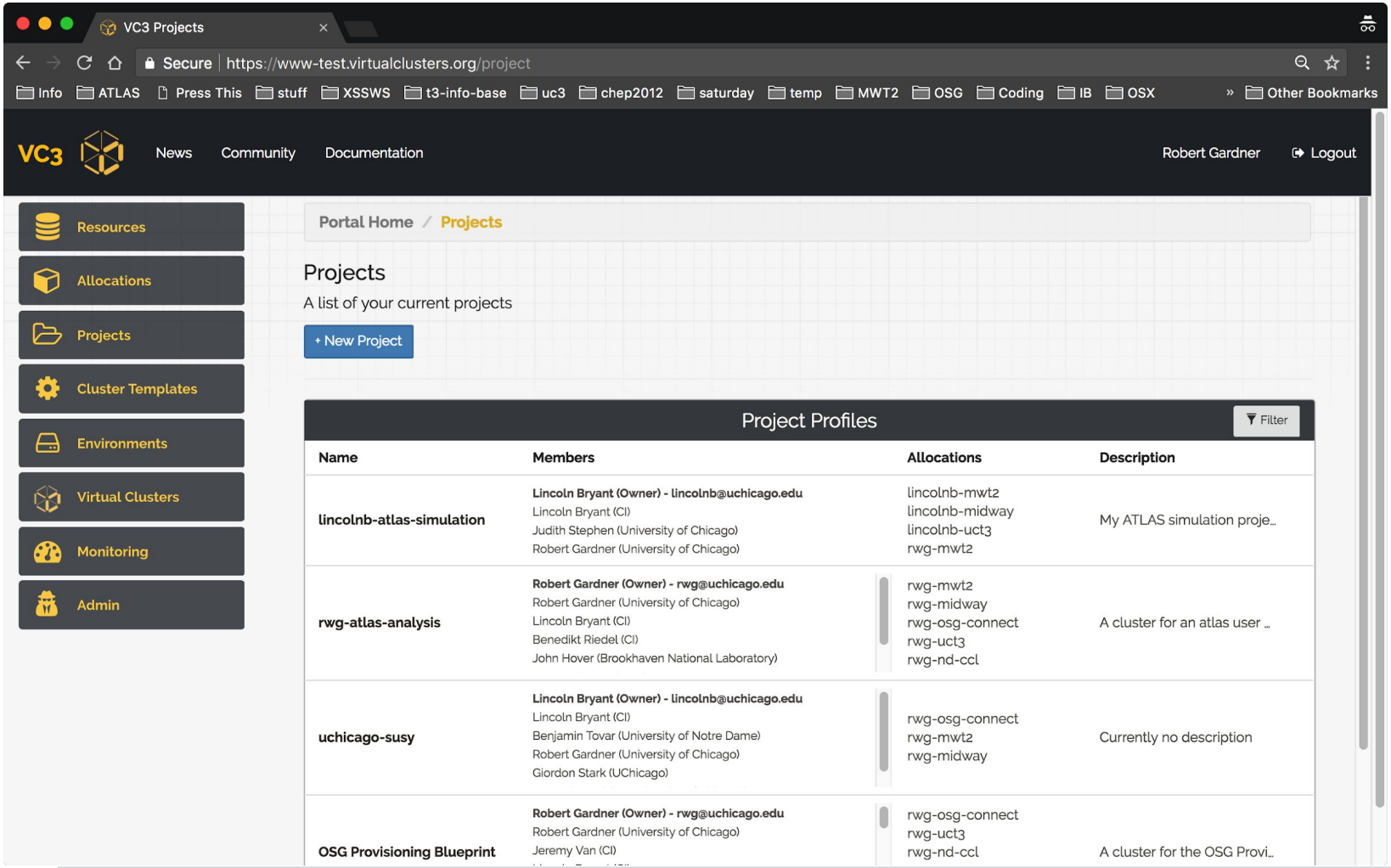
How do user's authenticate to a head node?

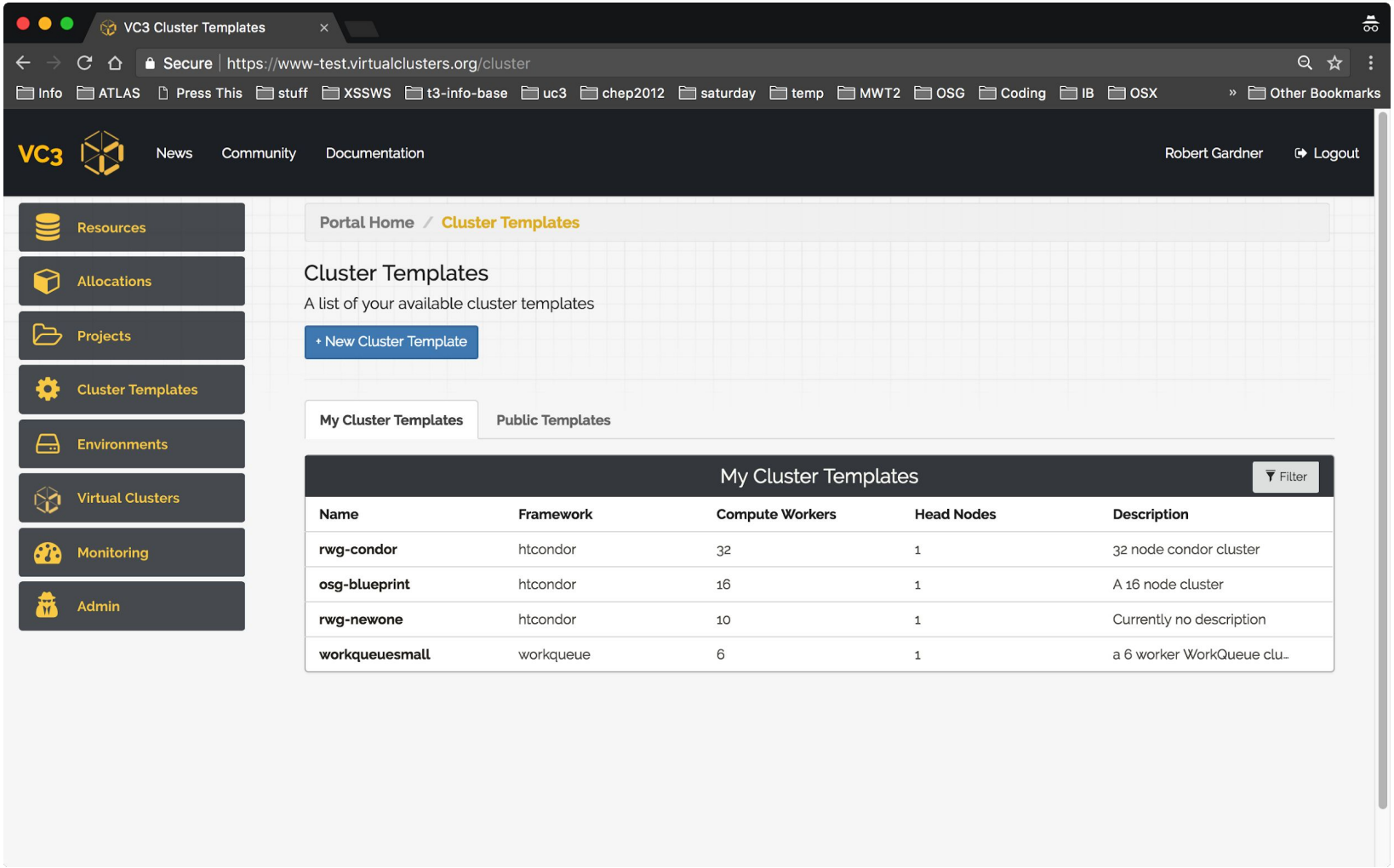
Every user puts their *public* SSH key into their VC3 user profile. It is added to relevant head nodes.

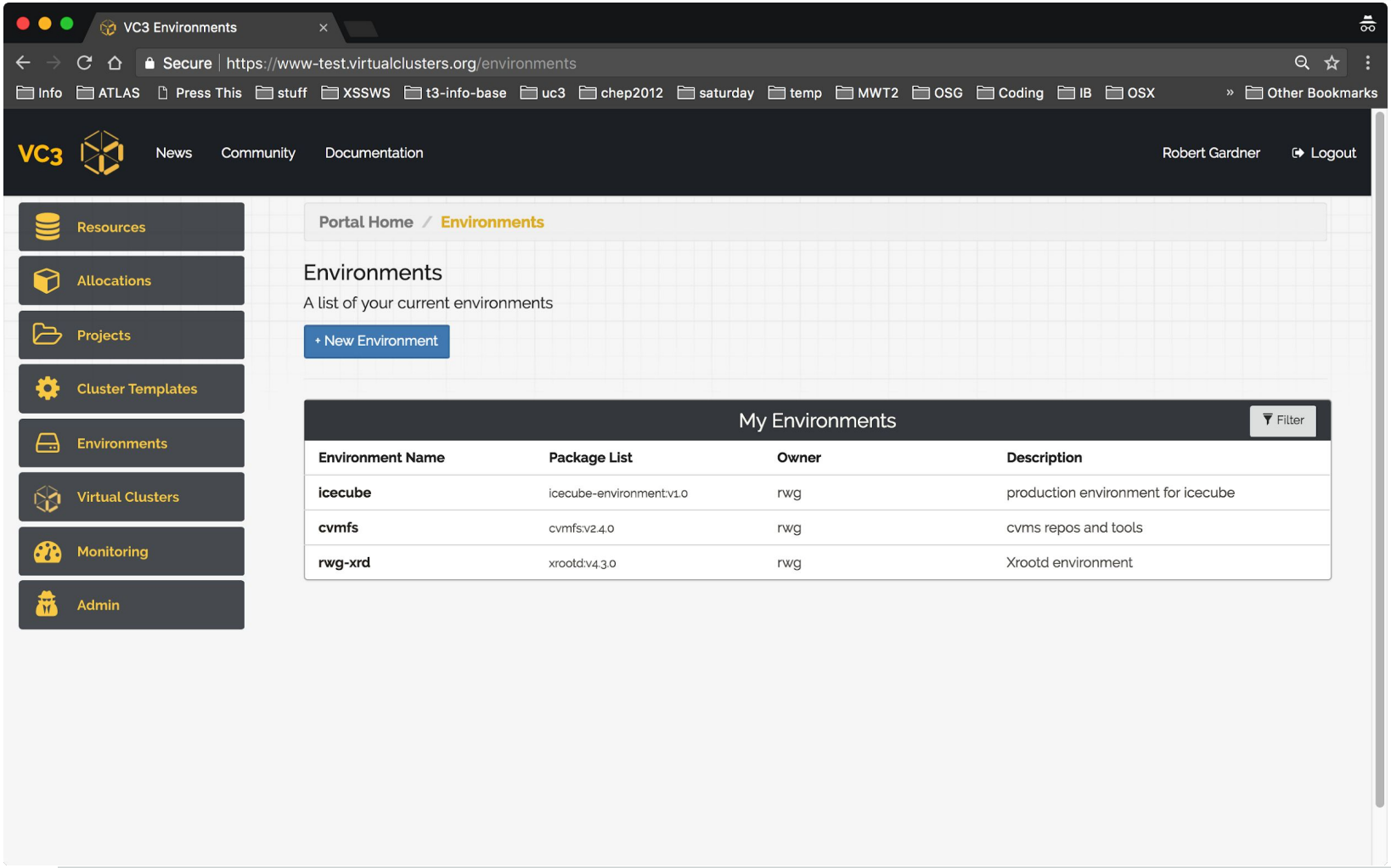

VC3 User: rwg (Robert Gardne
Secure | https://www-test.virtualclusters.org/profile
Info
ATLAS
Press This
stuff
XSSWS
t3-info-base
uc3
chep2012
saturday
temp
MWT2
OSG
Coding
IB
OSX
Other Bookmarks
VC3
News
Community
Documentation
Robert Gardner
Logout
Resources
Allocations
Projects
Cluster Templates
Environments
Virtual Clusters
Monitoring
Admin
Portal Home / Edit Profile
Edit Profile
Username: rwg
* = INDICATES REQUIRED FIELD
FIRST NAME *
Robert
LAST NAME *
Gardner
EMAIL *
rwg@uchicago.edu
INSTITUTION *
University of Chicago
SSH KEY *
ssh-dss
AAAAB3NzaC1kc3MAAACBALnbSbYpV+07kGqzkTAKUDTU62xGIFRl75wq
oaY5/QPiowpU1wH5lvQUTukBu/kEsDwGtVTOEreRjYQlA678qFjrdDD5TVl
Adding or generating your SSH key:
Already have an SSH key?
How to create a new SSH key...

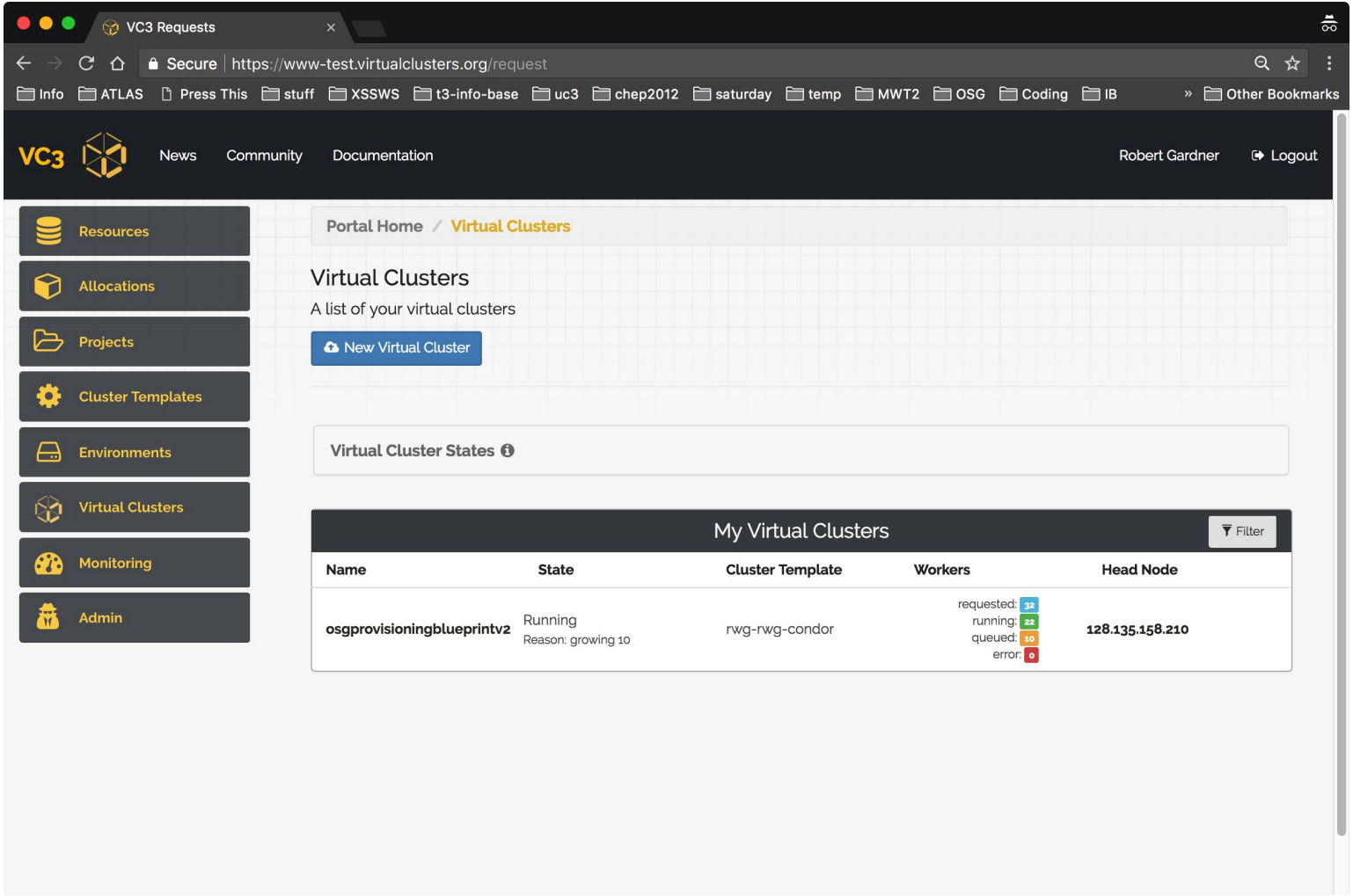

VC3 Resources
Secure | https://www-test.virtualclusters.org/resource
Info ATLAS Press This stuff XSSWS t3-info-base uc3 chep2012 saturday temp MWT2 OSG Coding IB OSX Other Bookmarks
VC3
News
Community
Documentation
Robert Gardner
Logout
Resources
Allocations
Projects
Cluster Templates
Environments
Virtual Clusters
Monitoring
Admin
Portal Home / Resources
Resources
Browse current list of known resource profiles
Connected Resources
To Be Connected Soon
Resource Profiles
Filter
Name Status Organization
MWT2 Healthy Midwest Tier 2
Midway Healthy University of Chicago Research Computing Center (RCC)
CoreOS Healthy University of Chicago
UCT3 Healthy University of Chicago
ND CCL Healthy University of Notre Dame Cooperative Computing Lab
VC3 Test Pool Healthy VC3
OSG Connect Healthy Open Science Grid

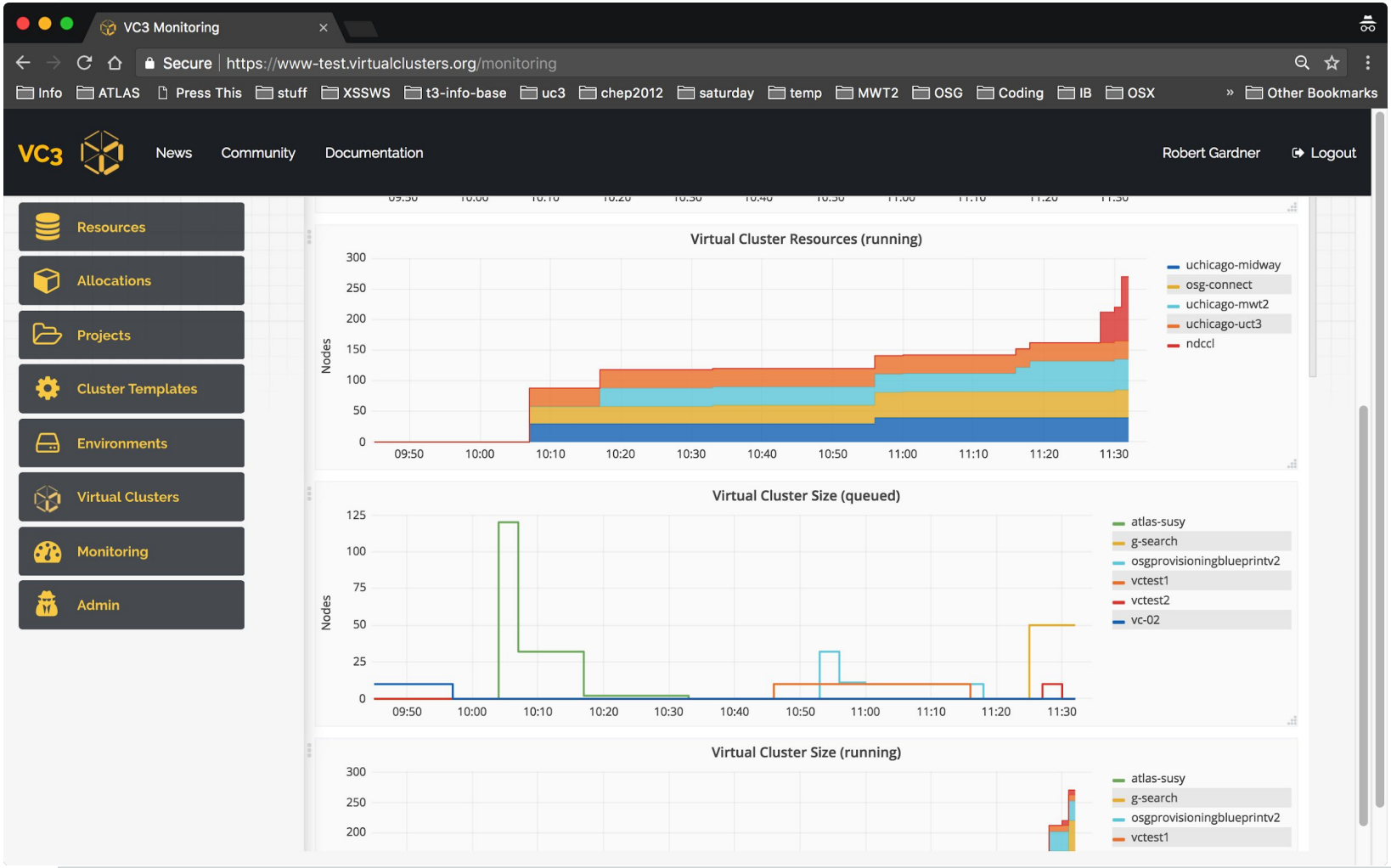

VC3 Allocations
Secure | https://www-test.virtualclusters.org/allocation
Info ATLAS Press This stuff XSSWS t3-info-base uc3 chep2012 saturday temp MWT2 OSG Coding IB OSX Other Bookmarks
VC3
News
Community
Documentation
Robert Gardner
Logout
Resources
Allocations
Projects
Cluster Templates
Environments
Virtual Clusters
Monitoring
Admin
Portal Home / Allocations
Allocations
A list of your current allocations
+ New Allocation
Allocation States
My Allocations
Filter
Allocation Profile | State | Resource | Description
rwg-osg-connect | Ready | OSG Connect | OSG connect to opportuni...
rwg-uct3 | Ready | UCT3 | UC Tier3 allocation
rwg-nd-ccl | Ready | ND CCL | this is the notre dame clus...
rwg-mwt2 | Ready | MWT2 | Midwest Tier2 account
rwg-midway | Ready | Midway | My Midway allocation

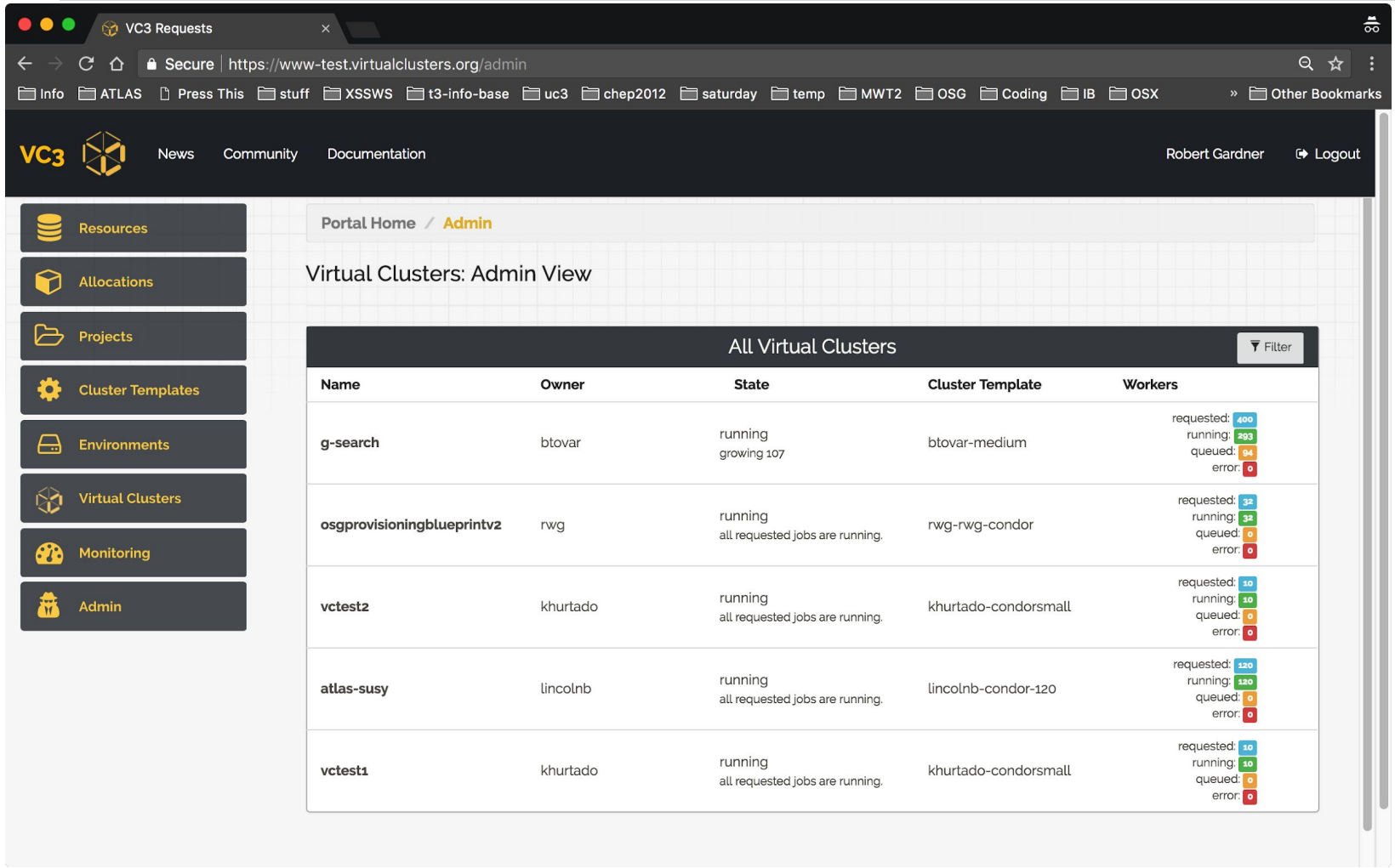

VC3 Projects
Secure | https://www-test.virtualclusters.org/project
Info ATLAS Press This stuff XSSWS t3-info-base uc3 chep2012 saturday temp MWT2 OSG Coding IB OSX Other Bookmarks
VC3
News
Community
Documentation
Robert Gardner
Logout
Resources
Allocations
Projects
Cluster Templates
Environments
Virtual Clusters
Monitoring
Admin
Portal Home / Projects
Projects
A list of your current projects
+ New Project
Project Profiles
Filter
Name
Members
Allocations
Description
lincolnb-atlas-simulation
Lincoln Bryant (Owner) - lincolnb@uchicago.edu
Lincoln Bryant (CI)
Judith Stephen (University of Chicago)
Robert Gardner (University of Chicago)
lincolnb-mwt2
lincolnb-midway
lincolnb-uct3
rwg-mwt2
My ATLAS simulation proje...
rwg-atlas-analysis
Robert Gardner (Owner) - rwg@uchicago.edu
Robert Gardner (University of Chicago)
Lincoln Bryant (CI)
Benedikt Riedel (CI)
John Hover (Brookhaven National Laboratory)
rwg-mwt2
rwg-midway
rwg-osg-connect
rwg-uct3
rwg-nd-ccl
A cluster for an atlas user ...
uchicago-susy
Lincoln Bryant (Owner) - lincolnb@uchicago.edu
Lincoln Bryant (CI)
Benjamin Tovar (University of Notre Dame)
Robert Gardner (University of Chicago)
Giordon Stark (UChicago)
rwg-osg-connect
rwg-mwt2
rwg-midway
Currently no description
OSG Provisioning Blueprint
Robert Gardner (Owner) - rwg@uchicago.edu
Robert Gardner (University of Chicago)
Jeremy Van (CI)
rwg-osg-connect
rwg-uct3
rwg-nd-ccl
A cluster for the OSG Provi...


VC3 Cluster Templates
Secure | https://www-test.virtualclusters.org/cluster
Info ATLAS Press This stuff XSSWS t3-info-base uc3 chep2012 saturday temp MWT2 OSG Coding IB OSX Other Bookmarks
VC3
News
Community
Documentation
Robert Gardner
Logout
Resources
Allocations
Projects
Cluster Templates
Environments
Virtual Clusters
Monitoring
Admin
Portal Home / Cluster Templates
Cluster Templates
A list of your available cluster templates
+ New Cluster Template
My Cluster Templates
Public Templates
My Cluster Templates
Filter
Name Framework Compute Workers Head Nodes Description
rwg-condor htcondor 32 1 32 node condor cluster
osg-blueprint htcondor 16 1 A 16 node cluster
rwg-newone htcondor 10 1 Currently no description
workqueuesmall workqueue 6 1 a 6 worker WorkQueue clu…

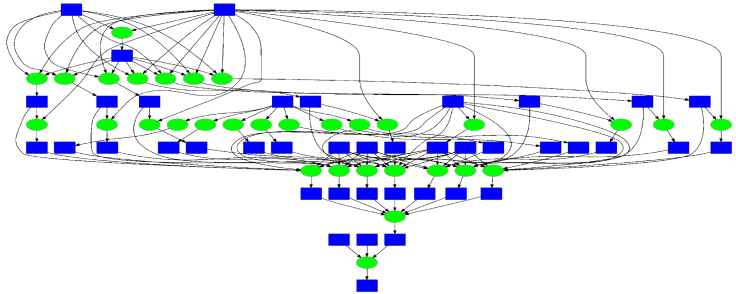

VC3 Environments
Secure | https://www-test.virtualclusters.org/environments
Info ATLAS Press This stuff XSSWS t3-info-base uc3 chep2012 saturday temp MWT2 OSG Coding IB OSX Other Bookmarks
VC3
News
Community
Documentation
Robert Gardner
Logout
Resources
Allocations
Projects
Cluster Templates
Environments
Virtual Clusters
Monitoring
Admin
Portal Home / Environments
Environments
A list of your current environments
+ New Environment
My Environments
Filter
Environment Name | Package List | Owner | Description
icecube | icecube-environment:v1.0 | rwg | production environment for icecube
cvmfs | cvmfs:v2.4.0 | rwg | cvms repos and tools
rwg-xrd | xrootd:v4.3.0 | rwg | Xrootd environment

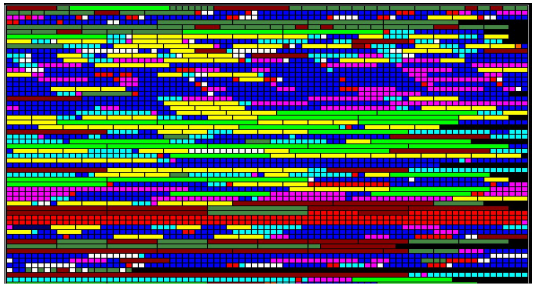

VC3 Requests
Secure | https://www-test.virtualclusters.org/request
Info ATLAS Press This stuff XSSWS t3-info-base uc3 chep2012 saturday temp MWT2 OSG Coding IB Other Bookmarks
VC3
News
Community
Documentation
Robert Gardner
Logout
Resources
Allocations
Projects
Cluster Templates
Environments
Virtual Clusters
Monitoring
Admin
Portal Home / Virtual Clusters
Virtual Clusters
A list of your virtual clusters
New Virtual Cluster
Virtual Cluster States
My Virtual Clusters
Filter
Name
State
Cluster Template
Workers
Head Node
osgprovisioningblueprintv2
Running
Reason: growing 10
rwg-rwg-condor
requested: 32
running: 22
queued: 10
error: 0
128.135.158.210


VC3 Monitoring
Secure | https://www-test.virtualclusters.org/monitoring
Info ATLAS Press This stuff XSSWS t3-info-base uc3 chep2012 saturday temp MWT2 OSG Coding IB OSX Other Bookmarks
VC3
News
Community
Documentation
Robert Gardner
Logout
Resources
Allocations
Projects
Cluster Templates
Environments
Virtual Clusters
Monitoring
Admin
Virtual Cluster Resources (running)
Nodes
300
250
200
150
100
50
0
09:50
10:00
10:10
10:20
10:30
10:40
10:50
11:00
11:10
11:20
11:30
uchicago-midway
osg-connect
uchicago-mwt2
uchicago-uct3
ndccl
Virtual Cluster Size (queued)
Nodes
125
100
75
50
25
0
atlas-susy
g-search
osgprovisioningblueprintv2
vctest1
vctest2
vc-02
Virtual Cluster Size (running)
300
250
200
atlas-susy
g-search
osgprovisioningblueprintv2
vctest1


VC3 Requests
Secure | https://www-test.virtualclusters.org/admin
Info ATLAS Press This stuff XSSWS t3-info-base uc3 chep2012 saturday temp MWT2 OSG Coding IB OSX Other Bookmarks
VC3
News
Community
Documentation
Robert Gardner
Logout
Resources
Allocations
Projects
Cluster Templates
Environments
Virtual Clusters
Monitoring
Admin
Portal Home / Admin
Virtual Clusters: Admin View
All Virtual Clusters
Filter
Name Owner State Cluster Template Workers
g-search btovar running growing 107 btovar-medium requested: 400 running: 293 queued: 94 error: 0
osgprovisioningblueprintv2 rwg running all requested jobs are running. rwg-rwg-condor requested: 32 running: 32 queued: 0 error: 0
vctest2 khurtado running all requested jobs are running. khurtado-condorsmall requested: 10 running: 10 queued: 0 error: 0
atlas-susy lincolnb running all requested jobs are running. lincolnb-condor-120 requested: 120 running: 120 queued: 0 error: 0
vctest1 khurtado running all requested jobs are running. khurtado-condorsmall requested: 10 running: 10 queued: 0 error: 0

You have developed a complex workload which runs successfully at one site, perhaps your home university.





Now, you want to migrate and expand that application to national-scale infrastructure.
And allow others to easily access and run similar workloads.

Traditional HPC Facility

Open Science Grid

Distributed HTC Facility

Commercial Cloud

Caveats and Limitations

Network access: System is an overlay, so network access from worker nodes is required.

Human-level MFA: If a site requires per-access human MFA confirmation, that is very hard to automate over. Various 7-day proxy schemes may mitigate this. Also negotiating to use connection source as part of MF.

Cloud Credentials: VC3 would need some private security token to act on the user's behalf for these targets. Lesser-privileged IAM-type accounts OK for cloud.

Accounting: Still to come, and is very hard to implement in a way guaranteed to always catch over-use. But everyone operating on allocation-based resources faces this.