

---

# VC3

## Virtual Clusters for Community Computation

DOE Collaborative Projects Meeting  
Sponsored by the U.S. Department of Energy  
Office of Advanced Scientific Computing Research  
September 23–25, 2019

---

Douglas Thain, University of Notre Dame  
Rob Gardner, University of Chicago  
John Hover, Brookhaven National Lab

---

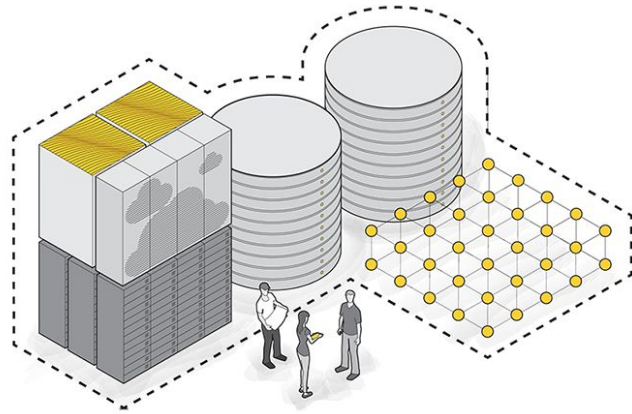
# Outline

---



- Intro to the Virtual Cluster Concept
- Self-service Provisioning of Middleware Across Heterogeneous Resources
- Dealing with Software Environments
- Clustering Middleware Examples
- New Applications and Configurations
- Thoughts and Lessons Learned

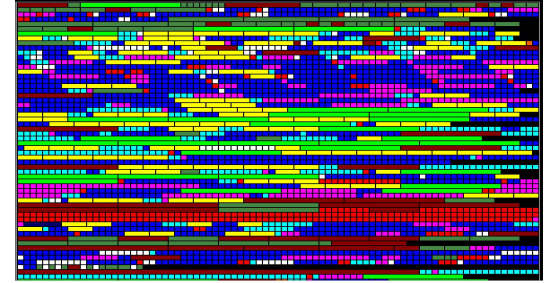
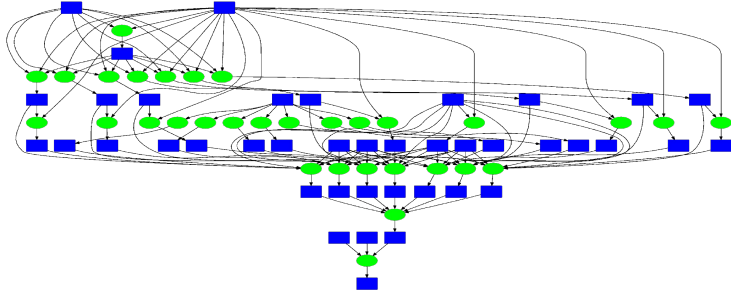
# Introduction



VC3: A platform for provisioning cluster frameworks over heterogeneous resources for collaborative science teams



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



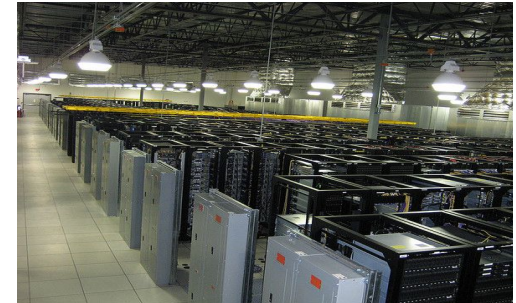
Now, you want to migrate and expand that application to national-scale infrastructure.  
And allow others to easily access and run similar workloads.



Leadership HPC Facility



Distributed Computing Facility

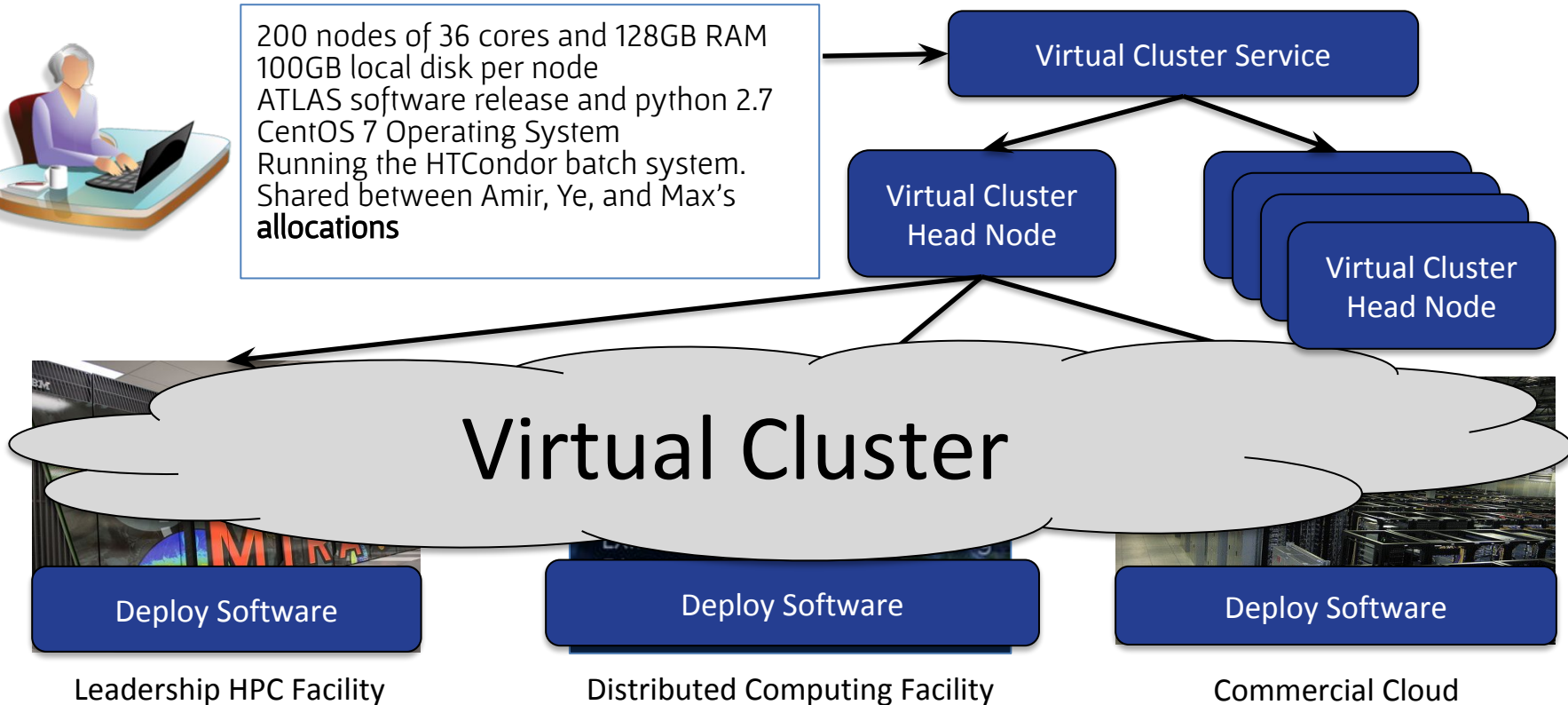


Commercial Cloud

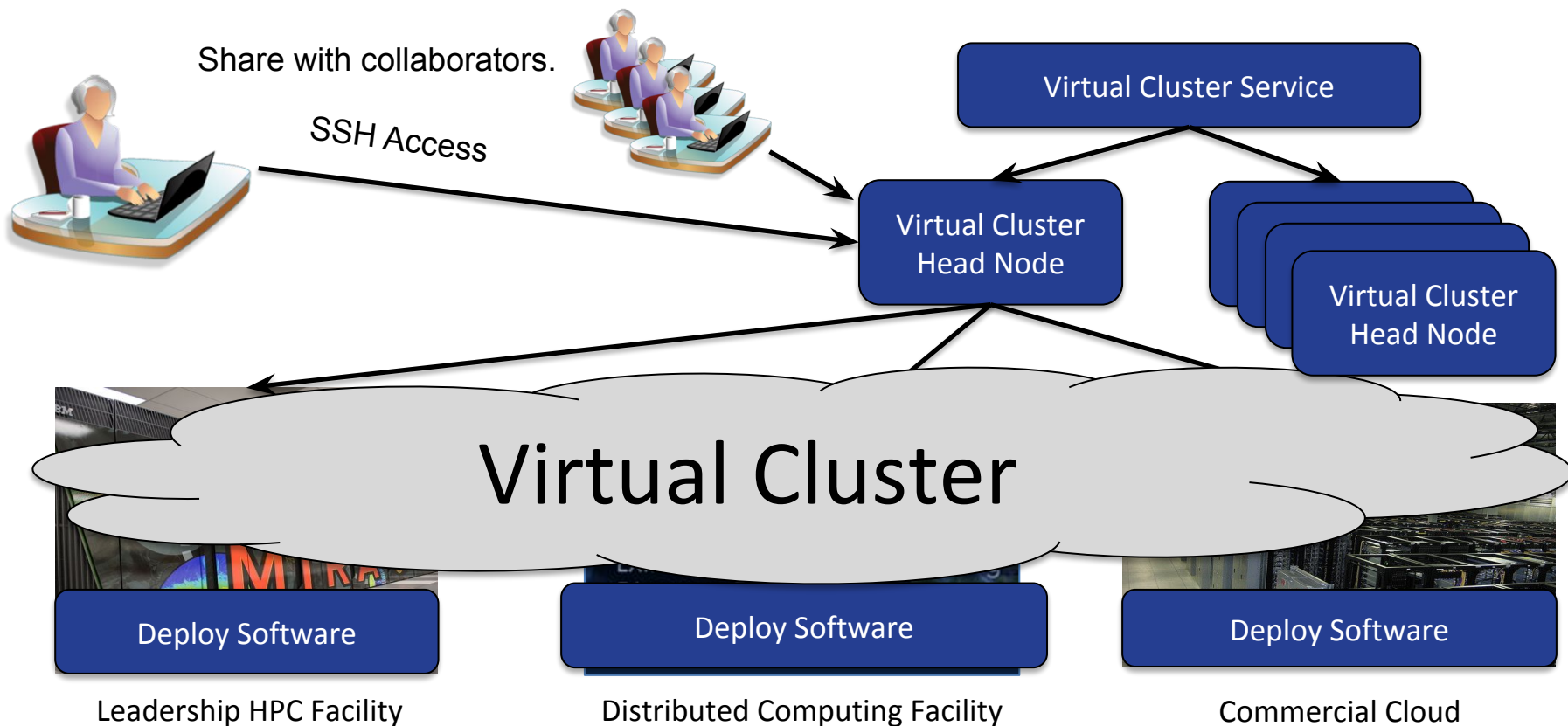
# Concept: Virtual Cluster



200 nodes of 36 cores and 128GB RAM  
100GB local disk per node  
ATLAS software release and python 2.7  
CentOS 7 Operating System  
Running the HTCondor batch system.  
Shared between Amir, Ye, and Max's  
**allocations**



# Concept: Virtual Cluster



# VC3: Virtual Clusters for Community Computation



- **VC3 is an interactive service for creating/sharing/using virtual clusters.**
- A virtual cluster consists of:
  - 1 x **head node** for interactive access to the cluster. (SSH, Jupyter,..)
  - N x **worker nodes** for executing your workload.
  - **Middleware** to manage the cluster. (HTCondor, Makeflow, Spark, ...)
  - **Application** software to do real work. (BLAST, CMSSW, etc...)
- A virtual cluster is created using:
  - Your standard accounts/credentials on existing facilities.
  - Plain ssh/qsub/srun access on each facility.
  - Container technology (if available) or user-level software builds (otherwise).
  - (No special privileges or admin access required on the facility.)



# VC3 Managed Infrastructure

## Static

vc3-website  
User Interface



vc3-client  
Service Library

```

user-create()
user-list()
user-delete()
project-create()
project-adduser()
resource-create()
resource-list()
resource-delete()
allocation-create()
.
.
.

```

```

create new vc3 user
list vc3 user(s)
delete a vc3 user
create new vc3 proj
add user to vc3 proj
create new vc3 rsc
list vc3 resource(s)
delete vc3 resource
create new vc3 alloc
.
.
.

```

vc3-info-service  
Object Persistence Store

User
Project
Resource
Cluster
Allocation

vc3-master  
Task Execution

```

TaskSets/Tasks:
AddFactoryConfiguration
CheckAllocations
CheckResourceAccess
HandleAllocations
HandleHeadNodes
HandleRequests
InitInstanceAuth
InitResources
.
.
.

```

Private CA

3. Master creates all info for Cluster request: configs, auth. Periodically updates...

vc3-factory  
Provisioning Engine

```

queues.conf
:
[AResX]
[AResZ]
[BResZ]
[BResY]
.
.
.

```

condor-c

Config Plugin

5. Factory config pulled from InfoService.

## Dynamic

Head Node A  
HTCondor Head Node

Head Node B  
Spark Head Node

4. Master creates cluster head nodes on Openstack...

## Distributed Resources

Resource X

vc3-builder  
HTCondor glidein

Resource Y

vc3-builder  
Spark Worker

Resource Z

vc3-builder  
HTCondor glidein

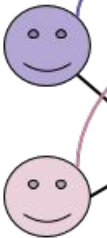
vc3-builder  
Spark Worker

6. Factory submits vc3-builders via SSH to User Resources

7. Users log into Head Nodes and do work!

1. Users make Cluster requests via web interface...

2. Request info placed in Infoservice for processing.



1. Users make Cluster requests via web interface...

2. Request info placed in Infoservice for processing.

5. Factory config pulled from InfoService.

4. Master creates cluster head nodes on Openstack...

6. Factory submits vc3-builders via SSH to User Resources

7. Users log into Head Nodes and do work!

# Self-Service Provisioning





The screenshot shows the homepage of the Virtual Clusters for Community Computation website. At the top left is the VC3 logo. A dark navigation bar contains the following links: News, Resources, Team, Community, Documentation, and Login. The main content area features a background image of server racks with a large, glowing VC3 logo in the center. Below the logo, the text reads "Virtual Clusters for Community Computation" and "Deploy HTCondor and WorkQueue clusters". At the bottom of the main content area are two buttons: "ABOUT VC3" and "SIGN UP".

login with a vc3 account

# Institutional Identity (CI-Logon)



Globus Account Log In

Log in to use VC3

Use your existing organizational login

e.g., university, national lab, facility, project

University of Notre Dame

Didn't find your organization? Then use [Globus ID to sign in](#). ([What's this?](#))

Continue



Globus uses CILogon to enable you to Log In from this organization. By clicking Continue, you agree to the [CILogon privacy policy](#) and you agree to share your username, email address, and affiliation with CILogon and Globus. You also agree for CILogon to issue a certificate that allows Globus to act on your behalf.

Or

 Sign in with Google

 Sign in with ORCID ID



# Curated Resources



Resource Profiles <span>Filter</span>							
Name	Organization	Description	Cores	Memory	Storage	Native OS	Features
<b>Cori</b>	National Energy Research Scientific Computing Center (NERSC)	Cori Supercomputer at NERSC	32	4000 MB	10000 MB	suse.v12	Shifter
<b>MWT2</b>	Midwest Tier 2	ATLAS Midwest Tier 2 Center job gateway (UChicago)	4	1000 MB	1000 MB	scientificlinux.v6.9	N/A
<b>Midway</b>	University of Chicago Research Computing Center (RCC)	Midway cluster at the University of Chicago Research Computing Center (RCC)	64	4000 MB	10000 MB	scientificlinux.v6.7	N/A
<b>Stampede 2</b>	Texas Advanced Computing Center (TACC)	Stampede 2 Super Computer	96	2000 MB	10000 MB	centos.v7.4	Singularity
<b>CoreOS</b>	University of Chicago	CoreOS Cluster	4	1000 MB	1000 MB	scientificlinux.v6.9	Singularity
<b>UCT3</b>	University of Chicago	UChicago ATLAS Tier 3	4	1000 MB	1000 MB	scientificlinux.v6.9	N/A
<b>ND CCL</b>	University of Notre Dame Cooperative Computing Lab	ND-CCL login none	4	1000 MB	10000 MB	redhat.v7	Singularity
<b>Bridges</b>	Pittsburgh Supercomputing Center	Bridges Supercomputer at PSC	28	4000 MB	35000 MB	centos.v7.3	Singularity
<b>VC3 Test Pool</b>	VC3	VC3 Test Pool	4	1000 MB	1000 MB	centos.v6.9	N/A
<b>UCLA Hoffman2</b>	University of California, Los Angeles	UCLA Hoffman2	8	1000 MB	10000 MB	centos.v6.9	N/A
<b>OSG Connect</b>	Open Science Grid	Open Science Grid (SL7)	4	1000 MB	1000 MB	Unknown	N/A

# Allocations



## Step 1: Log Into Resource

In a terminal, type:

```
ssh btovar@cc1vm05.crc.nd.edu
```

## Step 2: Access Resource

Enter your password for `cc1vm05.crc.nd.edu` for access

## Step 3: Add Allocation SSH Public Key to Resource

Once the SSH key is generated below, click 'Copy to Clipboard' and paste the following line into your SSH session. You will only need to do this once per allocation.

```
s*00rwwcgmrkubdkwmcncgcawcrvx7anigrw10wccdq9fsc13604r0c9g7e9k  
/GTjh8YrCyX6UhqG+S3nOxOf+ewxx3RSIMf9lsFzPDNdXwJl1YD1dyRCYy8TwNhBggGikCxEKMqfOgo  
L6ROpicuUhFY6yT9apKGox1mPSM  
/94ETHxIkBmNK8Phg26fuT+F+QQToSQVovgoghWLGidNoztW8OUkSFzZ6uZE5zfPp0xq45a4*FYE  
TorlJRapgPsjmSjmSB7TeD+qs1ECilwrrg3JPoRBOEMMeLf7rwgDxjtzkBUQ7zlkq5lXTUAYeuoCbGgll  
Q7ZHGHRNTyKkSPL7rXEI7nnz6ofgUJCU3L7hr2VKKy84RcHPsfep64qV3jIOcw1o6SPvu6iwRYeqhfe  
Aoo  
/yKp1vapyfM7Ptuy+6yWZ7grZlb9AtBolcoBColpig64MR8T4D8Rkp1960nCG5ltXwC4mmPSgffQofOl  
WJom7TudG+yTWouWikipoieObZX5w8SKFcoH
```

Copy to Clipboard

## Step 4: Validate Allocation

# Projects



Project Profiles <span>Filter</span>			
Name	Members	Allocations	Description
<b>vc3-team</b>	<b>Benjamin Tovar (Owner) - btovar@nd.edu</b> Lincoln Bryant (UChicago) Jeremy Van (UChicago) Robert Gardner (UChicago) Kenyi Hurtado (University of Notre Dame)	btovar-ndccl khurtado-osgconnect lincolnb-midway	Currently no description
<b>btovar</b>	<b>Benjamin Tovar (Owner) - btovar@nd.edu</b> Benjamin Tovar (University of Notre Dame)	btovar-ndccl	Currently no description

# Launching a Virtual Cluster



VIRTUAL CLUSTER NAME

CLUSTER TEMPLATE \*

ENVIRONMENT

ALLOCATIONS \*

Nothing selected

Select Allocations for Virtual Cluster

Select All Deselect All

btovar-ndccl

khurtado-osgconnect

lincoln-btcondor-10-workers

lincoln-midway

A brief description...

shared cluster definition

workers will have this environment installed

allocations available in this project

# Cluster Status



My Virtual Clusters <span>Filter</span>				
Name	State	Cluster Template	Workers	Head Node
<b>my-virtual-cluster</b>	<div style="background-color: #28a745; color: white; padding: 2px 5px; border-radius: 3px;">Running</div> <p>All requested compute workers are running.</p>	lincoln-btcondor-10-workers	Requested: <span style="background-color: #17a2b8; color: white; padding: 0 2px;">10</span> Running: <span style="background-color: #28a745; color: white; padding: 0 2px;">7</span> Queued: <span style="background-color: #ffc107; color: white; padding: 0 2px;">3</span> Error: <span style="background-color: #dc3545; color: white; padding: 0 2px;">0</span>	<b>128.135.158.187</b>

# Workers from many sites



```
[btovar@btovar-my-virtual-cluster ~]$ ip addr | grep 128.135.158.187
    inet 128.135.158.187/25 brd 128.135.158.255 scope global dynamic eth0
[btovar@btovar-my-virtual-cluster ~]$ condor_status
```

Name	OpSys	Arch	State	Activity	LoadAv	Mem	ActvtyTime
slot1@glidein_21791@camd01.crc.nd.edu	LINUX	X86_64	Unclaimed	Idle	5.120	4013	0+00:19:37
slot1@glidein_29106@camd01.crc.nd.edu	LINUX	X86_64	Unclaimed	Idle	5.120	4013	0+00:19:37
slot1@glidein_91802@camd05.crc.nd.edu	LINUX	X86_64	Unclaimed	Idle	5.260	4013	0+00:19:37
slot1@glidein_39133@iut2-c257.iu.edu	LINUX	X86_64	Unclaimed	Idle	34.620	3223	0+00:19:48
slot1@glidein_61297@lnxfarm275.colorado.edu	LINUX	X86_64	Unclaimed	Idle	6.990	3002	0+00:14:36
slot1@glidein_28373@midway091.rcc.local	LINUX	X86_64	Unclaimed	Idle	8.170	2013	0+00:19:36
slot1@glidein_71179@midway098.rcc.local	LINUX	X86_64	Unclaimed	Idle	7.480	2013	0+00:19:36
slot1@glidein_46364@midway260.rcc.local	LINUX	X86_64	Unclaimed	Idle	8.040	2013	0+00:19:35
slot1@glidein_39282@midway324.rcc.local	LINUX	X86_64	Unclaimed	Idle	8.750	2013	0+00:19:36
slot1@glidein_39133@uct2-c373.mwt2.org	LINUX	X86_64	Unclaimed	Idle	34.080	2415	0+00:19:33

Machines Owner Claimed Unclaimed Matched Preempting Drain

X86_64/LINUX	10	0	0	10	0	0	0
Total	10	0	0	10	0	0	0

```
[btovar@btovar-my-virtual-cluster ~]$
```

Notre  
Dame

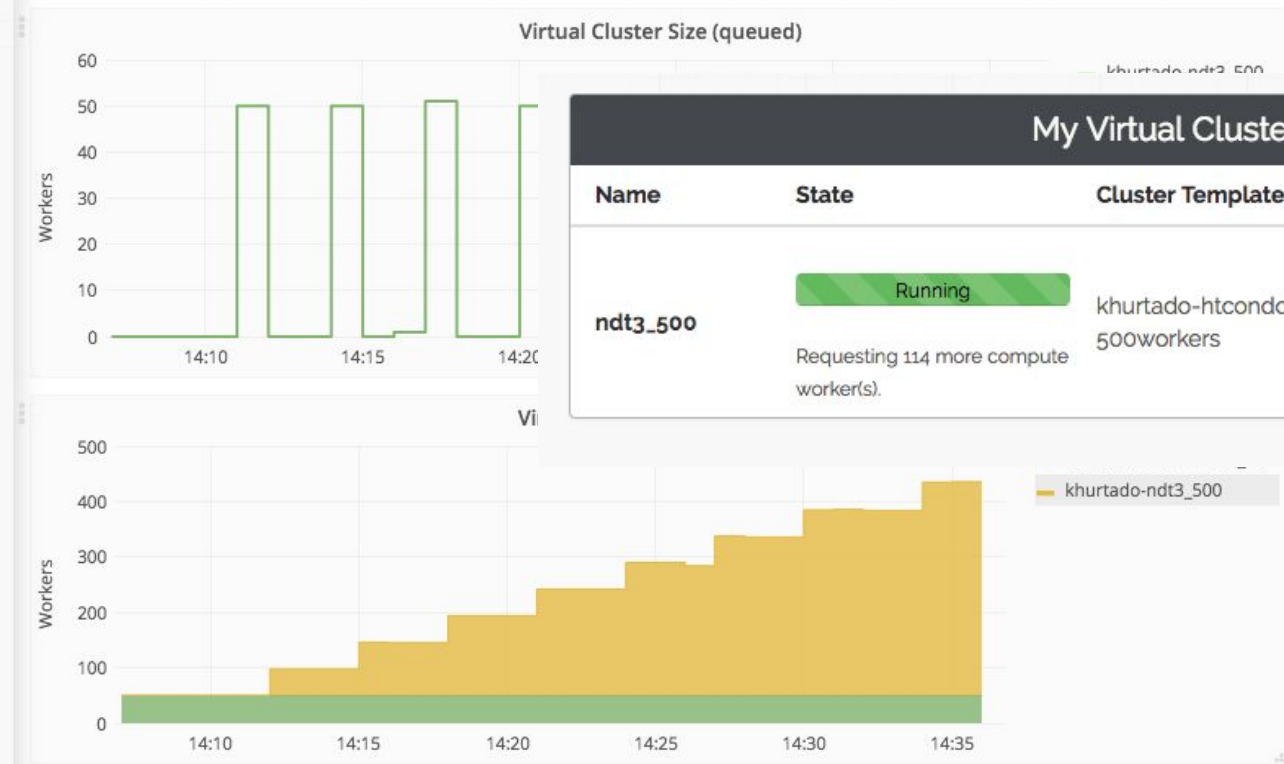
OSG

UChicago

1  
8



## VC3 Monitoring



### My Virtual Clusters Filter

Name	State	Cluster Template	Workers	Head Node
ndt3_500	<span>Running</span> Requesting 114 more compute worker(s).	khurtado-htcondor-500workers	Requested: 500 Running: 386 Queued: 48 Error: 0	128.135.158.178

# Dealing with Software Environments



# Deploying Software Environments

---



The **vc3-builder**, a command-line tool for deploying software environments on clusters.

```
vc3-builder
  --require-os centos:7
  --mount /scratch=/data
  --require /cvmfs
  --require python:2.7 -- myapp ...my args...
```

# Basic Use Case: Application Software



```
vc3-builder --require ncbi-blast
```

```
..Plan:  ncbi-blast => [, ]
```

```
..Try:   ncbi-blast => v2.2.28
```

```
....Plan:  perl => [v5.008_1
```

```
....Try:   perl => v5.
```

```
....could not add any
```

```
....Try:   perl => v5.
```

```
....could not add any
```

```
....Try:   perl => v5..
```

```
.....Plan:  perl-vc3-
```

```
.....Try:   perl-vc3-r
```

```
.....Success: perl-vc
```

```
....Success: perl v5.2
```

```
....Plan:  python => [v2.006, ]
```

```
....Try:   python => v2.6.0
```

**(New Shell with Desired Environment)**

```
bash$ which blastx
```

```
/tmp/test/vc3-root/x86_64/redhat6/ncbi-blast/v2.2.28/bin  
/blastx
```

```
bash$ blastx --help
```

```
USAGE
```

```
blastx [-h] [-help] [-import_search_strategy filename]
```

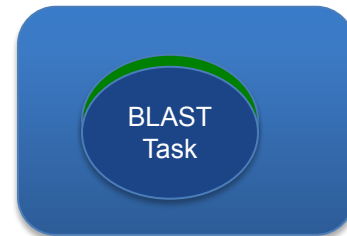
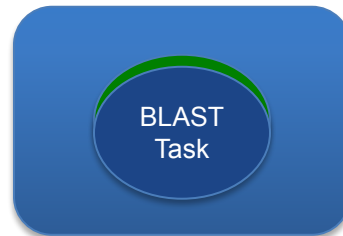
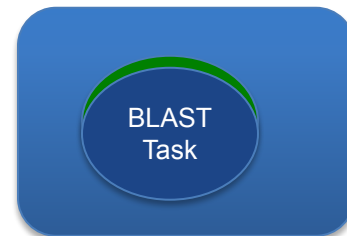
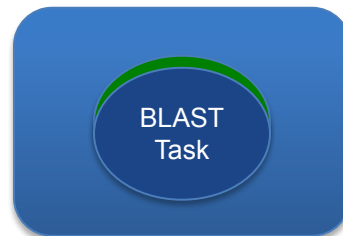
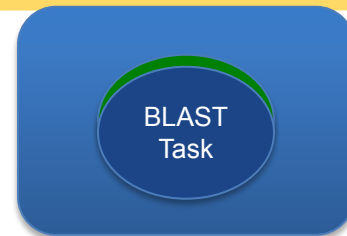
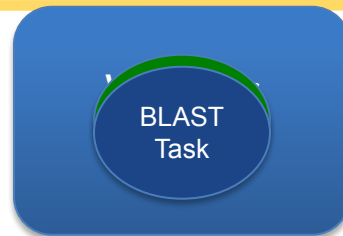
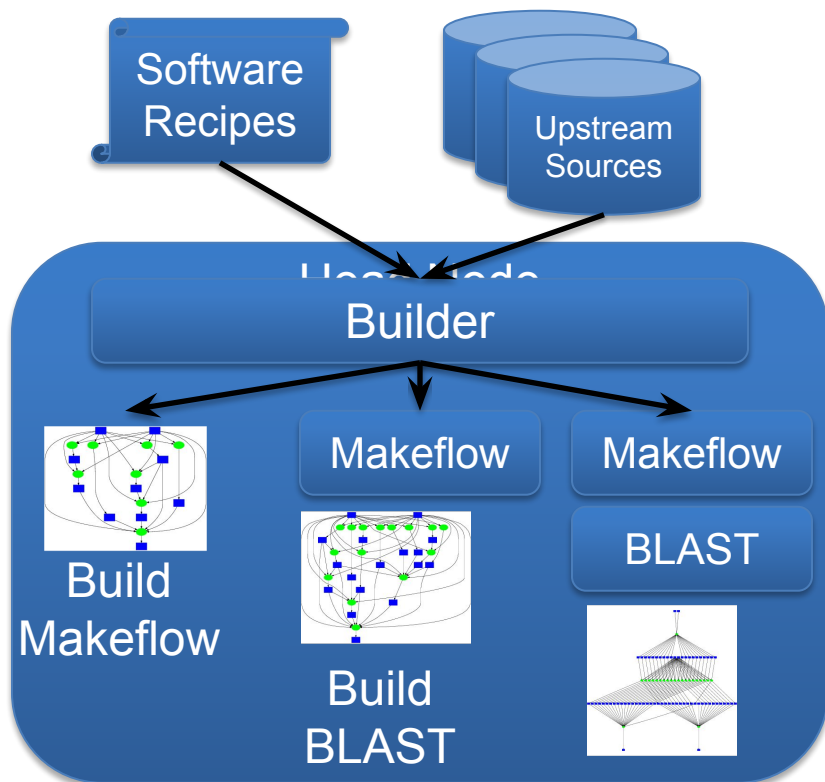
# Problem: Long Build on Head Node

---



- Many facilities limit the amount of work that can be done on the head node, so as to maintain quality of service for everyone.
- Solution: Move the build jobs out to the cluster nodes. (Which may not have network connections.)
- Idea: Reduce the problem to something we already know how to do: *Workflow!* But how do we bootstrap the workflow software? *With the builder!*

# Bootstrapping Workflows and Apps





# Controlling Cluster Size

# Dynamic Cluster Resizing

---



VC3 Supports several mechanisms for setting the number of workers in a given virtual cluster.

- Set static size for cluster upon creation.
- Manually change cluster size via web interface. The provisioning factory automatically adds or removes worker jobs at resources.
- (For HTCondor cluster) Scale the cluster dynamically based on idle jobs on the head node.

# Current Resizing Cases



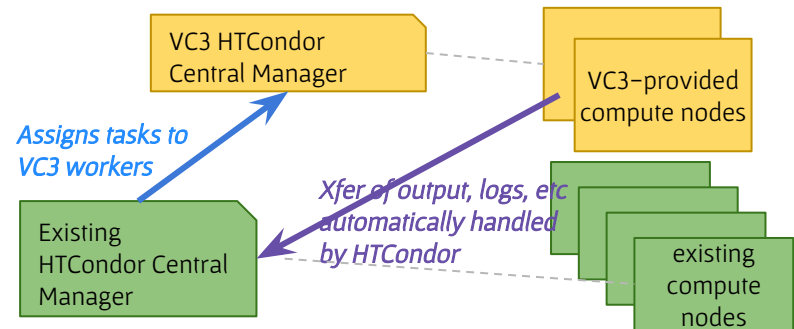
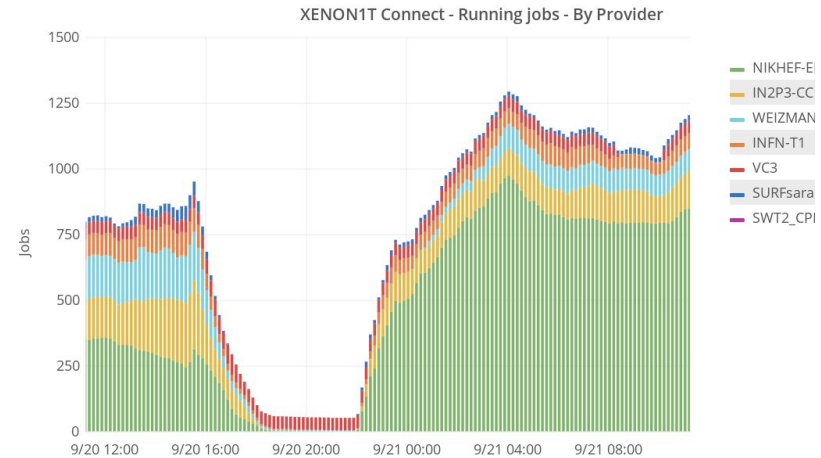
		Job Removal Setting	
		Peaceful = True	Peaceful = False
Source of desired number of workers	User (via web interface)	<p><b>Increasing:</b> Works as expected.</p> <p><b>Decreasing:</b> Workers will terminate when they have finished current job. By default, the younger workers go away first.</p>	<p><b>Decreasing:</b> Running jobs may be killed. But cluster can be reduced rapidly.</p>
	VC3 (automatic, dynamic, based on user's job pressure on head node.)	<p><b>Increasing:</b> Works as expected.</p> <p><b>Decreasing:</b> Workers will terminate when they have finished current job. By default, the younger workers go away first.</p>	



# Scaling Out Production Clusters



- Collaborations who have existing HTCondor pools can extend them by adding more worker nodes via VC3
- Add XSEDE resources, Open Science Grid, and campus HPC clusters
- End-users can transparently use additional resources



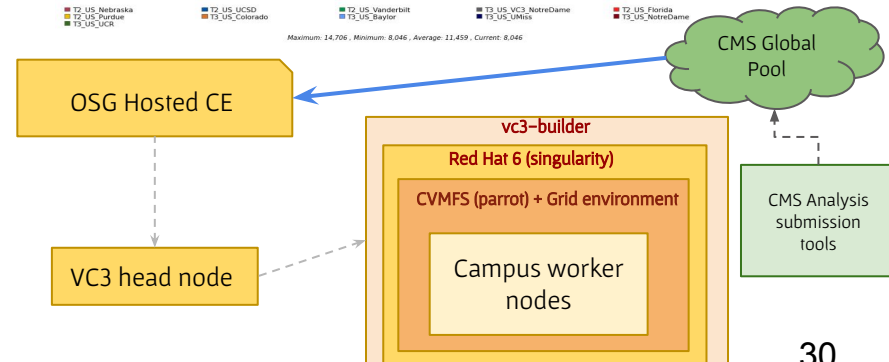
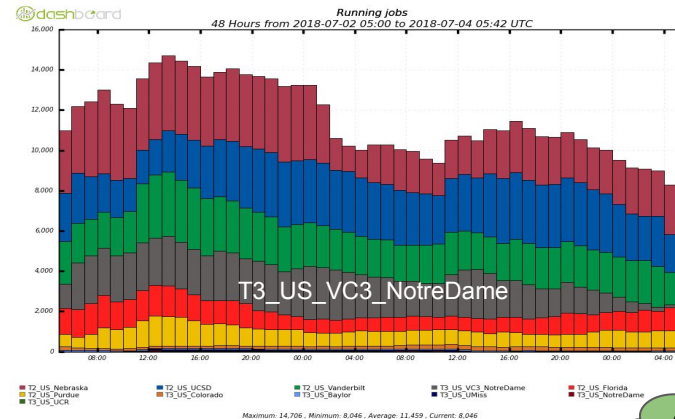
# Building Tier 3s on top of campus resources



- Although different computing resources are often available at universities, meeting all requirements to deploy a valid Tier 3 able to run CMS workflows on the grid is not an easy task to achieve without the intervention of a system administrator with root access.
- VC3 allows the provisioning at user-level of:
  - The CERN File System (CVMFS) (via parrot)
  - The OSG grid environment on the worker nodes (via CVMFS)
  - Customized Operating Systems (via singularity)
- The OSG Compute Element (CE) is then integrated with the VC3 submit host, allowing the creation of a CMS Tier 3 using Notre Dame opportunistic campus resources without any root access level.

## CMS Analysis activity per Tier Site

Note T3\_US\_VC3\_NotreDame performing at the scale of the Tier 2s!



# Examples of clustering middleware

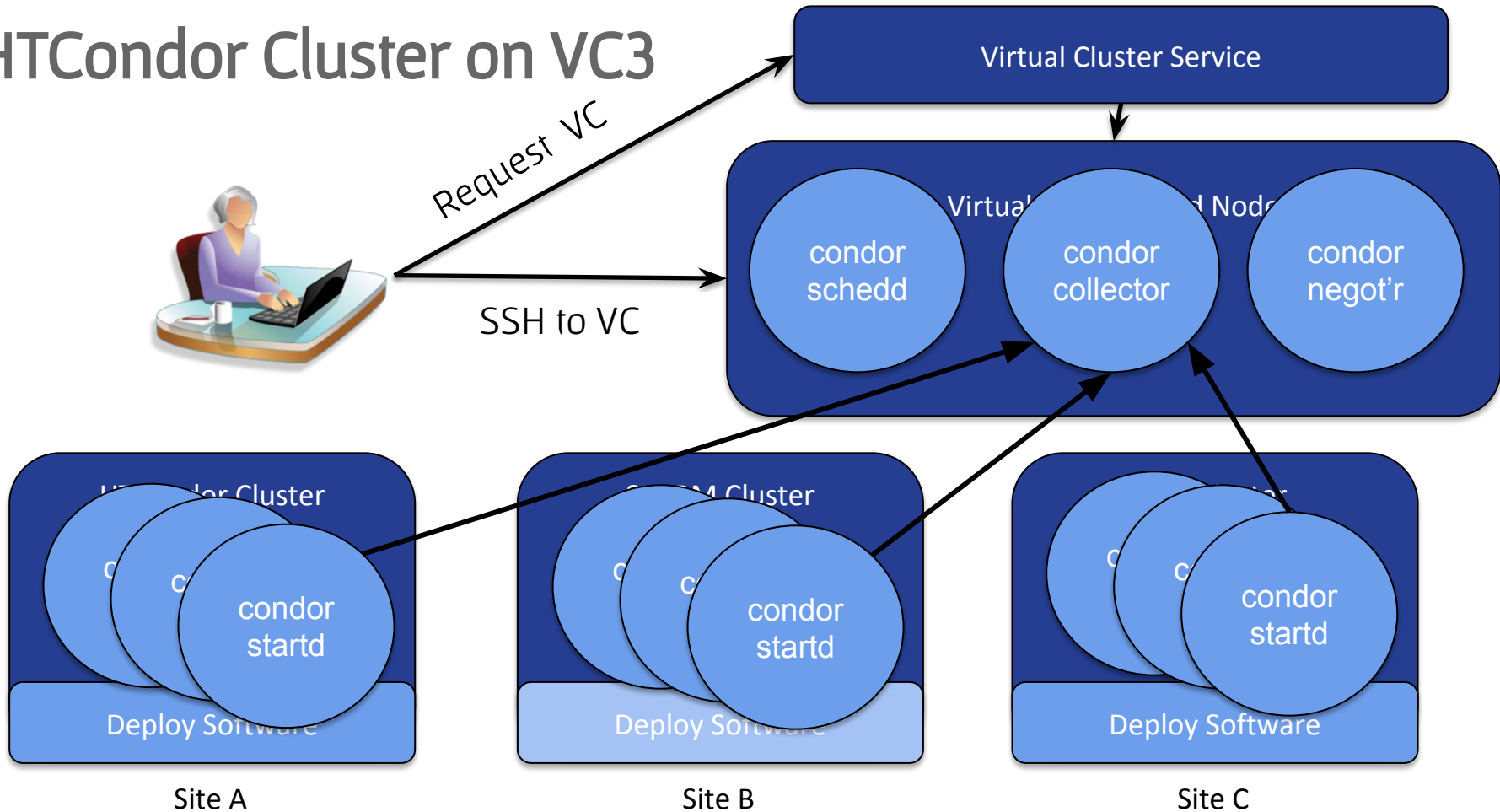
# Use Cases

---

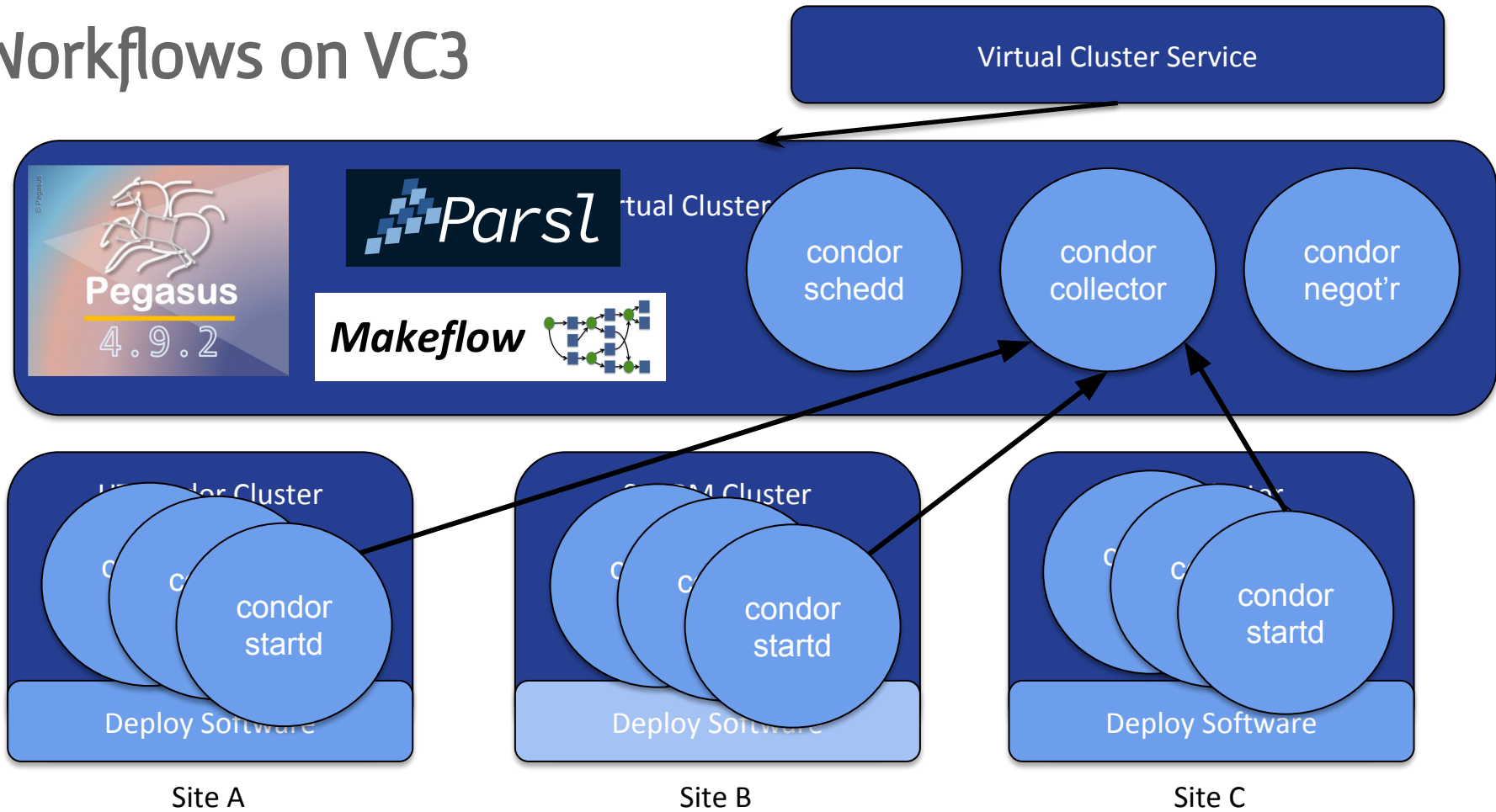


- HTCondor Pool
  - Already demonstrated self-service, fully automated provisioning clusters, binding allocations into projects
- Workflow Technologies
  - Makeflow + Work Queue
  - Pegasus + HTCondor,
  - Parsl + iPyParallel
- Spark
  - **D**eployment of a Spark cluster for data analysis on top of existing schedulers
- JupyterLab
  - **P**opular **i**nteractive analytics environment – provisionable by VC3
- SCAILFIN + REANA
  - Complex set of REANA services deployed on minikube on head node + HTCondor on Stampede, Blue Waters and PSC.
- KOTO
  - Helping a new collaboration with no establishes resources to run on BNL HPC, KEK HPC + OSG HTC + campus resources.

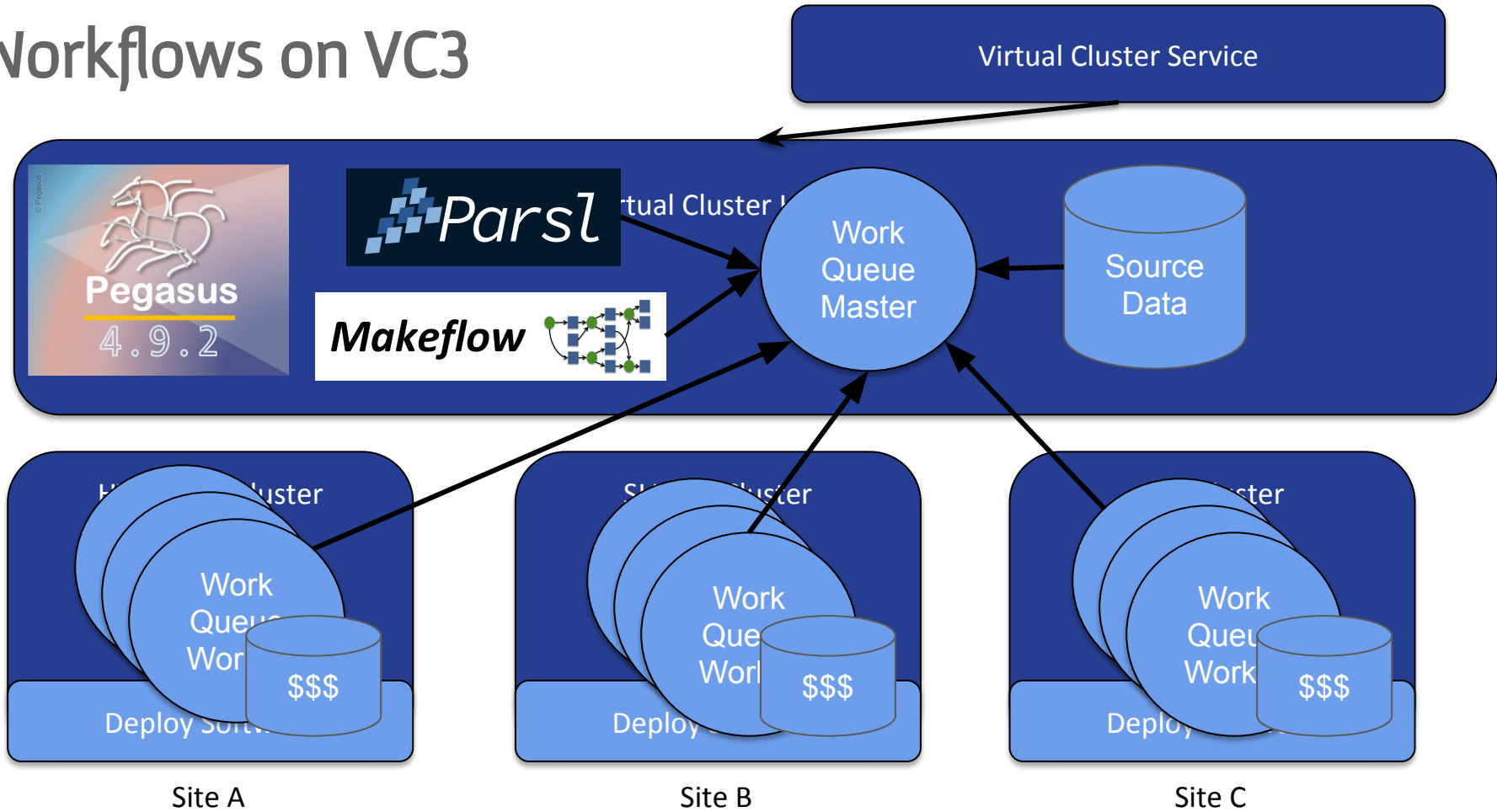
# HTCondor Cluster on VC3



# Workflows on VC3



# Workflows on VC3



# Spark Middleware Integration

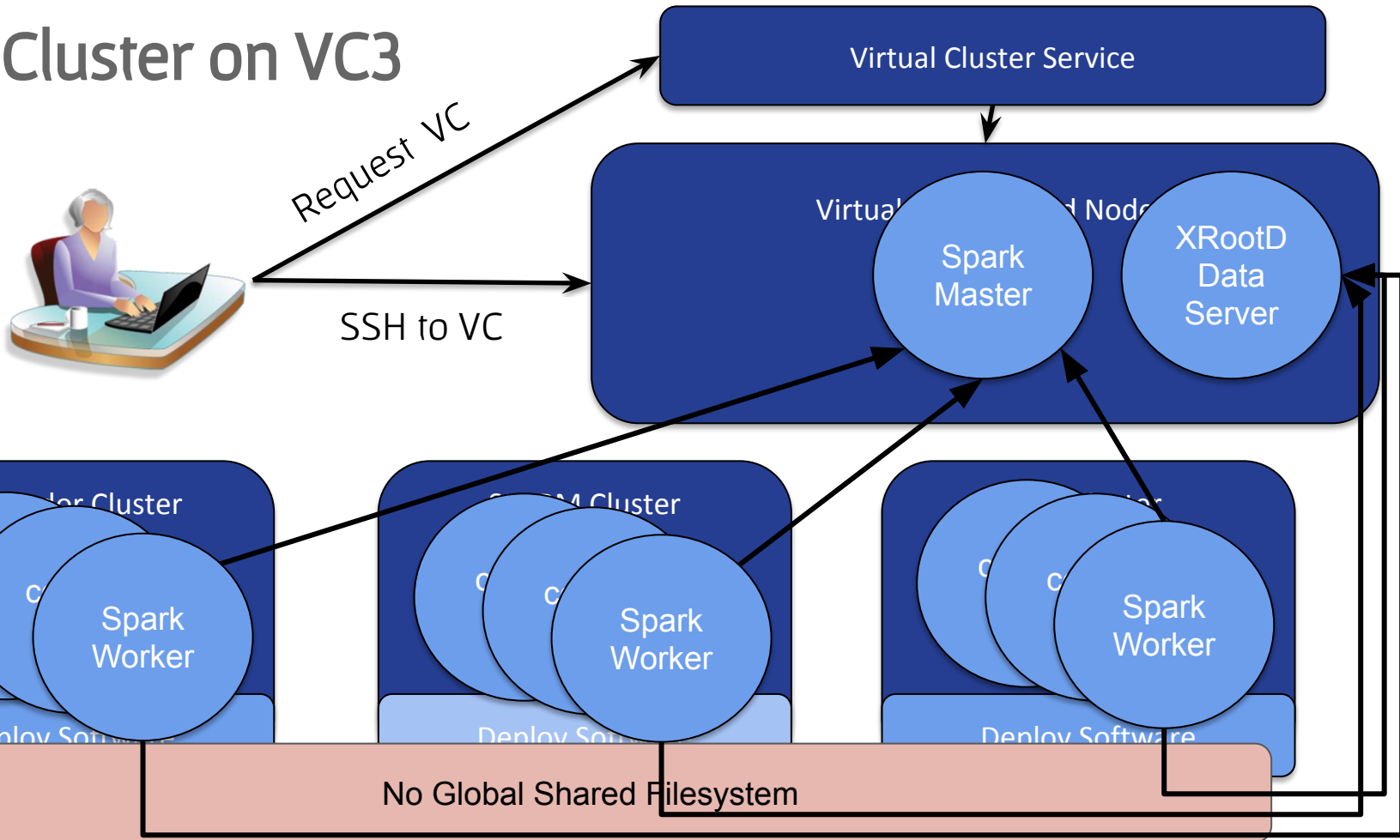
---



- Apache Spark now a first-class supported middleware in VC3
- Spark master runs on the virtual cluster head-node
- Virtual cluster workers are spark slaves
  - Java JRE, spark, scala, pyspark are deployed as-needed by the target resource.
- A shared secret secures connection to the master from workers and application clients
- No shared filesystem, use of an IO driver is needed.



# Spark Cluster on VC3



# Spark & CMS Analysis

---



- Late-stage custom analysis code processing a Mini-AOD file to produce a histogram.
- Application written in python, using pyspark
- Data IO with the spark-xrootd plugin
  - Automatically installed in headnode and workers
- Ran with 20 workers, 4 cores each.
- Maximum concurrency was 80 tasks.

# Spark & CMS Analysis



## Spark Master at spark://128.135.158.221:7077

URL: spark://128.135.158.221:7077  
REST URL: spark://128.135.158.221:8066 (cluster mode)  
Alive Workers: 17  
Cores in use: 68 Total, 68 Used  
Memory in use: 895.3 GB Total, 17.0 GB Used  
Applications: 1 Running, 0 Completed  
Drivers: 0 Running, 0 Completed  
Status: ALIVE

### Workers

#### Worker Id

worker-201809191231112-10.32.83.11-44930  
worker-20180919123157-10.32.79.15-40509  
worker-20180919123157-10.32.80.39-36998  
worker-20180919123209-10.32.79.65-42856  
worker-20180919123211-10.32.79.65-41106  
worker-20180919123220-10.32.79.45-44809  
worker-20180919123232-10.32.79.43-39239  
worker-20180919123233-10.32.80.47-45735  
worker-20180919123245-10.32.72.4-36789  
worker-20180919123250-10.32.77.80-44497  
worker-20180919123250-10.32.77.80-46283  
worker-20180919123307-10.32.76.62-41504  
worker-20180919123346-10.32.80.71-37519  
worker-20180919123414-10.32.76.64-35737  
worker-20180919123414-10.32.76.64-36614  
worker-20180919123456-10.32.76.74-40008  
worker-20180919123458-10.32.76.74-44468

### Running Applications

Application ID	Name
app-20180919163039-0000	Zpeak_NanoAOD-SF (kill) Zpeak_NanoAOD-SF

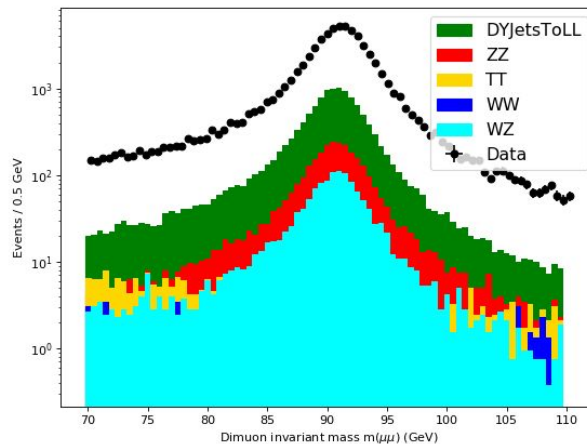
## Virtual Cluster: btovar-zpeak

Terminate Cluster

### STATE OF VIRTUAL CLUSTER



All requested compute workers are running.



### Status

BTOVAR-ZPEAK  
Cluster Framework:

Requested 20  
Running 20  
Queued 0  
Error 0

Resize Workers

Duration  
4.5 min

# Provisioning JupyterLab Notebooks

---



- JupyterLab-based head nodes now launchable in VC3 development branch
- Web-based notebook interface for Python
- Uses Globus identity plugin to login
- Integration with HTCondor and other middleware types

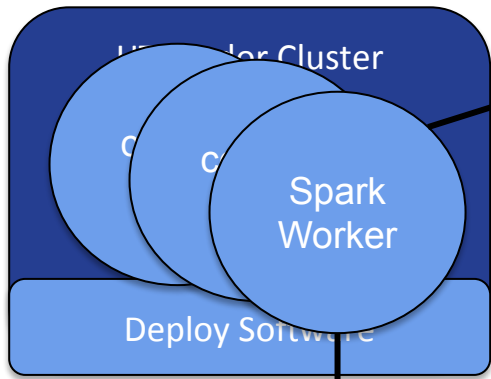
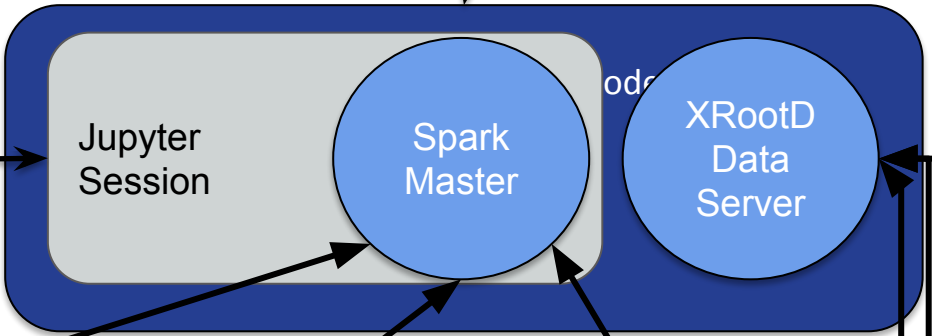
# Jupyter + Spark on VC3



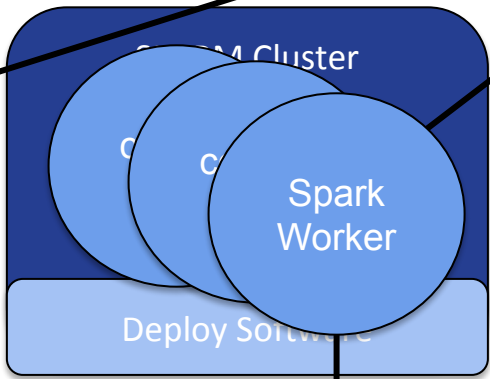
Request VC

WWW Access

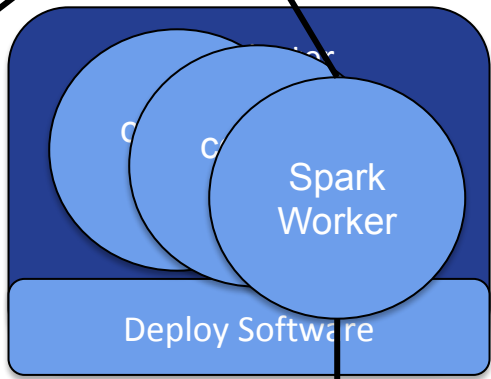
Virtual Cluster Service



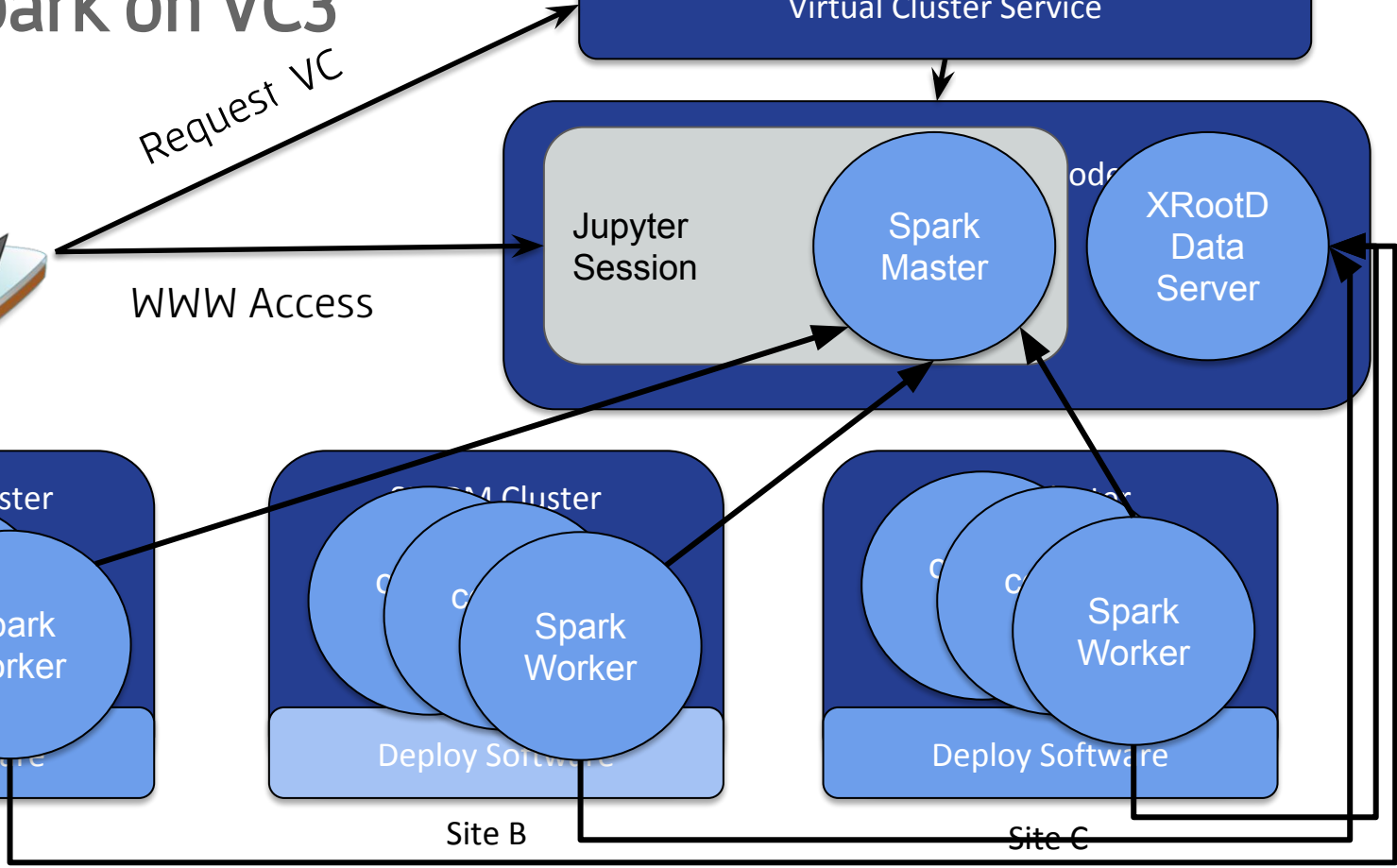
Site A



Site B



Site C

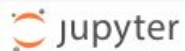


# Provisioning JupyterLab Notebooks



My Virtual Clusters

Name	Project	Head Node	Workers
btovar-chiroptera3	btovar	128.135.158.235	<div style="display: flex; flex-direction: column; align-items: center;"><div style="margin-bottom: 2px;">1 Requested</div><div style="margin-bottom: 2px;">1 Running</div><div style="margin-bottom: 2px;">0 Queued</div><div style="margin-bottom: 2px;">0 Error</div></div>



Files

Running

Clusters

Select items to perform actions on them.

- 0
- 07-Kalman-Filter-Math.ipynb
- Balancing Reactions.ipynb

Home x Balancing Reactions x

Not secure | <https://128.135.158.235:8080/user/btovar/notebooks/Balancing%20Reactions.ipynb>

jupyter Balancing Reactions (autosaved) Logout Control Panel

File Edit View Insert Cell Kernel Help Trusted Python 3

equation that determines the stoichiometric coefficient for a particular chemical species. In this case, since this example refers to the combustion of methane, an obvious basis is to set the stoichiometric coefficient of methane to -1.

```
In [3]: basis = [sympy.Eq(vCH4, -1)]

for eqn in atomBalances + basis:
    print(eqn)

sympy.solve(atomBalances + basis)

Eq(vCH4 + vC02, 0)
Eq(4*vCH4 + 2*vH2O, 0)
Eq(2*vC02 + vH2O + 2*vO2, 0)
Eq(vCH4, -1)

Out[3]: {vCH4: -1, vC02: 1, vH2O: 2, vO2: -2}
```

**Example: Hypergolic Reactions**  
Hypergolic reactions are reactions where the reactants spontaneously ignite. They are frequently used in space propulsion where it desired to have a thruster that can be readily controlled over a range of operating conditions.

# Credential Delegation



Request VC

WWW Access

Virtual Cluster Service

Jupyter Session

Spark Master

XRootD Data Server

## Credential Delegation:

- Globus Auth to VC3
- VC3 WWW to Jupyter
- VC3 WWW to Site SSH
- User to Jupyter
- Worker to Master
- Worker to XRootD

Each one a different technology!

Site A

Site B

Site C

# Scalable cyberinfrastructure on HPC facilities



## SCAILFIN: Scalable CyberInfrastructure for Artificial Intelligence and Likelihood Free Inference

The SCAILFIN project aims to deploy artificial intelligence and likelihood-free inference techniques and software using scalable cyberinfrastructure (CI) that is developed to be integrated into existing CI elements, such as the

REANA system, to work on HPC facilities.



**reana**

Reproducible research data analysis platform

Launch New Virtual Cluster

Project: bwtest1

\* - INDICATES REQUIRED FIELD

VIRTUAL CLUSTER NAME (A-Z, 0-9, \_ AND -)\*

reanabwv1

CLUSTER TEMPLATE FRAMEWORK \*

REANA+HTCondor

NUMBER OF COMPUTE WORKERS: \*

2



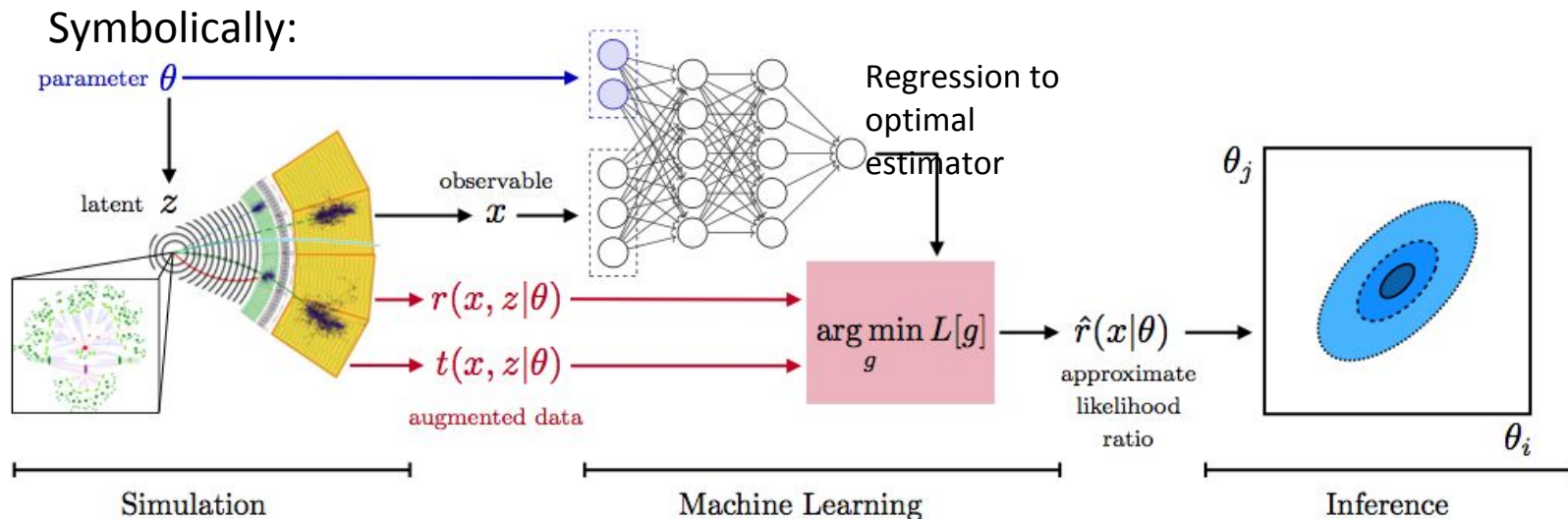
VC3 headnode

```
(reana) [khurtado@khurtado-reanabwv1 ~]$ reana-cluster status
COMPONENT STATUS
job-controller Running
server Running
db Running
workflow-controller Running
message-broker Running
REANA cluster is ready.
(reana) [khurtado@khurtado-reanabwv1 ~]$ kubectl get pods
NAME READY STATUS RESTARTS
AGE
batch-serial-7e79ee48-836f-4049-87ee-a3dc66d8a1da-t17zd 0/1 Completed 0
5h54m
db-69744557df-wg4mt 1/1 Running 0
5h55m
job-controller-5c7f4c8b4f-sgnj6 1/1 Running 0
5h55m
message-broker-b7d66cf55-m9p4n 1/1 Running 0
5h55m
server-58dc985c77-n2qpn 2/2 Running 0
5h55m
workflow-controller-668f69d4bc-x62w7 2/2 Running 0
5h55m
```

REANA is automatically deployed and integrated with HTCondor in VC3

REANA components are started via kubernetes(minikube)

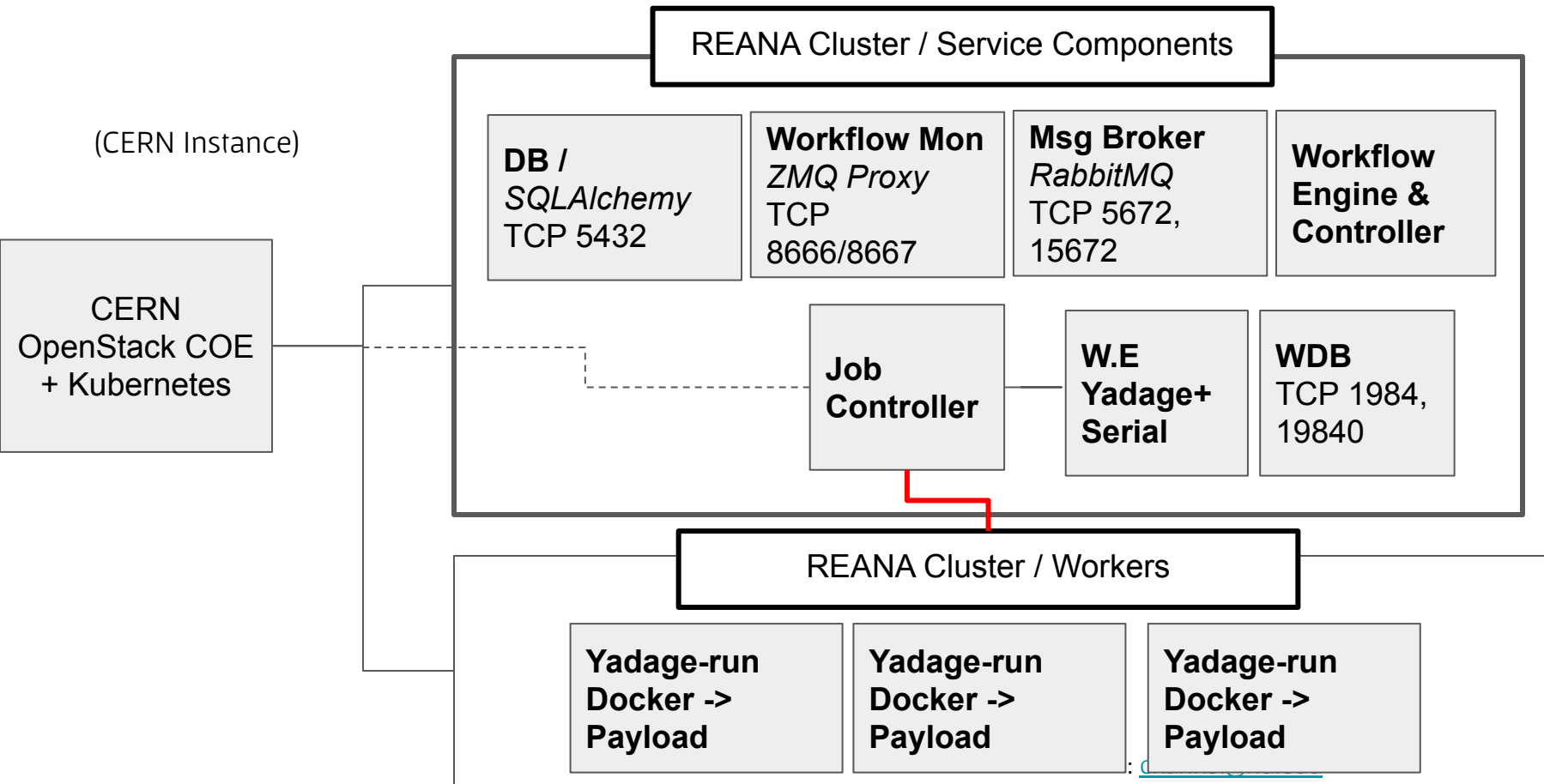




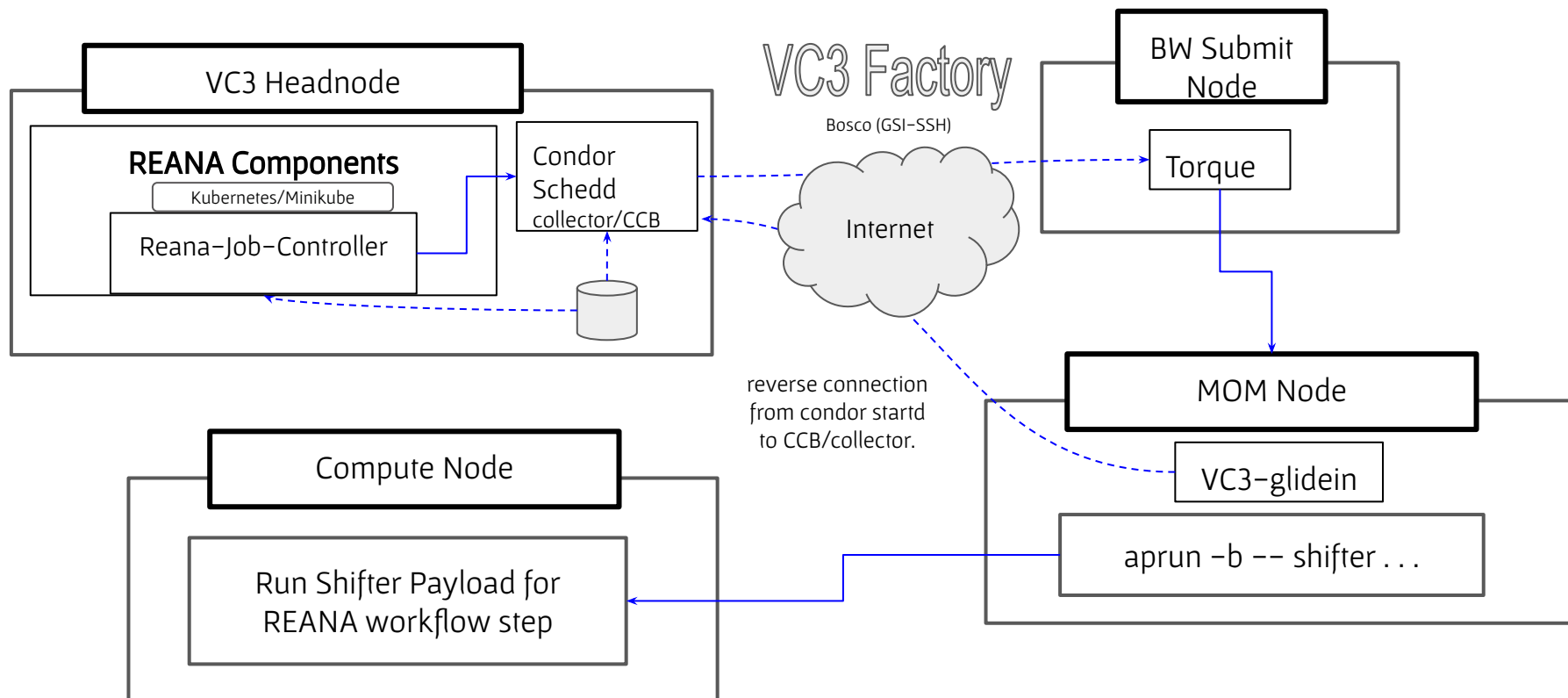
Estimation of optimal estimator lends itself to ML methods:

- Training data derived from simulations
- Can be guided by optimal sampling based on phase space density of generator, sensitivity to physics under study

# reana-cluster - Simplified Diagram



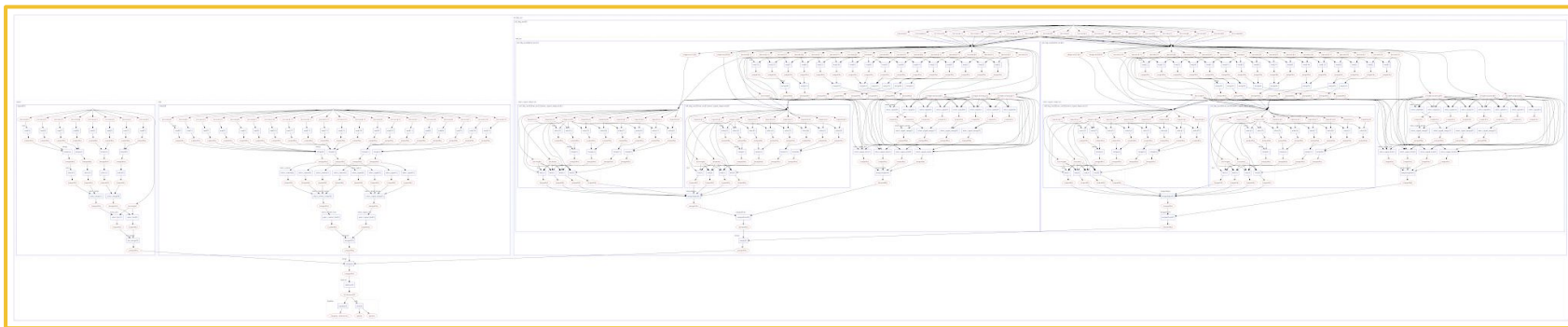
# SCAILFIN on Blue Waters via VC3



# SCAILFIN on Blue Waters via VC3



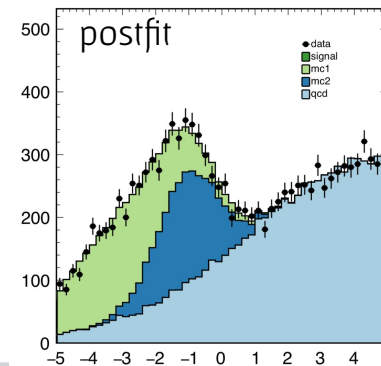
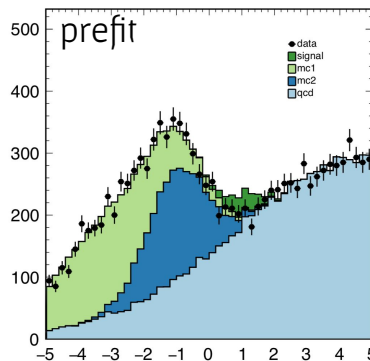
Complex Beyond Standard Model example (ran at BW via REANA + HTCONDO Rvia VC3)



Data is generated/emulated according to Standard Model expectations.

After processing, a statistical model involving both signal and control regions is built and the model is fitted against the observed data.

The signal sample is scaled down significantly to fit the data, which is expected since the data was emulated in accordance with a SM-only scenario



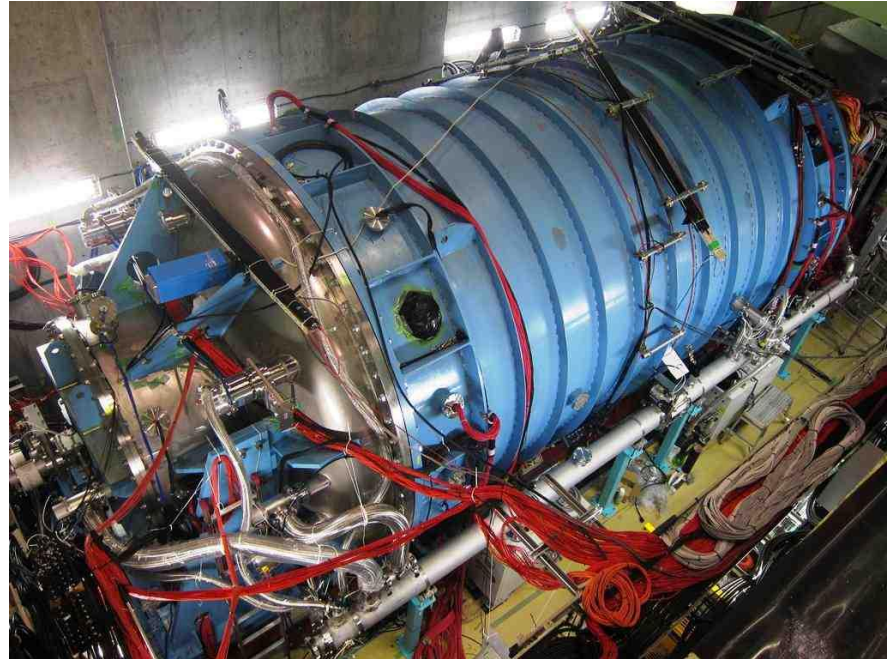
# KOTO experiment

---



High energy physics experiment at the J-PARC (Japan Proton Accelerator Research Complex) Facility in Japan - <http://koto.kek.jp/>

Experiment measures the decay rate of neutral K-mesons (kaons) into neutral pi-mesons and a pair of neutrinos



# Requirements

---



**Storage:** Estimated storage for accommodating the MC and MC/Analysis portions of the pipeline for a full experiment cycle is in the order of 200 TB per experiment cycle

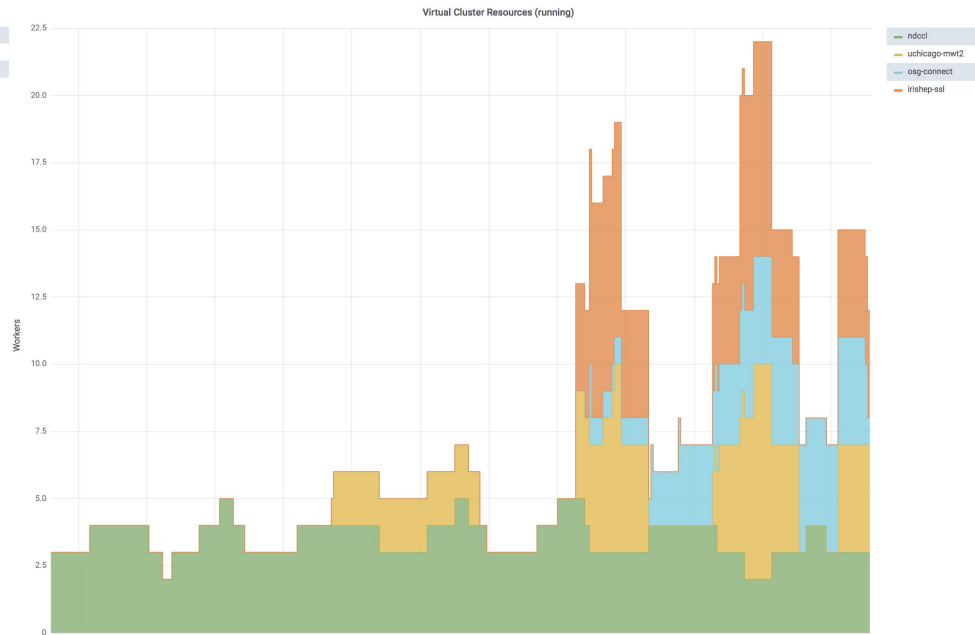
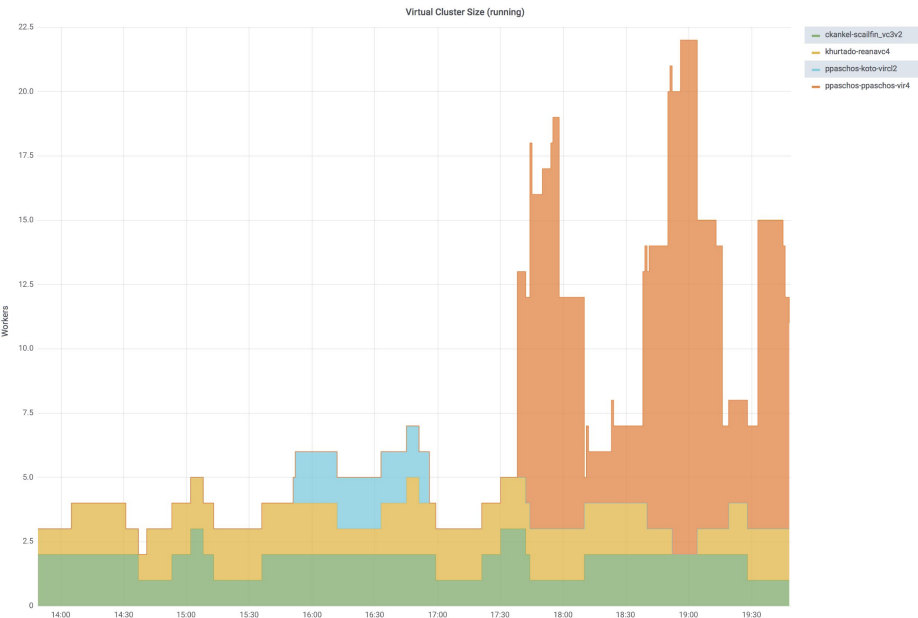
**Software stack:** There are two applications running on the KEKCC machine that needed to be rebuilt on the OSG side under RHEL7. A KOTO GEANT package and an analysis collection of tools. Most of the effort so far is to provide the environment for the software to be built and run on remote OSG sites

**Submission scripts:** KEKCC uses the LSF scheduler (bsub). Submission scripts needed to be modified to HTCondor

# VC3 job submission for KOTO



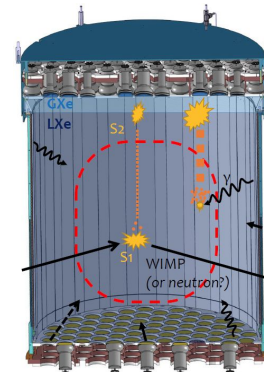
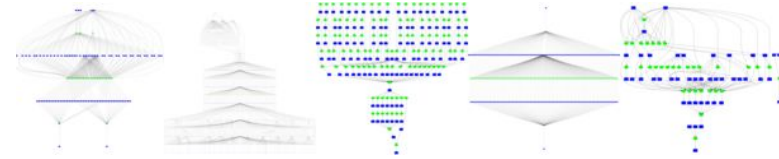
- OSG Connect (blue), UC (yellow), IRIS-HEP-SSL (orange)
- VC3 evenly distributes the load across resources



# Working **Middleware** and **Applications**



- Various Bioinformatics Workflows
  - Makeflow + **HTCondor** + BWA, Shrimp, BLAST.
- Lobster CMS Data Analysis
  - **Work Queue** + Builder + CVMFS
- Lifemapper
  - **Work Queue** + Builder
- **Spark** CMS Data Analysis
- 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**





# Thoughts and Lessons Learned

# Successful Ideas that Worked

---



- Automated Provisioning of Cluster Frameworks Across Heterogeneous HPC Resources
- Automated Deployment of User-Level Software
  - (Build itself may require substantial cluster resources)
- Internal Architecture for Robust Service Deployment
  - (Surprisingly complex state machine.)
- Multi-Layer Configuration Enables Collaborations
  - Credentials go deep into the framework

# Things That were Harder than Expected

---



- New Security Requirements (MFA) Spread Very Quickly
  - Working Idea: Let user login and “pull” the cluster blueprint to the local site where it can be deployed.
- Difficulty of Curating Site/Cluster Details due to Churn
  - Choice: Either invest more in active curation, or involve users in the configuration process.
- Debugging Complex Service Deployments
  - Working Idea: Test before using, expose stages of success to user, give concrete feedback for fixing.

# Portable Ideas to Keep and Use Elsewhere

---



- Importance of the Head Node as Infrastructure
  - Allocatable, Personal, Shareable, Configurable
- Self-Service Deployment Model
  - Generalization of Infrastructure-as-Code:
  - OpSys + Software + Middleware + Entry Point
- A Model for Sharing Institutional Resources
  - Classic: Submit My Work to Remote Environments
  - VC3: Deploy My Environment on Remote Resources

# Technology Evolution 2016–2020

---



## Evolution from VMs to Containers:

- Swap out OpenStack in favor of Kubernetes, which is lightweight and better supports process automation.
- Docker is being supplanted by Singularity, and other container environments.

## Evolution from credentials to capabilities:

- ssh-keygen → Upload Pubkey → RSA-SSH (Standalone)
- Globus Auth → GSI Token → GSI-SSH (DCDE Fed.)
- Globus Auth → Web Tokens → WebTokens-SSH (Web Native)
- SciTokens? ([scitokens.org](http://scitokens.org))

# Towards Federated Ops with SLATE



- Remotely manage edge services at sites by **expert teams** from **trusted organizations**
- Deploy updates more quickly & introduce new services more easily
- Save time and effort for local site admins -- towards **OSG NoOps**
- Edge federation via lightweight server/client overlay using **Kubernetes**, the industry leading container orchestration platform
- Software catalog, with push button deploy using vetted **Helm** charts



```
$ slate instance list
$ slate instance delete <instance name>
$ slate app install --group atlas-xcache --cluster uchicago-prod
--conf MWT2.yaml1 xcache
```

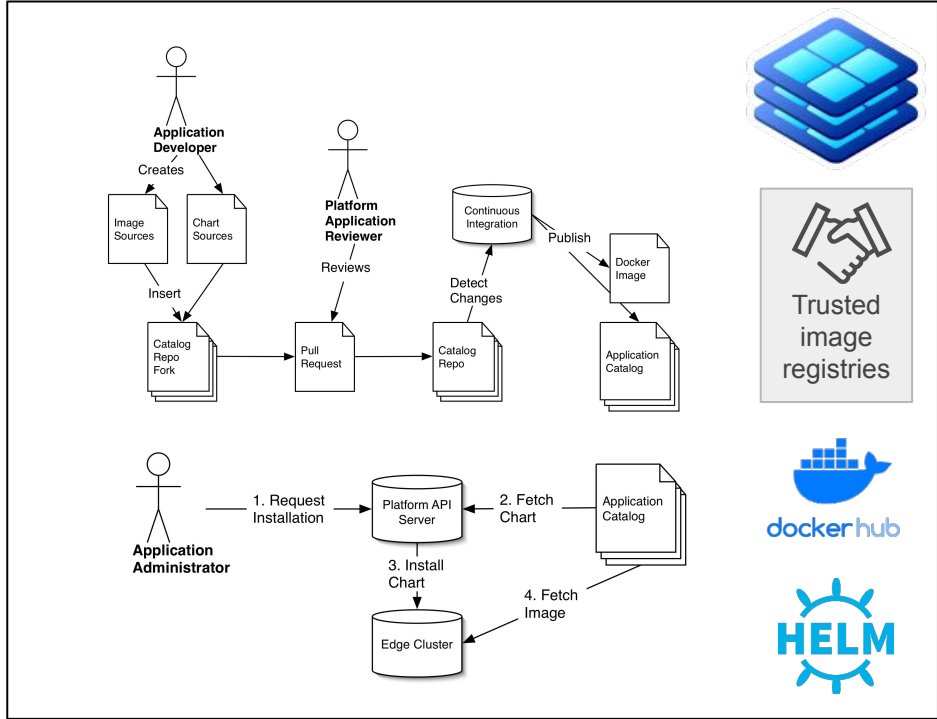
Overall System Load

Average Load

Last 3 hours

- uchicago-prod
- uchicago-river
- uchicago-test
- umich-prod
- usdsh-prod

- [SLATE](#) - a value added K8s distribution
  - Support for CVMFS, ingress controller (multi-tenant, scoped privileges), Prometheus monitoring, **curated application catalog w/ Jenkins CI**
- Site security & policy conscious
  - SLATE works as an unprivileged user
  - Single entrypoint via **institutional identity**
  - Site owner controls group whitelists & service apps; **retains full control**
- With OSG, WLCG, [trustedci.org](#) & others working to establish a "CISO compliant" security posture and **new trust delegation model**



SLATE creates secrets and XCache deployment on cluster

XCache Container Download

Kubernetes objects instantiated

Pod starts up, registers itself in AGIS



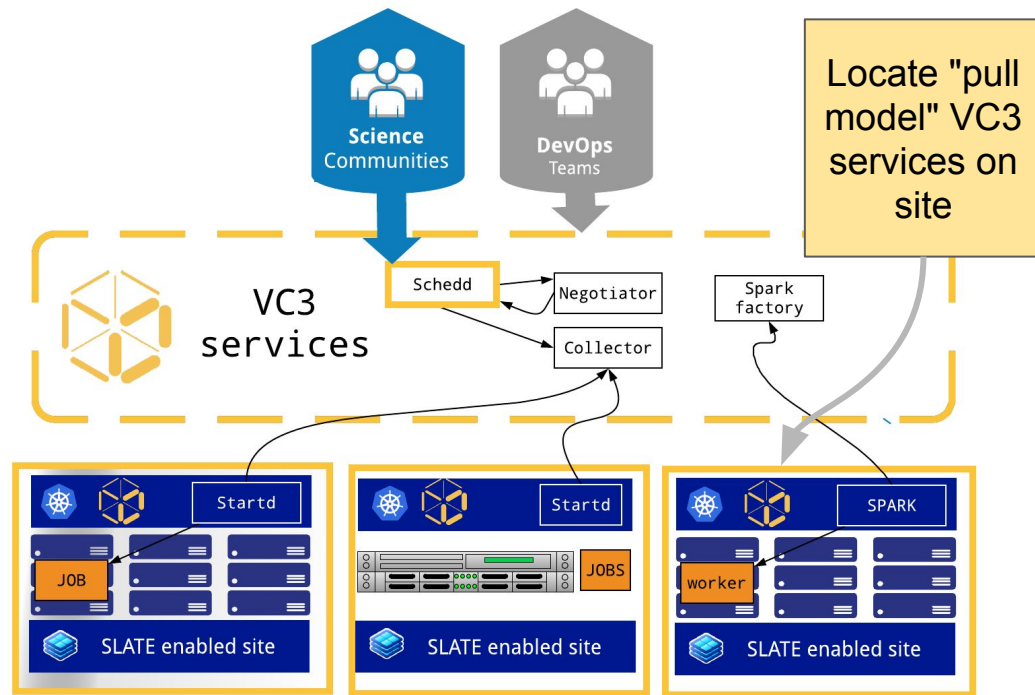
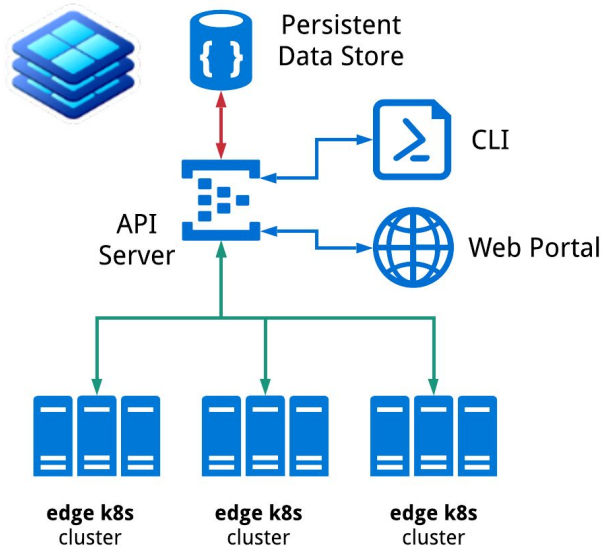
Upgrades are as simple as re-deploying.

**A data caching network deployed in less than 20 minutes.**

# VC3 and SLATE synergies



## SLATE: federated service orchestration in the SciDMZ





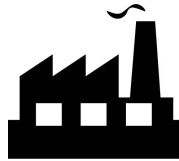
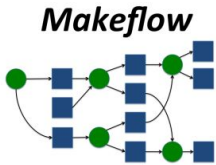
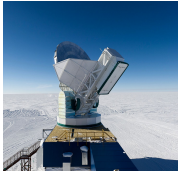
# Open Problems in Federated Systems

---



- Troubleshooting Across Systems
  - Many stakeholders, system churn, lack of evidence.
  - Troubleshooting posed as a distributed database query?
- State Management and Garbage Collection
  - Remote creation of containers for local sites, how long must they be kept around?
- Socio-Technical Understanding of Delegation
  - User X submits jobs to user Y's account on site S. Ok?
  - Technical ability to represent non-local account?

# Collaborators and Connections



AutoPyFactory



Science Gateways  
Community Institute



Open Science Grid



# VC3

## Virtual Clusters for Community Computation

<https://www.virtualclusters.org>

@virtualclusters

New users signup: <http://bit.ly/vc3-signup>

Register your HPC: <http://bit.ly/vc3-new-resource>



# VC3 Funding and Team



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

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

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

dev-team: Lincoln Bryant (UC), Benedikt Riedel (UC), Suchandra Thapa (UC), Jeremy Van (UC), Kenyi Hurtado Anampa (ND), Ben Tovar (ND), Jose Caballero Bejar (BNL).



U.S. DEPARTMENT OF  
**ENERGY**

Office of  
Science

Supported by the Department of Energy Office of Advanced Scientific Computing Research and Next Generation Networking Services, Solicitation DE-FOA 0001344 (DDRM), Proposal 0000219942  
**Rich Carlson, Program Manager**