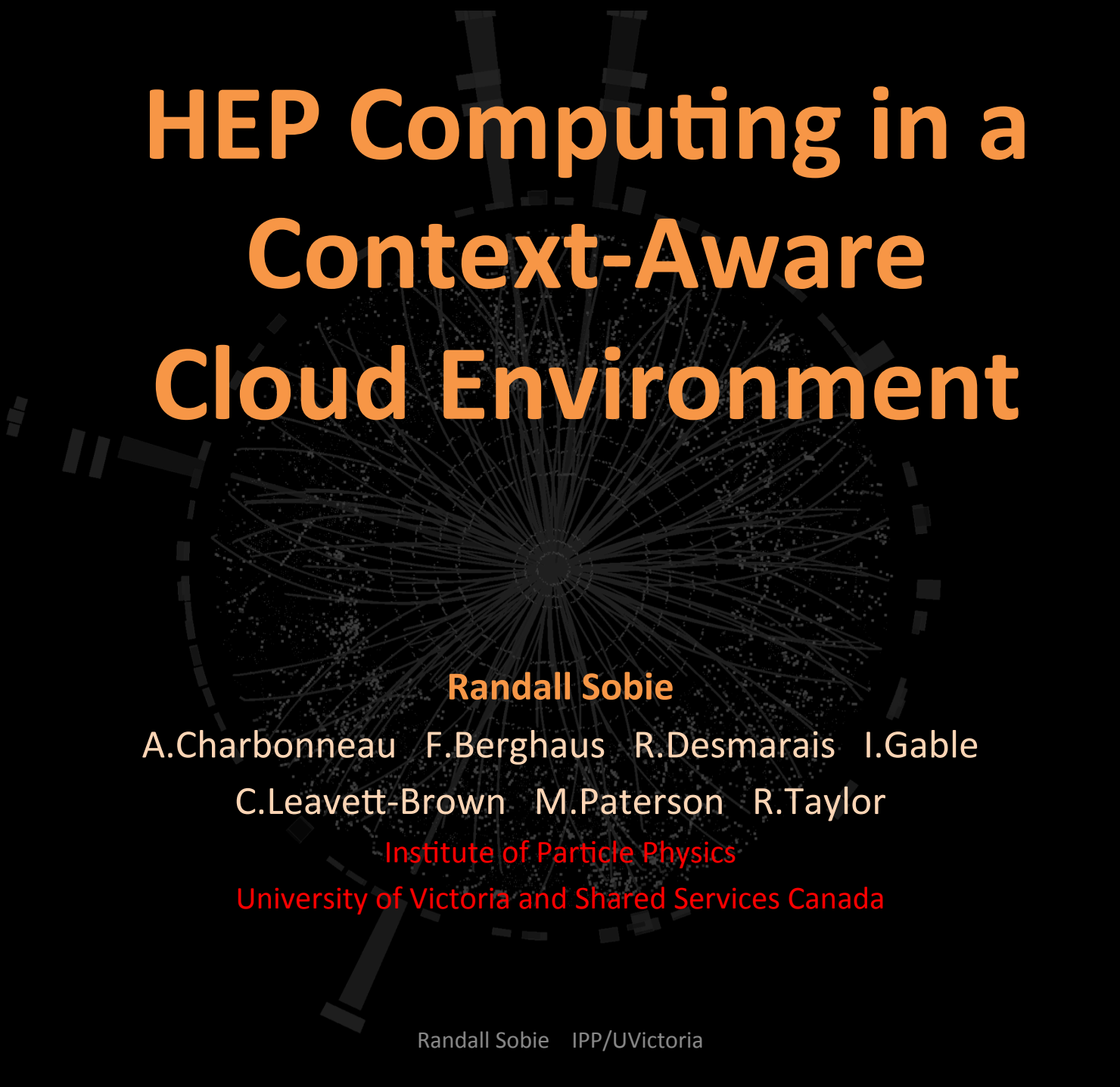


HEP Computing in a Context-Aware Cloud Environment



Randall Sobie

A.Charbonneau F.Berghaus R.Desmarais I.Gable

C.Leavett-Brown M.Paterson R.Taylor

Institute of Particle Physics

University of Victoria and Shared Services Canada

Motivation



Science is becoming a global endeavour

Particle physics (HEP) requires large machines that can not be funded by a single country

Use computing resources distributed around the world to analyze the vast data sets



CERN Laboratory in Geneva

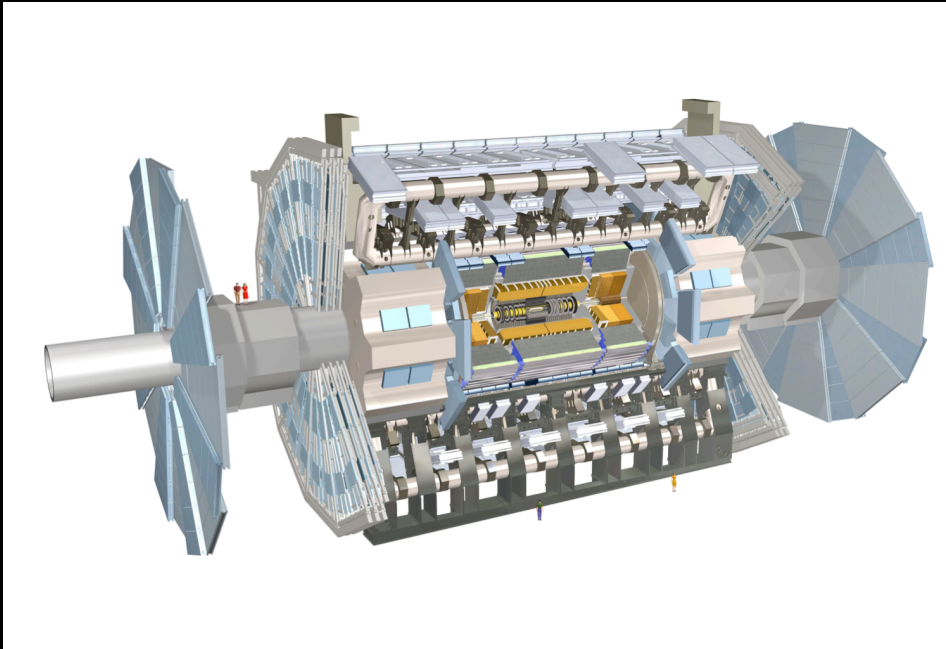
Largest man-made machine in the world

27 km circumference

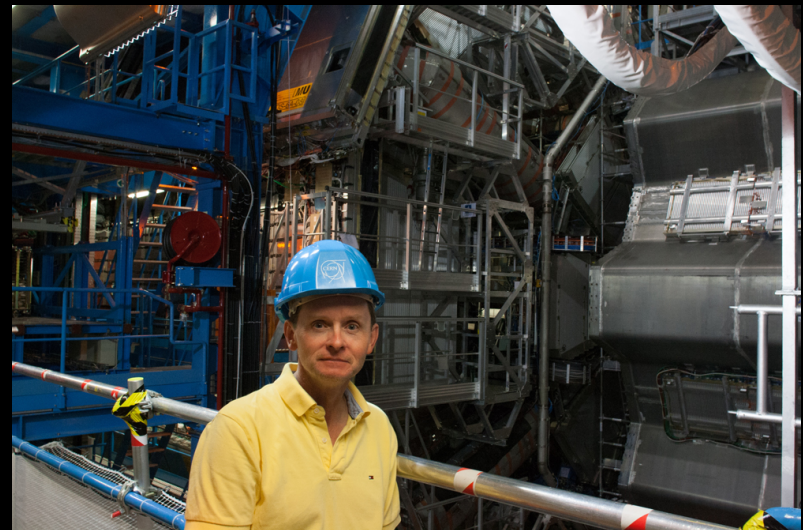
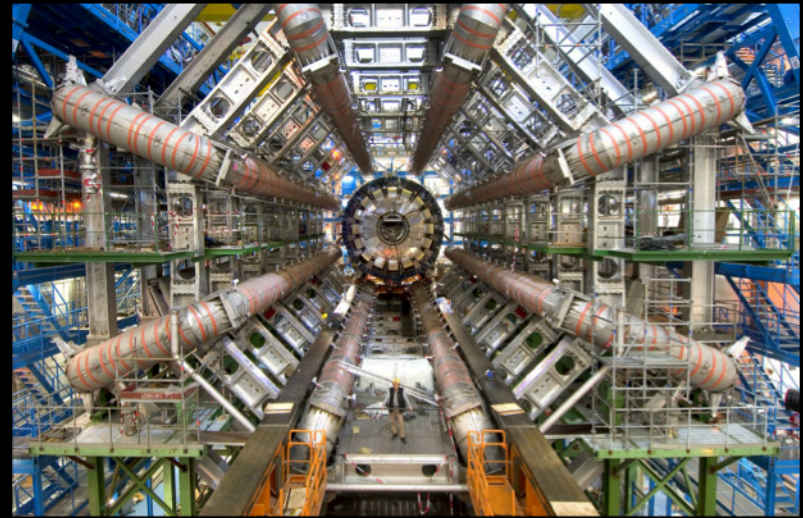
Colliding protons-on-protons at very high energy

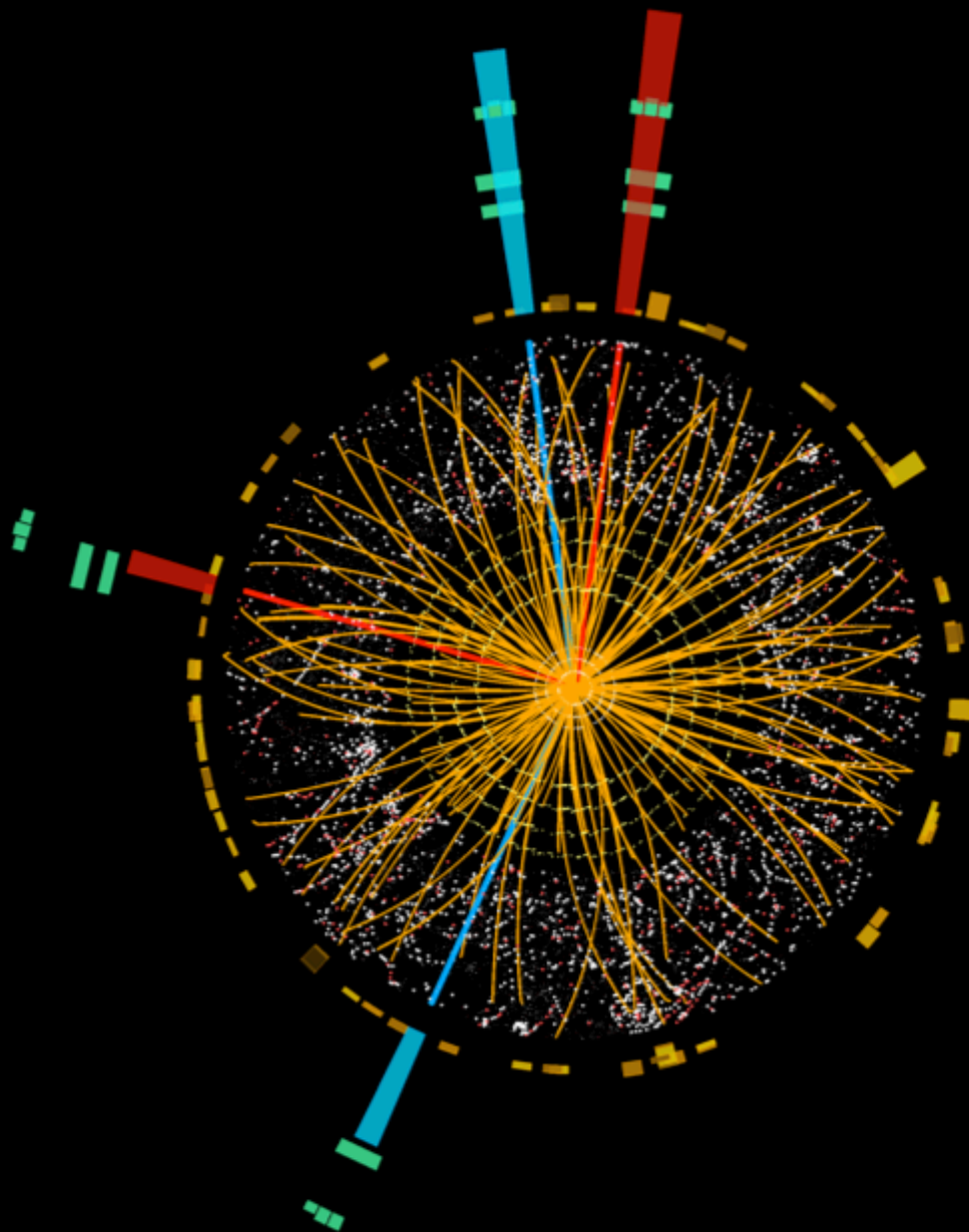
Goal is to find the Higgs boson, dark matter or other new physics particles

ATLAS Detector



44 m wide 25 m diameter





Event image (Higgs candidate)

In the plane perpendicular to the proton beams

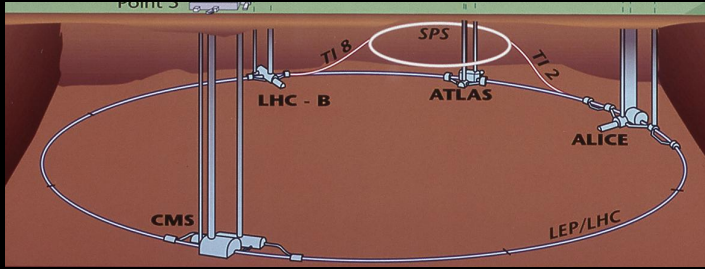
Yellow tracks are charged particles

Curvature indicates sign of the charge

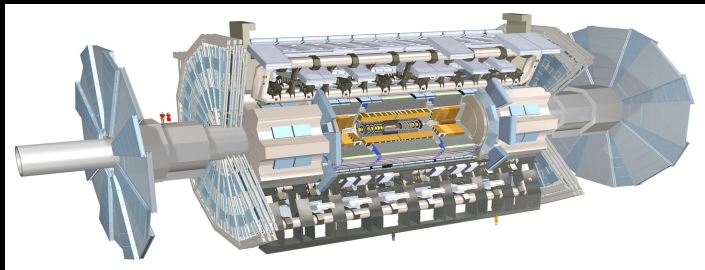
Coloured boxes indicate energy deposition

Each "event" is about 2 MB

Collect 200 events per second



40 million collisions per second



100,000 collisions selected



200 events per second

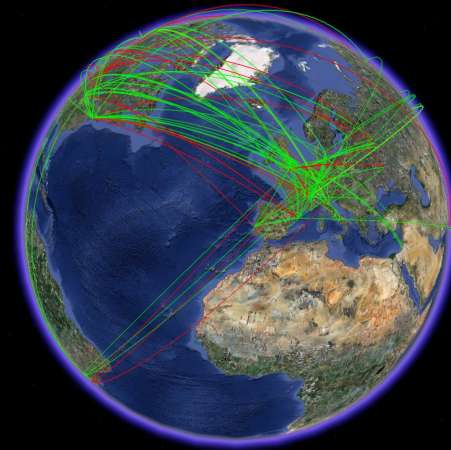
WLCG Computing Grid

CERN Tier 0

10 Tier1 sites

60+ Tier2 sites

150 PB data



Clouds in High Energy Physics

Long-term preservation of
software and data of HEP
experiments

Utilize special computing
resources attached to the
detectors

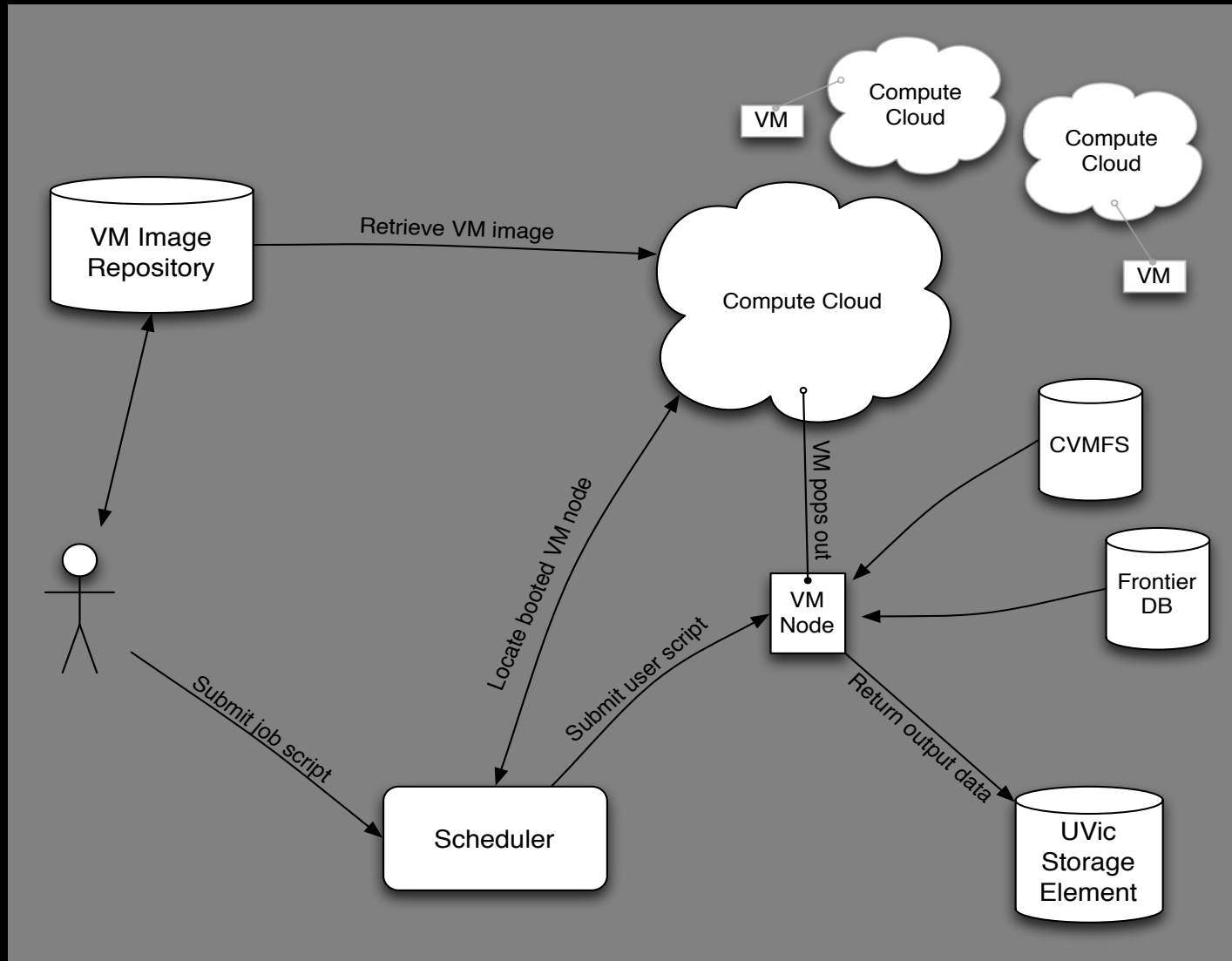
Simplify the management of
heterogeneous in-house
resources

Distributed cloud computing
using HEP and non-HEP
clouds

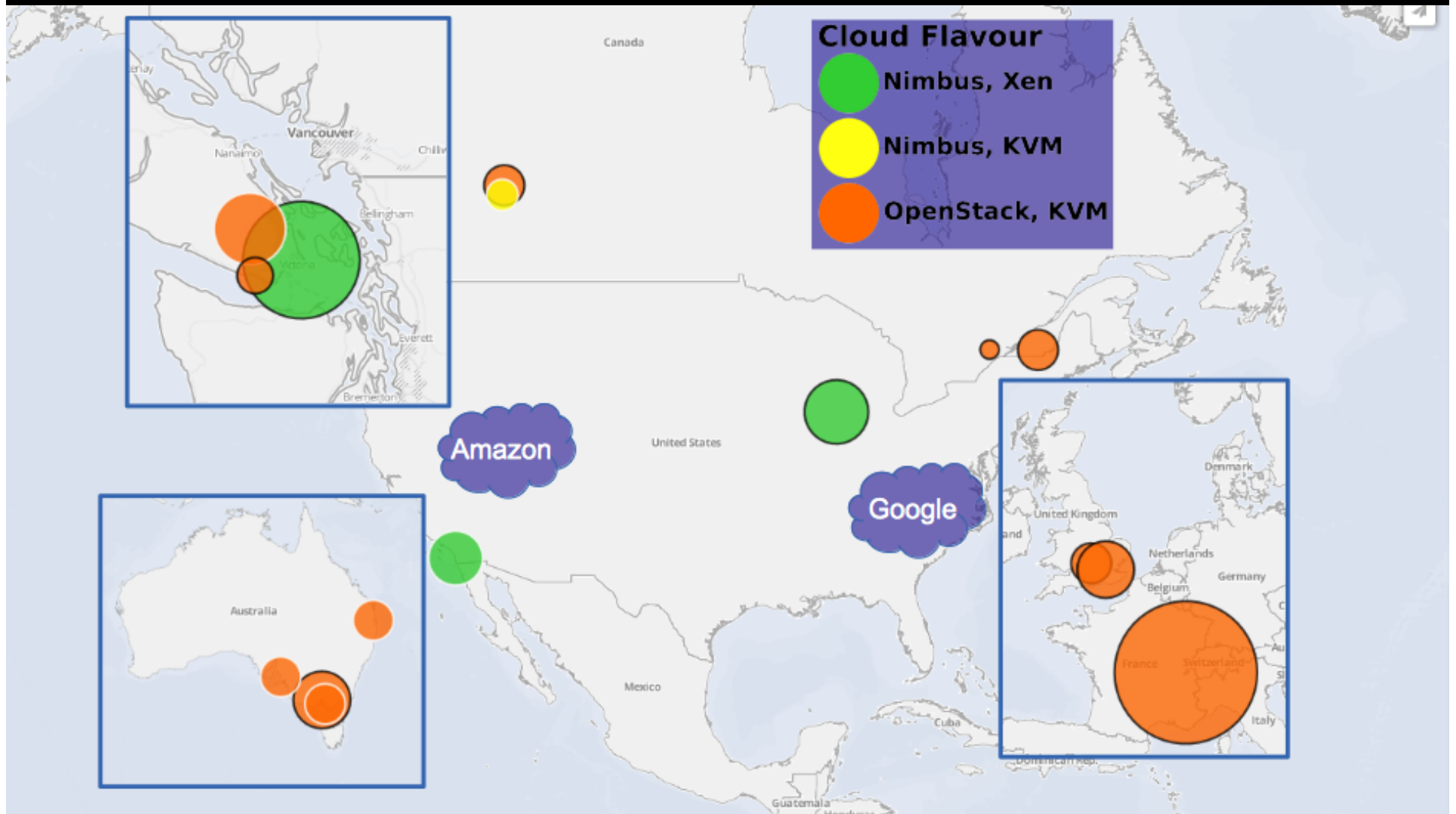
Use commercial clouds for
exceptional computing
demands

Distributed cloud computing
using HEP and non-HEP
clouds

Overview of distributed compute cloud



Clouds



Operate three HEP HTCondor/CS Instances

ATLAS

North America (2.5 years)

Europe (6 months)

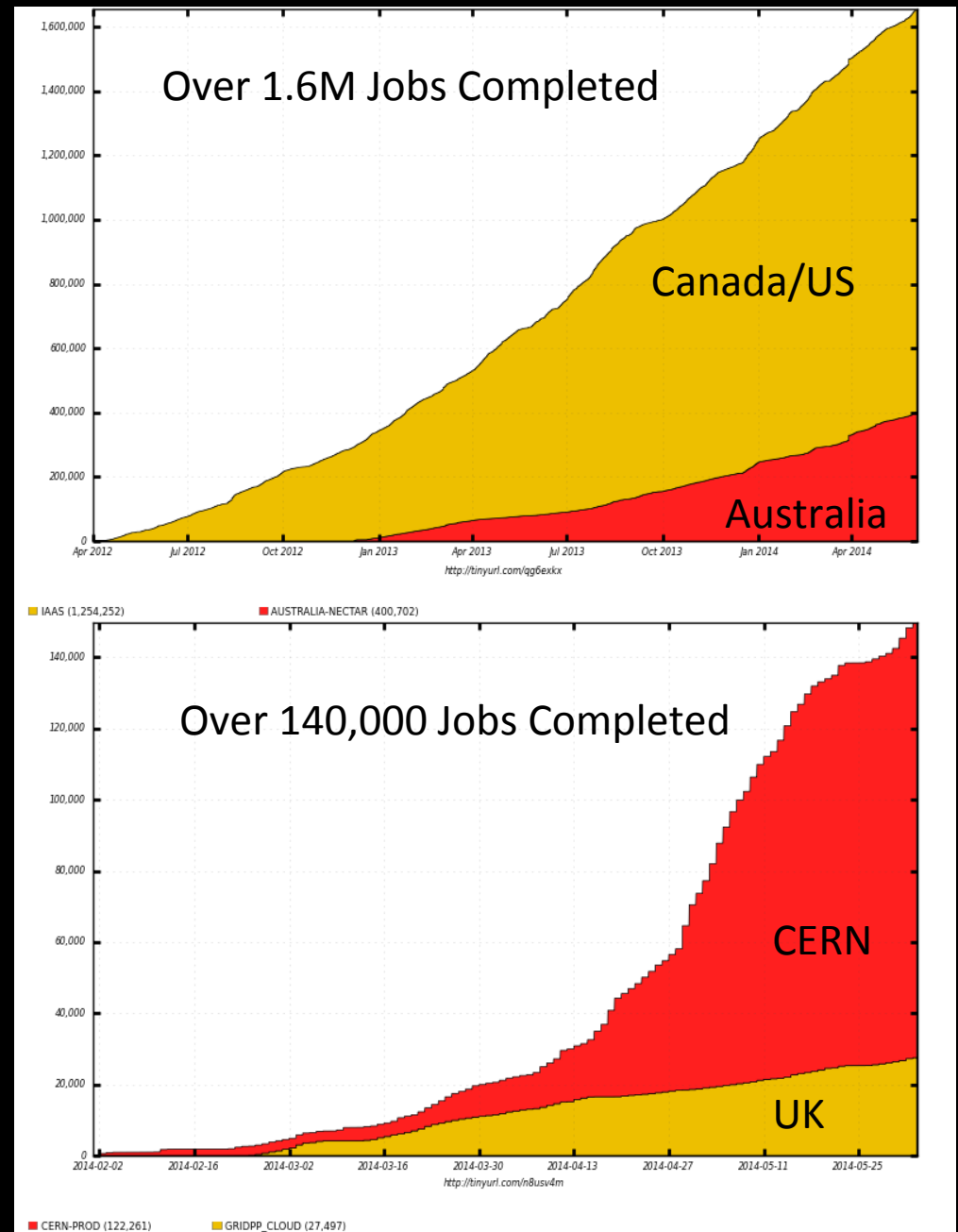
BelleII

North America (6 months)

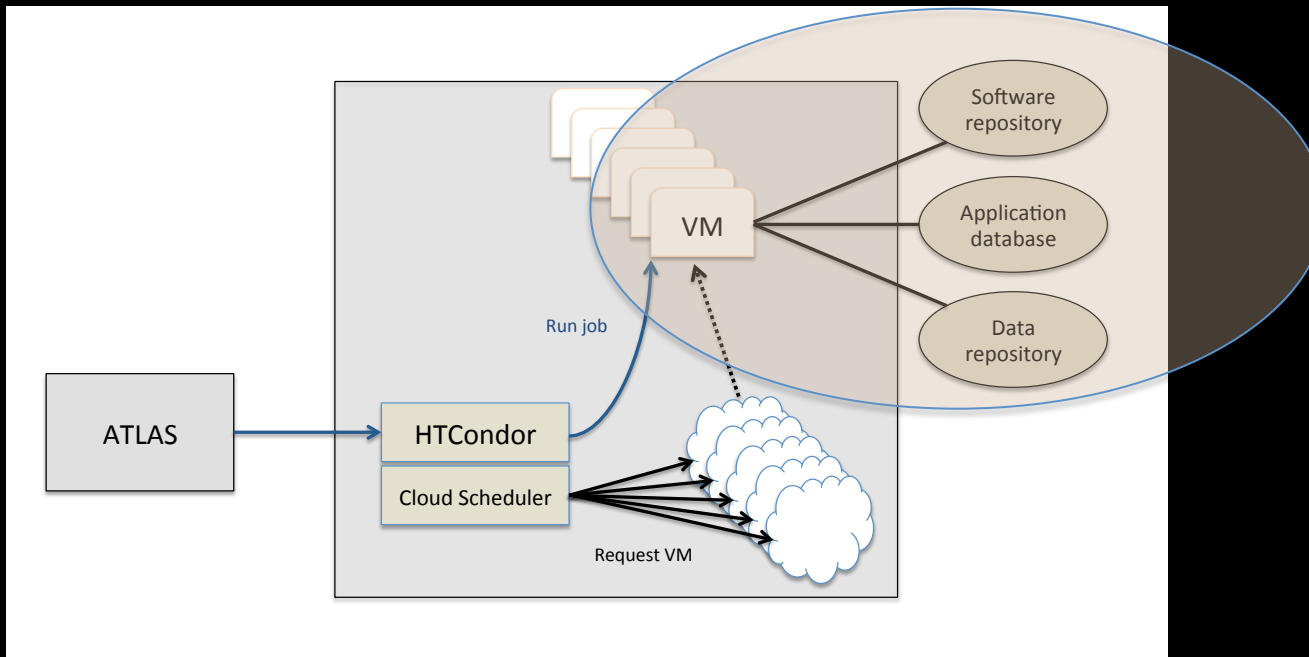
Total work loads average 1500 cores
Peak work loads 2500 cores

1000 cores on Amazon EC2

Working on API to Google GCE



Need for a *context-aware* cloud



VMs can appear anywhere
in the world

Software repository and
application database
servers are fixed nodes

No input data repository

Single output data
repository

We also have access to repositories distributed around the world
How can we make the system aware of these repositories?

Software distribution

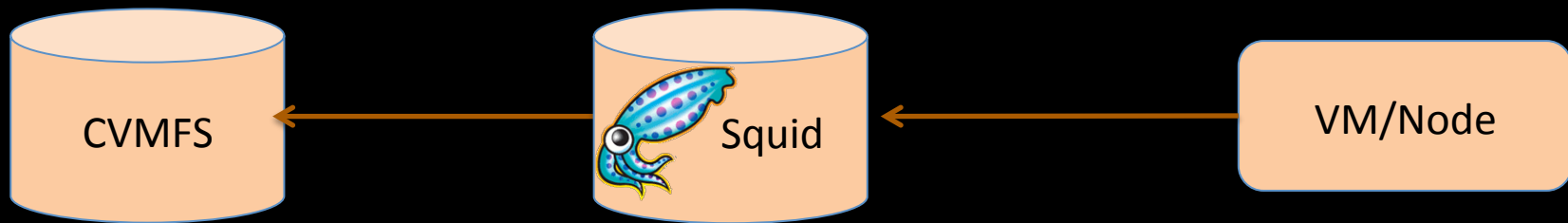


CernVM file system (CVMFS)

A network file system based on HTTP and optimized to deliver experiment software in a fast, scalable, and reliable way.

Files and file metadata are aggressively cached and downloaded on demand.

Software distribution



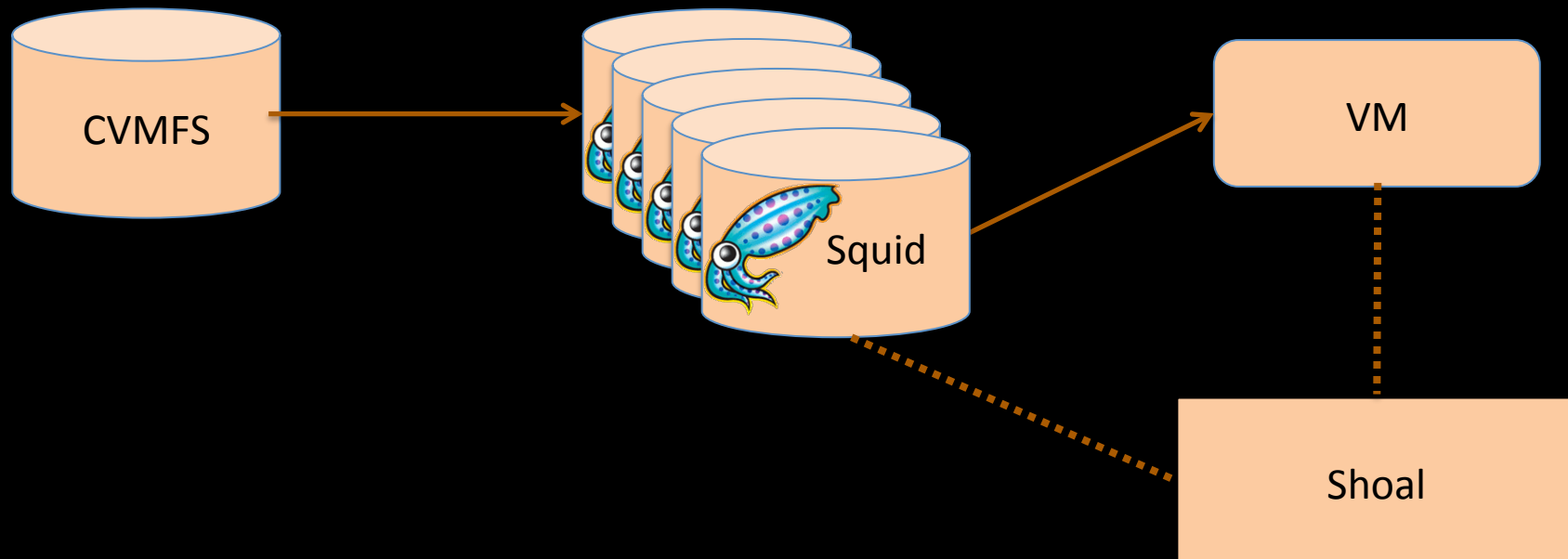
Squid HTTP Cache

Rather than query the CVMFS master repository we use Squid HTTP caches

Approximately 100 Squid caches distributed around the world

Squids distribute the world load but we still use a single Squid (per HTCondor/CS instance)

Software distribution in a context-aware cloud

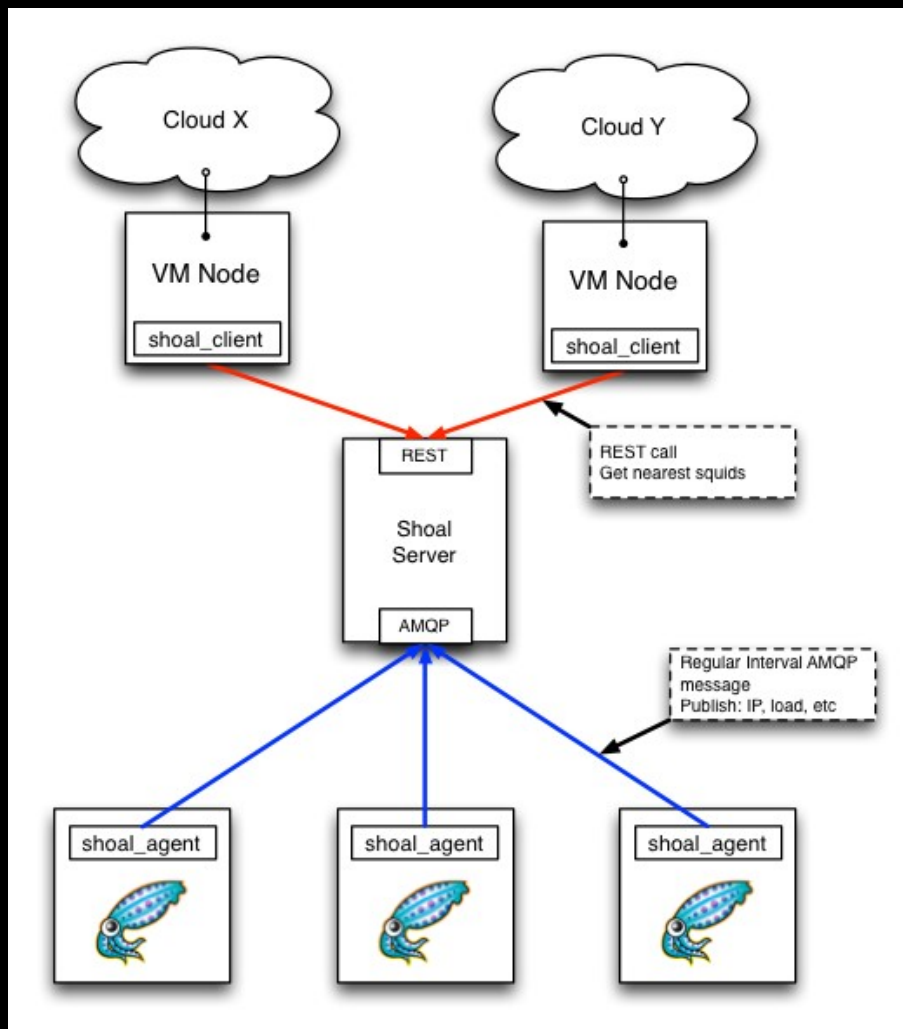


Shoal – a group of Squids

Shoal keeps a list of active Squid caches

Discovers new Squids and removes inactive Squids

VM queries Shoal for the location of the nearest squid (fastest network, lowest load)



Shoal

<http://arxiv.org/pdf/1311.0058v1.pdf>

github.com/hep-gc/shoal

Shoal used in production operation

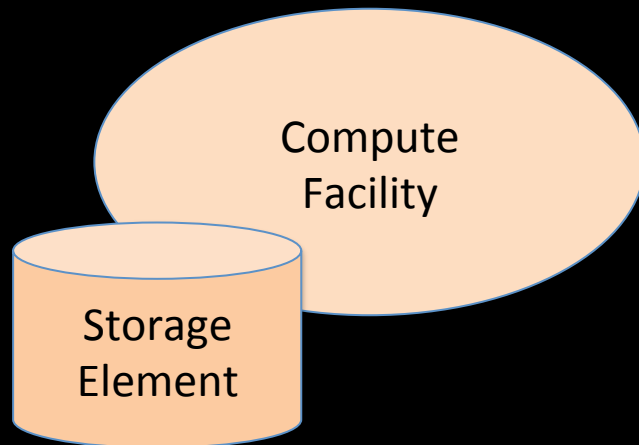
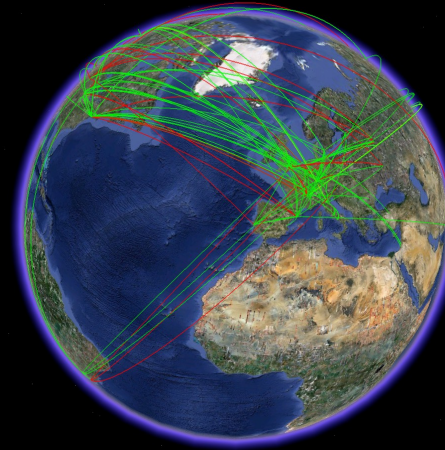
Allows for dynamic creation of Squids

A cloud could boot a Squid-VM

Data Management

Approximately 100 sites (150 PB of data)

Each site has CPU and storage (SE)
10-100G network



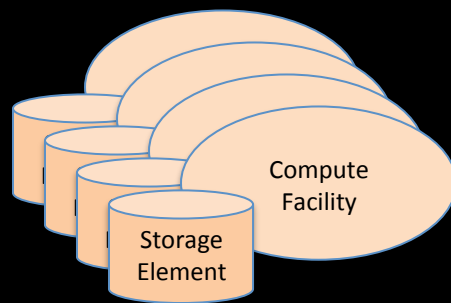
Two types of applications:

1. Analysis jobs are sent to the site with the data
2. Production jobs often have little input data
Primary use of cloud resources

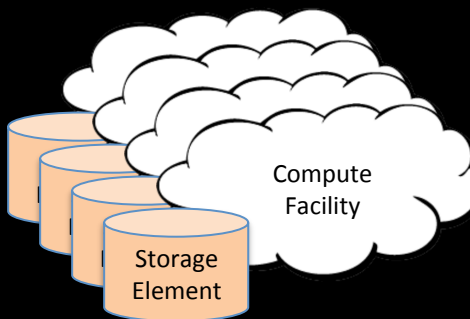
We have three types of facilities

Two of those types of facilities host the data samples

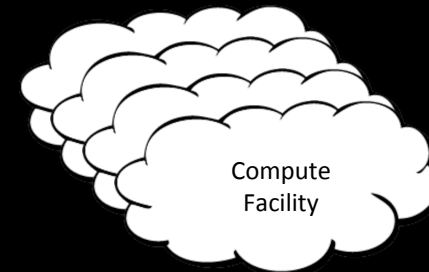
The samples at each SE are not identical



Non-cloud HEP



Cloud HEP



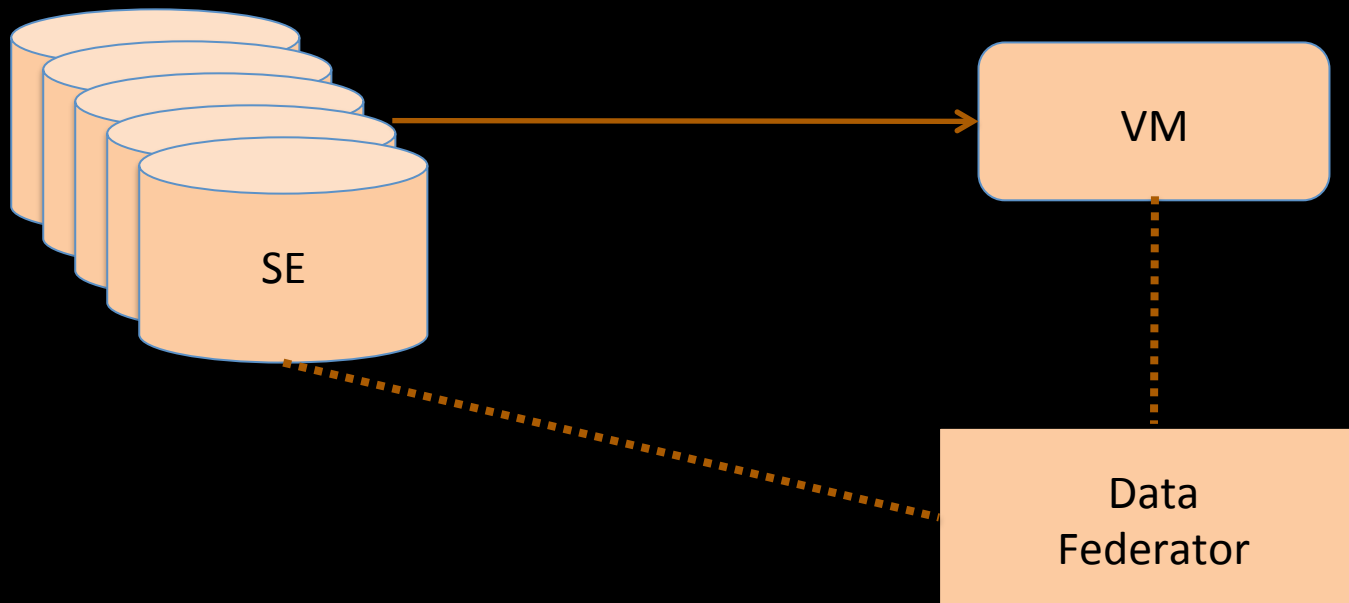
Non-HEP cloud

We want to run jobs with data on the cloud sites

The data is retrieved from the nearest SE

We need a “data federation”

HEP Data Federation



The LHC experiments operate an HEP-specific data federation – FAX

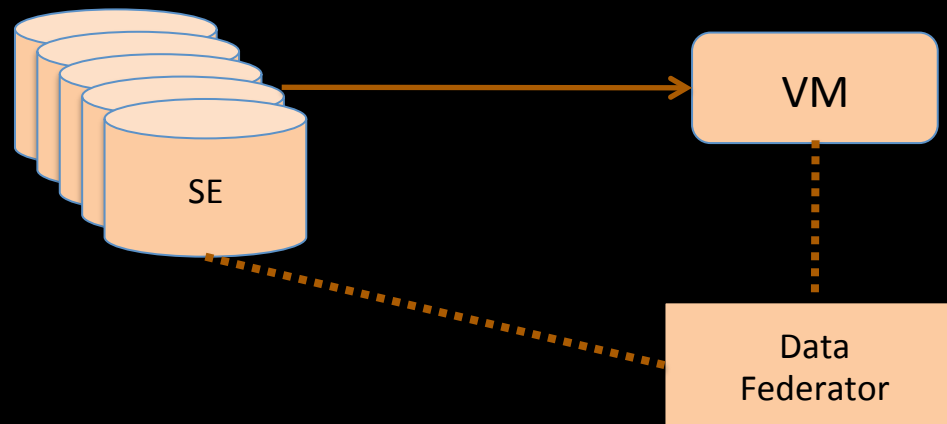
Primarily used for failover purposes on traditional computing facilities

We are testing a more general data federation

Targeted at clouds at both HEP and non-HEP sites

We have adopted WebDAV for file access

Open-source HTTP protocol standard



Custom (CERN) Data Federator

Presents a common view of the data file tree to the application

Directs the VM to the nearest SE with the input data

Uses GeoIP information to pick the closest site

Single file
tree of the
data

Browser address bar: /myfed/AOD/mc11_7TeV.105011.J2_pythia_jetjet.merge.AOD.e815_s1273_s1274_r2731_r2780_tid541566_00/ - Mozilla Firefox

Browser tabs: /myfed/AOD/mc11_7T...

Browser menu: Disable, Cookies, CSS, Forms, Images, Information, Miscellaneous, Outline, Resize, Tools

/myfed

/AOD/mc11_7TeV.105011.J2_pythia_jetjet.merge.AOD.e815_s1273_s1274_r2731_r2780_tid541566_00

Mode	UID	GID	Size	Modified	Name
-rwxrwxrwx	0	0	1.1G	Thu, 01 Jan 1970 00:00:00 GMT	AOD.541566.000001.pool.root.1
-rwxrwxrwx	0	0	1.2G	Thu, 01 Jan 1970 00:00:00 GMT	AOD.541566.000002.pool.root.1
-rwxrwxrwx	0	0	1.0G	Thu, 01 Jan 1970 00:00:00 GMT	AOD.541566.000003.pool.root.1
-rwxrwxrwx	0	0	1.0G	Thu, 01 Jan 1970 00:00:00 GMT	AOD.541566.000004.pool.root.1
-rwxrwxrwx	0	0	1.2G	Thu, 01 Jan 1970 00:00:00 GMT	AOD.541566.000005.pool.root.1
-rwxrwxrwx	0	0	1014.9M	Thu, 01 Jan 1970 00:00:00 GMT	AOD.541566.000006.pool.root.1
-rwxrwxrwx	0	0	1011.6M	Thu, 01 Jan 1970 00:00:00 GMT	AOD.541566.000007.pool.root.1
-rwxrwxrwx	0	0	1001.6M	Thu, 01 Jan 1970 00:00:00 GMT	AOD.541566.000008.pool.root.1
-rwxrwxrwx	0	0	1000.7M	Thu, 01 Jan 1970 00:00:00 GMT	AOD.541566.000009.pool.root.1
-rwxrwxrwx	0	0	1.0G	Thu, 01 Jan 1970 00:00:00 GMT	AOD.541566.000010.pool.root.1
-rwxrwxrwx	0	0	1.0G	Thu, 01 Jan 1970 00:00:00 GMT	AOD.541566.000011.pool.root.1
-rwxrwxrwx	0	0	1.0G	Thu, 01 Jan 1970 00:00:00 GMT	AOD.541566.000012.pool.root.1
-rwxrwxrwx	0	0	1.0G	Thu, 01 Jan 1970 00:00:00 GMT	AOD.541566.000013.pool.root.1
-rwxrwxrwx	0	0	1.0G	Thu, 01 Jan 1970 00:00:00 GMT	AOD.541566.000014.pool.root.1
-rwxrwxrwx	0	0	1.0G	Thu, 01 Jan 1970 00:00:00 GMT	AOD.541566.000015.pool.root.1
-rwxrwxrwx	0	0	1.0G	Thu, 01 Jan 1970 00:00:00 GMT	AOD.541566.000016.pool.root.1

Single file
tree of the
data

```

/myfed
/AOD/mc11_7TeV.105011.J2_pythia_jetjet.merge.AOD.e815_s1273_s12

```

The screenshot shows a terminal window with a directory listing and a gedit editor window. The directory listing has columns for Mode, UID, GID, Size, Modified, and Name. The gedit window is editing a file named _AOD_mc-5.metalink and contains XML content.

Mode	UID	GID	Size	Modified	Name
-rwxrwxrwx	0	0	1.1G	Thu, 01 Jan 1970 00:00:00 GMT	AOD.541566..000001.pool.root.1
-rwxrwxrwx	0	0	1.2G	Thu, 01 Jan 1970 00:00:00 GMT	AOD.541566..000002.pool.root.1
-rwxrwxrwx	0	0	1.0G	Thu, 01 Jan 1970 00:00:00 GMT	AOD.541566..000003.pool.root.1

The gedit window title is `_AOD_mc-5.metalink (/tmp) - gedit`. The file content is:

```
<?xml version="1.0" encoding="UTF-8"?>
<metalink version="3.0" xmlns="http://www.metalinker.
generator="lcgdm-dav" pubdate="Thu, 01 Jan 1970 00:00
<files>
<file name="/AOD/mc">
<size>1077899833</size>
```

Two copies
of this file



```
<?xml version="1.0" encoding="UTF-8"?>
<metalink version="3.0" xmlns="http://www.metalinker.org/" xmlns:lcgdm="LCGDM:"
generator="lcgdm-dav" pubdate="Thu, 01 Jan 1970 00:00:00 GMT">
<files>
<file name="/AOD/mc">
  <size>1077899833</size>
  <resources>
    <url type="https">http://elephant08.heprc.uvic.ca/atlas/AOD/
mc11_7TeV.105011.J2_pythia_jetjet.merge.AOD.e815_s1273_s1274_r2731_r2780_tid541566_00
AOD.541566._000010.pool.root.1</url>
    <url type="https">http://elephant10.heprc.uvic.ca/atlas/AOD/
mc11_7TeV.105011.J2_pythia_jetjet.merge.AOD.e815_s1273_s1274_r2731_r2780_tid541566_00
AOD.541566._000010.pool.root.1</url>
  </resources>
</file>
</files>
</metalink>
```

Data Federation Status

Currently testing the data federation with our two test clouds

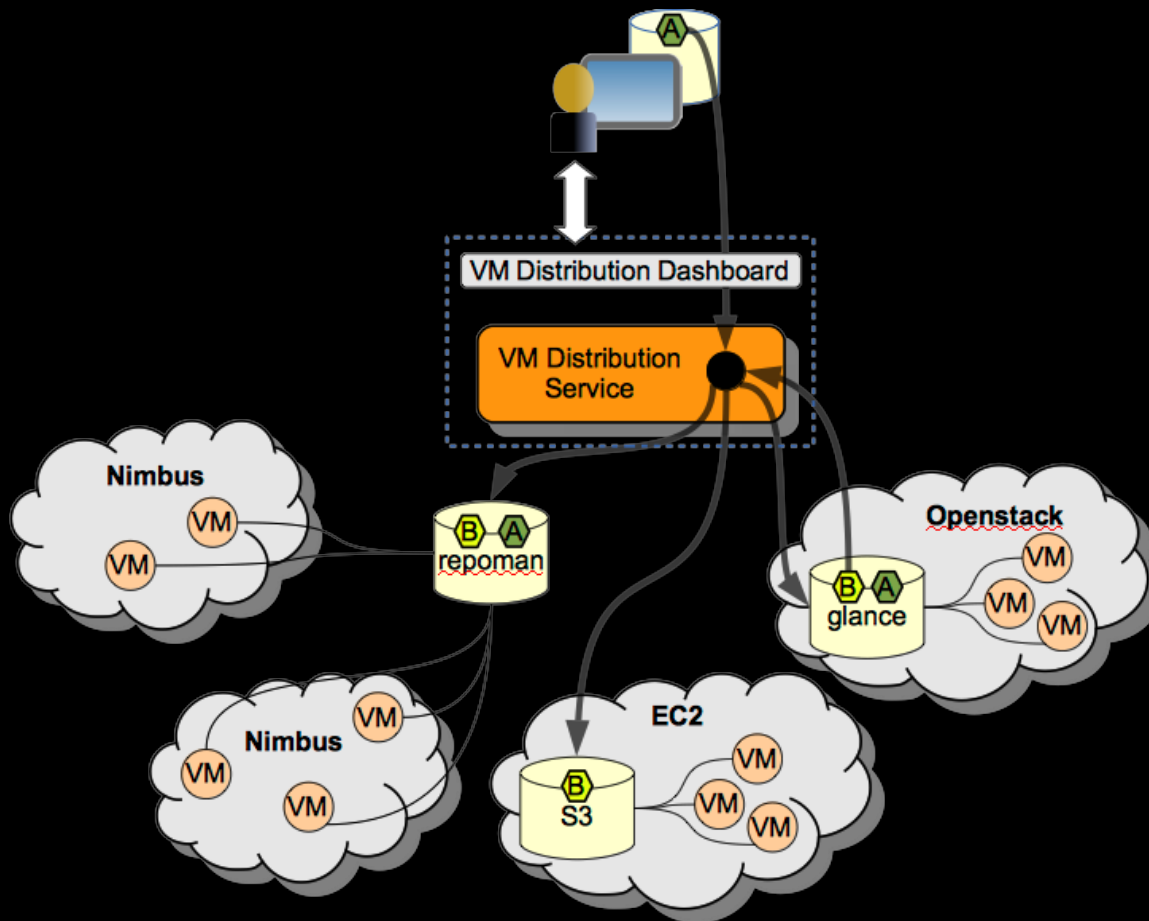
Sorted out some issues

Denial-of-Service vulnerability fixed

Disabled web crawling using “robots.txt”

Next step is to use it with the ATLAS application

Image distribution system



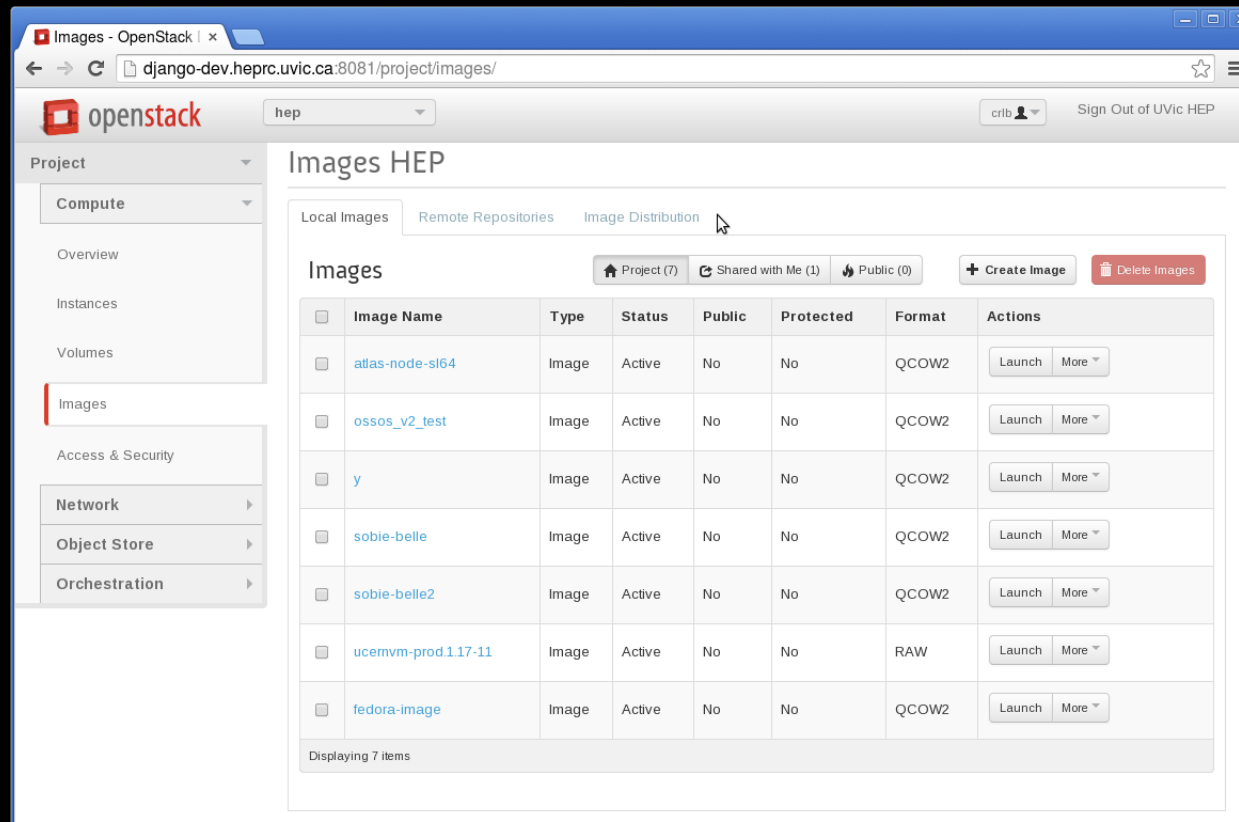
We need to manage VM images over multiple clouds

Nimbus clouds can access images from external repositories (such as our Repoman repository)

Openstack and **EC2** clouds require that we manually upload images to each cloud

Glnt

Created *Glnt* to integrate with the individual *Glance* repositories of each OpenStack cloud



Glnt is designed to be part of the OpenStack GUI

Glint manages images on multiple OpenStack clouds

The screenshot shows the OpenStack Images HEP interface. The browser address bar displays `django-dev.heprc.uvic.ca:8081/project/images/`. The interface includes a sidebar with navigation options: Project, Compute, Overview, Instances, Volumes, Images (selected), Access & Security, Network, Object Store, and Orchestration. The main content area is titled "Images HEP" and features tabs for Local Images, Remote Repositories, and Image Distribution. The Remote Repositories tab is active, showing a table of sites. Above the table are buttons for "+ Create Site" and "Delete Sites". A tooltip labeled "Create Site" is visible over the "+ Create Site" button. The table lists three sites: ALTO, DAIR Cloud, and Mouse, each with a checkbox, site name, cloud URL, auth port, cloud type, and actions (AddCredential and More).

☐	Site Name	Cloud URL	Auth Port	Cloud Type	Actions
☐	ALTO	<code>http://alto.cloud.nrc.ca</code>	5000	Openstack	AddCredential More
☐	DAIR Cloud	<code>http://nova-ab.dair-atir.canarie.ca</code>	5000	Openstack	AddCredential More
☐	Mouse	<code>http://mouse01.heprc.uvic.ca</code>	5000	Openstack	AddCredential More

Displaying 3 items

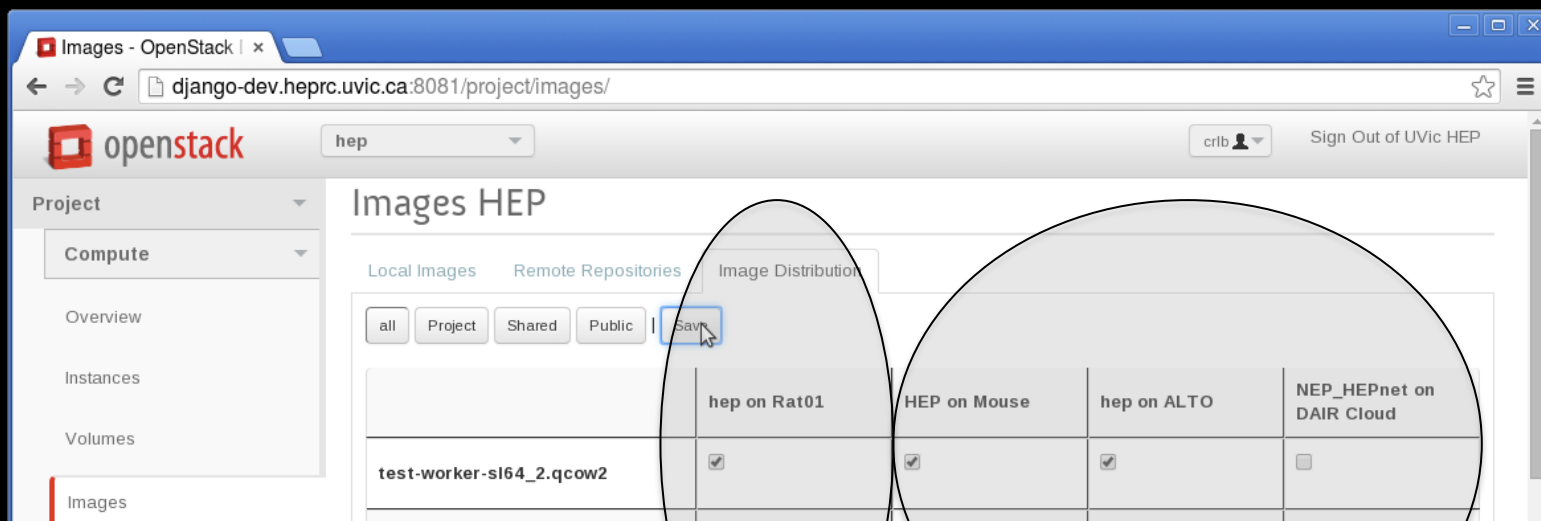
This image shows the remote repositories

The screenshot shows the OpenStack Images dashboard in a web browser. The browser's address bar displays the URL `django-dev.heprc.uvic.ca:8081/project/images/`. The dashboard has a sidebar with navigation links: Project, Compute, Overview, Instances, Volumes, Images (selected), Access & Security, Network, Object Store, and Orchestration. The main content area is titled 'Images HEP' and contains a 'Create Site' dialog box. The dialog box has the following fields: 'Name: *' (text input), 'Description: Specify a site for the Glint Service.' (text area), 'Url: *' (text input), 'Port: *' (text input), 'Format: *' (dropdown menu with 'Openstack' selected), and 'Description:' (text area). At the bottom of the dialog are 'Cancel' and 'Create Site' buttons.

And add other clouds

Add user credentials

The screenshot shows the OpenStack Images dashboard in a web browser. The browser's address bar displays the URL `django-dev.heprc.uvic.ca:8081/project/images/`. The dashboard has a sidebar with navigation links: Project, Compute, Overview, Instances, Volumes, Images (selected), Access & Security, Network, Object Store, and Orchestration. The main content area is titled 'Images HEP' and contains an 'Add Credential' dialog box. The dialog box has the following fields: 'Tenant: *' (text input), 'Description: Add a Credential to a Site using the Glint Service.' (text area), 'User Name: *' (text input), and 'Password: *' (text input). At the bottom of the dialog are 'Cancel' and 'Add Credential' buttons.



Home Repository
Not a master repository

Remote

The images on the Home Repository can be different than those
on the Remote Repositories
Potential for different images to use the same name

The screenshot shows the OpenStack Images HEP interface. The left sidebar contains navigation links: Project, Compute, Overview, Instances, Volumes, Images (highlighted), Access & Security, Network, Object Store, and Orchestration. The main content area is titled 'Images HEP' and has tabs for Local Images, Remote Repositories, and Image Distribution. Below these tabs are filters: all, Project, Shared, Public, and a 'Save' button (circled). The table below lists various images and their distribution status across different platforms.

	hep on Rat01	HEP on Mouse	hep on ALTO	NEP_HEPnet on DAIR Cloud
test-worker-sl64_2.qcow2	☒	☒	☒	☐
b_sl6_april24	☒	☒	☒	☐
atlas-node-sl64	☒	☒	☒	☒
ossos_v2_test	☒	☒	☐	☐
y	☒	☐	☐	☐
sobie-belle	☒	☒	☒	☐
sobie-belle2	☒	☒	☐	☐
ucernvm-prod 1.17.11	☒	☒	☐	☒
to_share_with_ron	☒	☒	☒	☒
fedora-image	☒	☒	☐	☐
Fedora-x86_64-19-20130627-sda.qcow2	☒	☒	☒	☒

Glint is under beta-test and will in production before September 2014
Integration into OpenStack is a key goal

Summary

We have been operating a distributed cloud computing system for scientific applications for many years

Used for ATLAS and BelleII HEP experiments and Canadian astronomy community

We are extending the functionality and reliability of the system by making it a context-aware distributed cloud

Utilizing distributed software caches

Accessing distributed data repositories

Managing VM images over multiple clouds

We expect the system to grow in scale to 5000 cores by 2015

Providing these new services will further enhance its capabilities

Acknowledgements



Contact information:

Web sites:

rsobie@uvic.ca

<http://rjs.phys.uvic.ca/>

<http://heprc.phys.uvic.ca/home/>

