Distributed Cloud Computing in High Energy Physics

Randall Sobie

Institute of Particle Physics

University of Victoria

High Energy Physics

(Particle Physics)

The area of physics that studies the fundamental particles of nature and their interactions.

Accelerators

SLAC Linear Accelerator

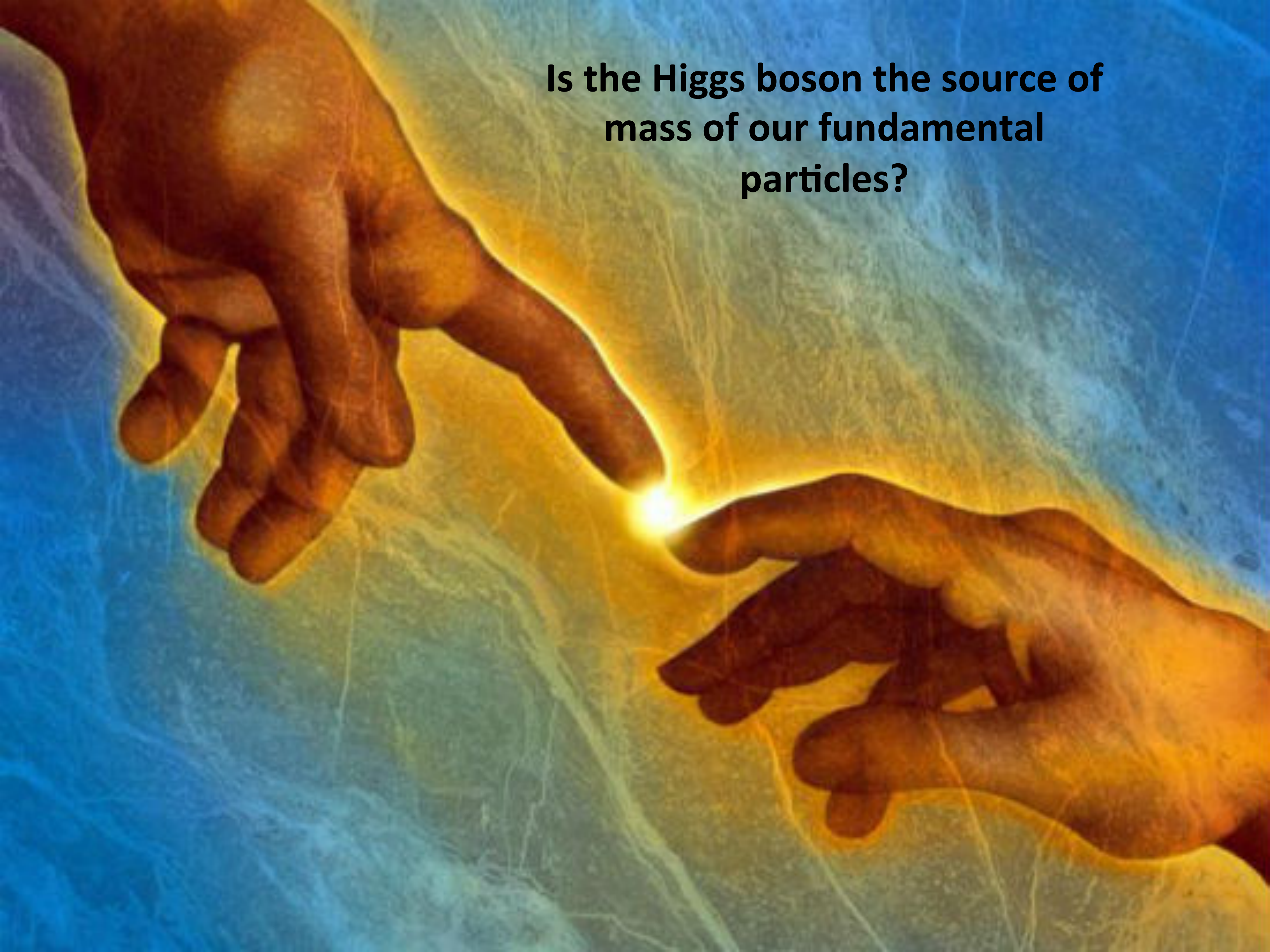
Underground labs

Sudbury Solar Neutrino Detector

Orbiting labs

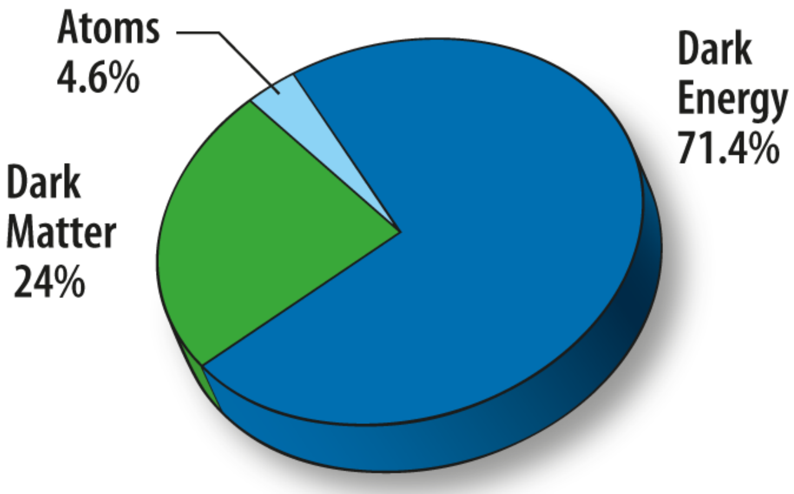
Alpha Magnetic Spectrometer
Space Station

Is the Higgs boson the source of mass of our fundamental particles?

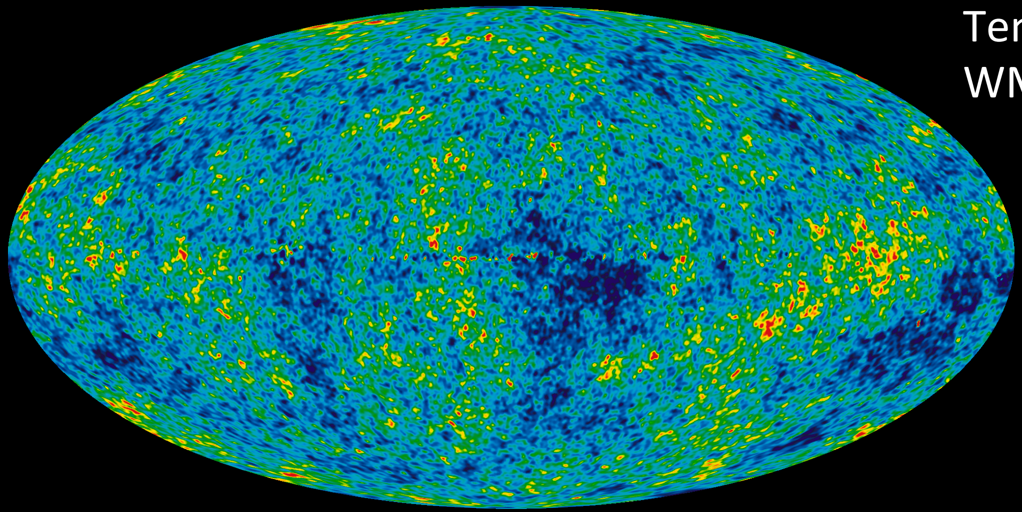


Why is the universe made of matter and not equal amounts of matter/antimatter?

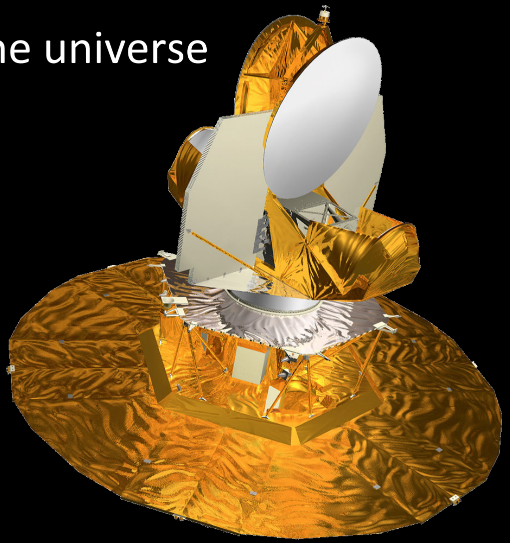
What is origin of Dark Matter and Dark Energy?



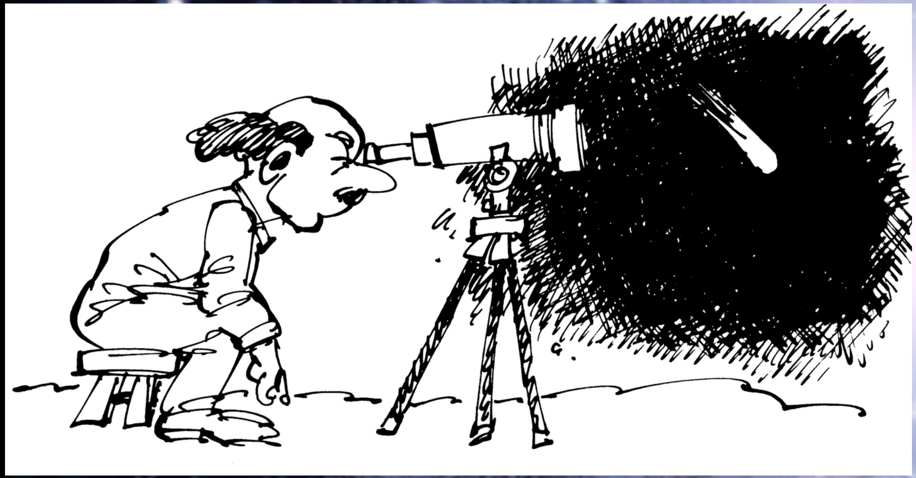
We do not know the composition of 95% of the universe

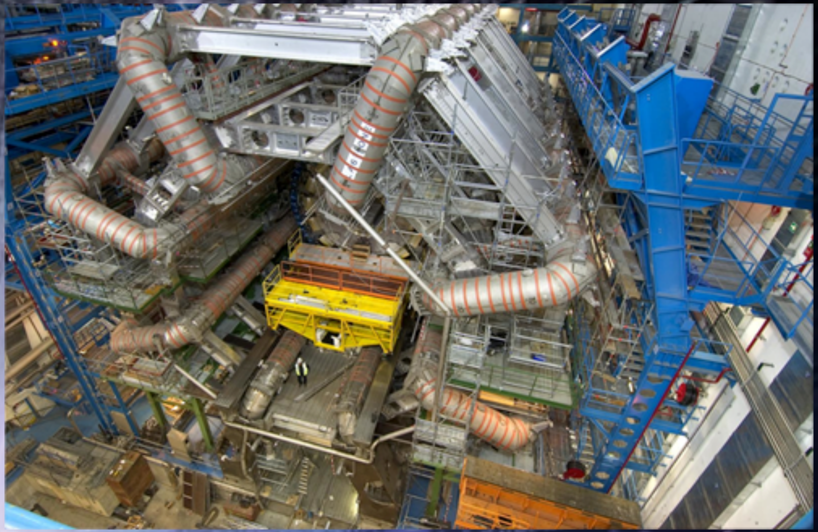


Temperature of the universe
WMAP satellite



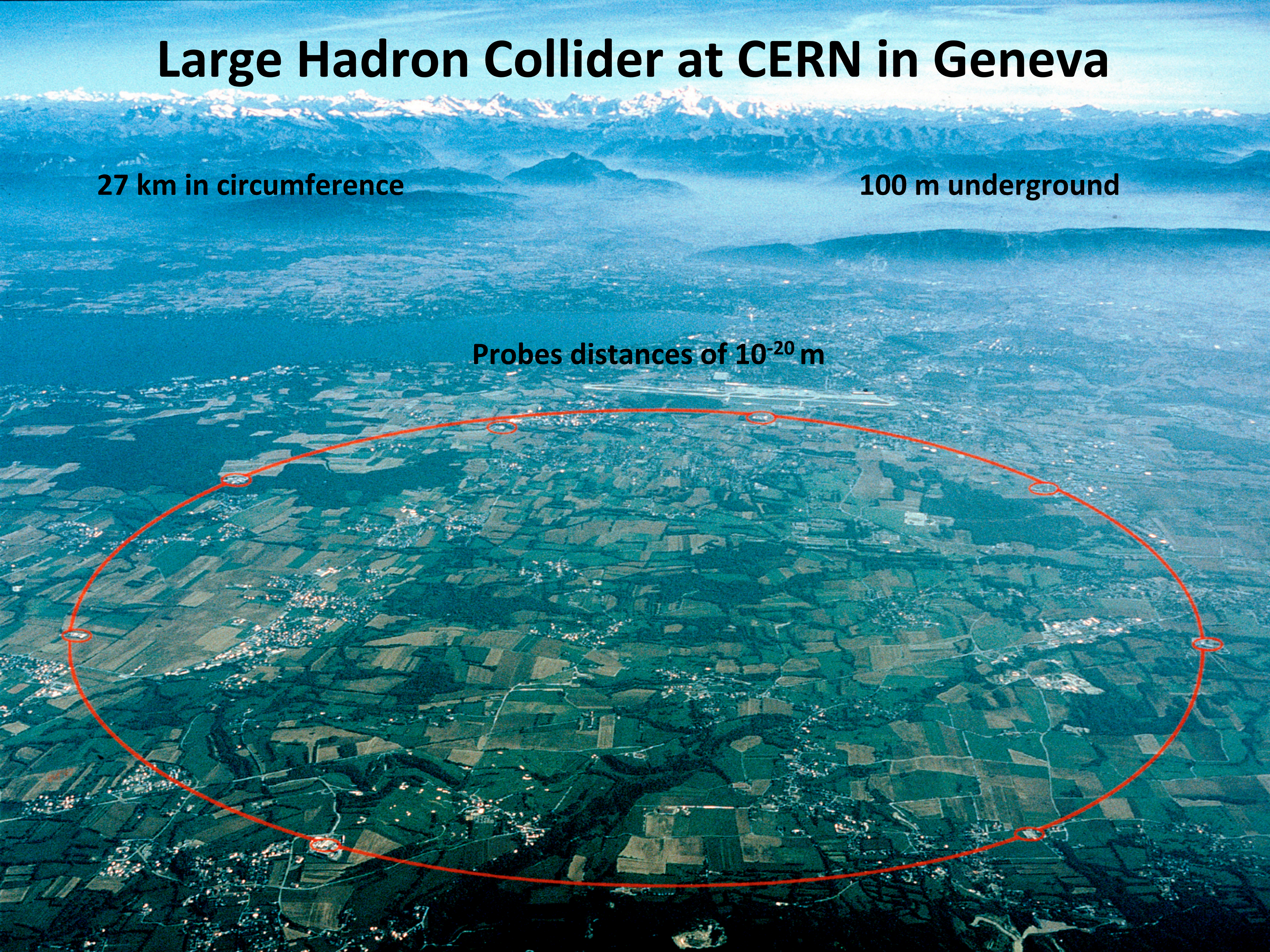
Understanding our World





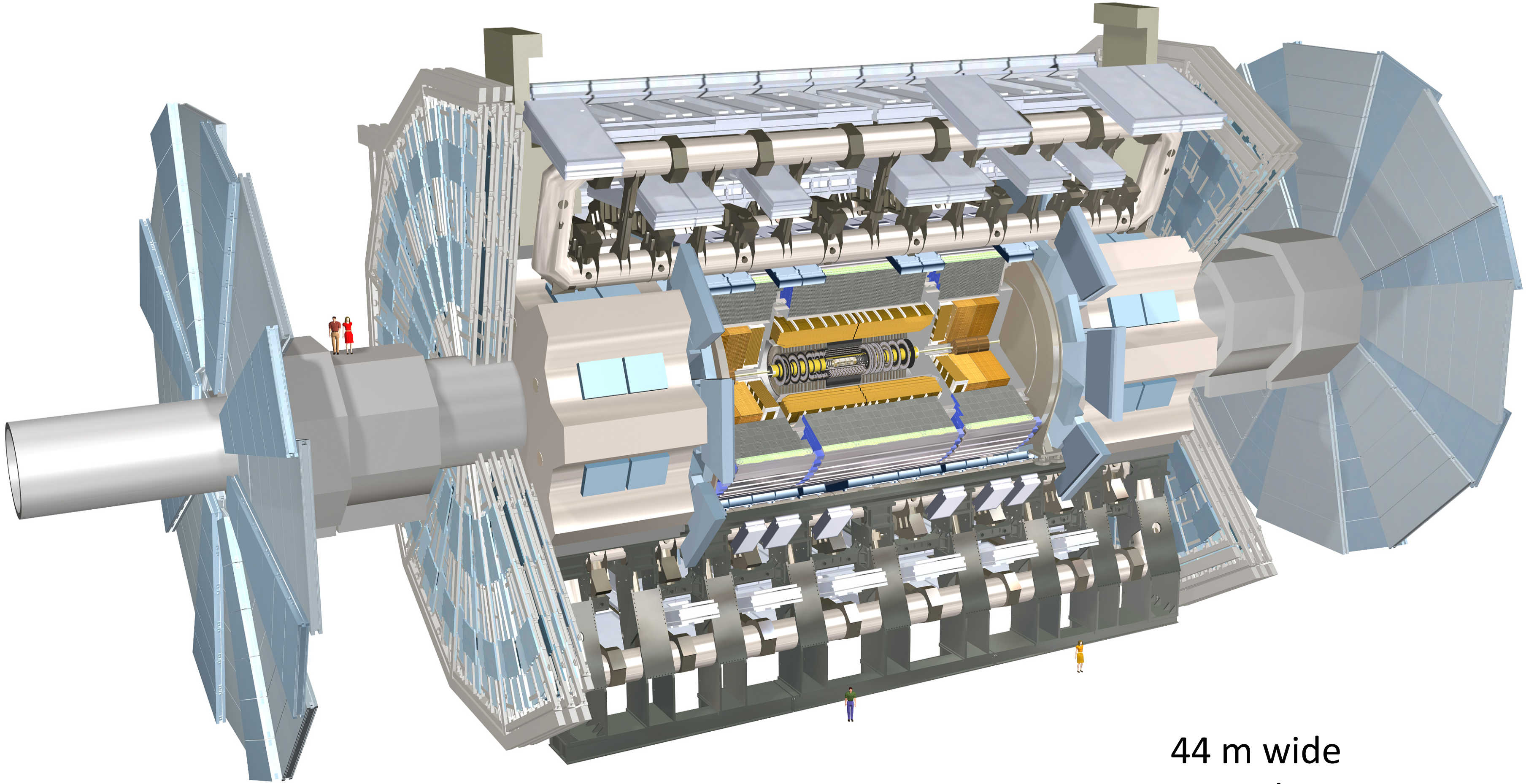

C
D

Large Hadron Collider at CERN in Geneva

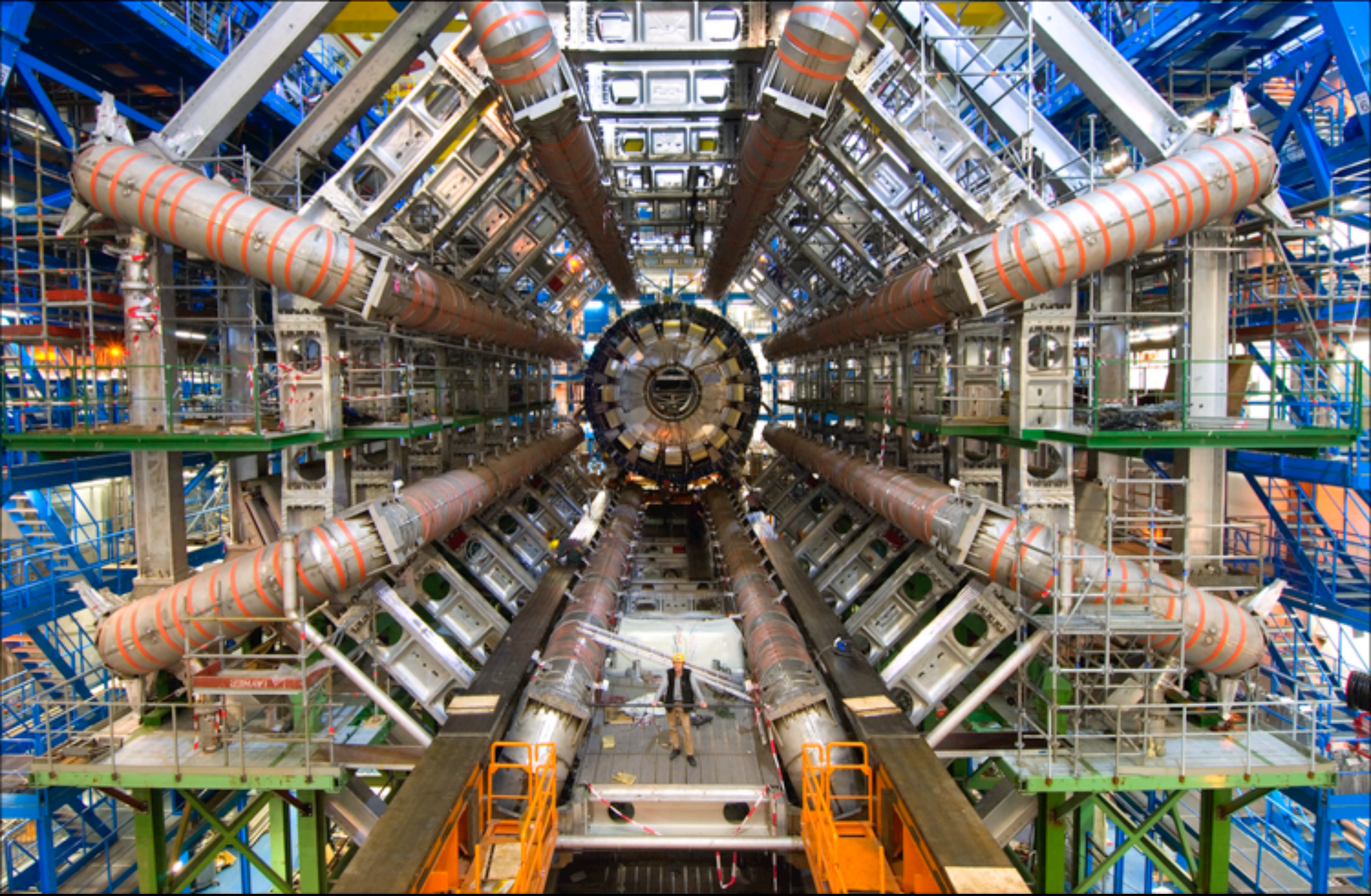


Blue tubes contain the two beam pipes and magnets at 1.8 degrees Kelvin

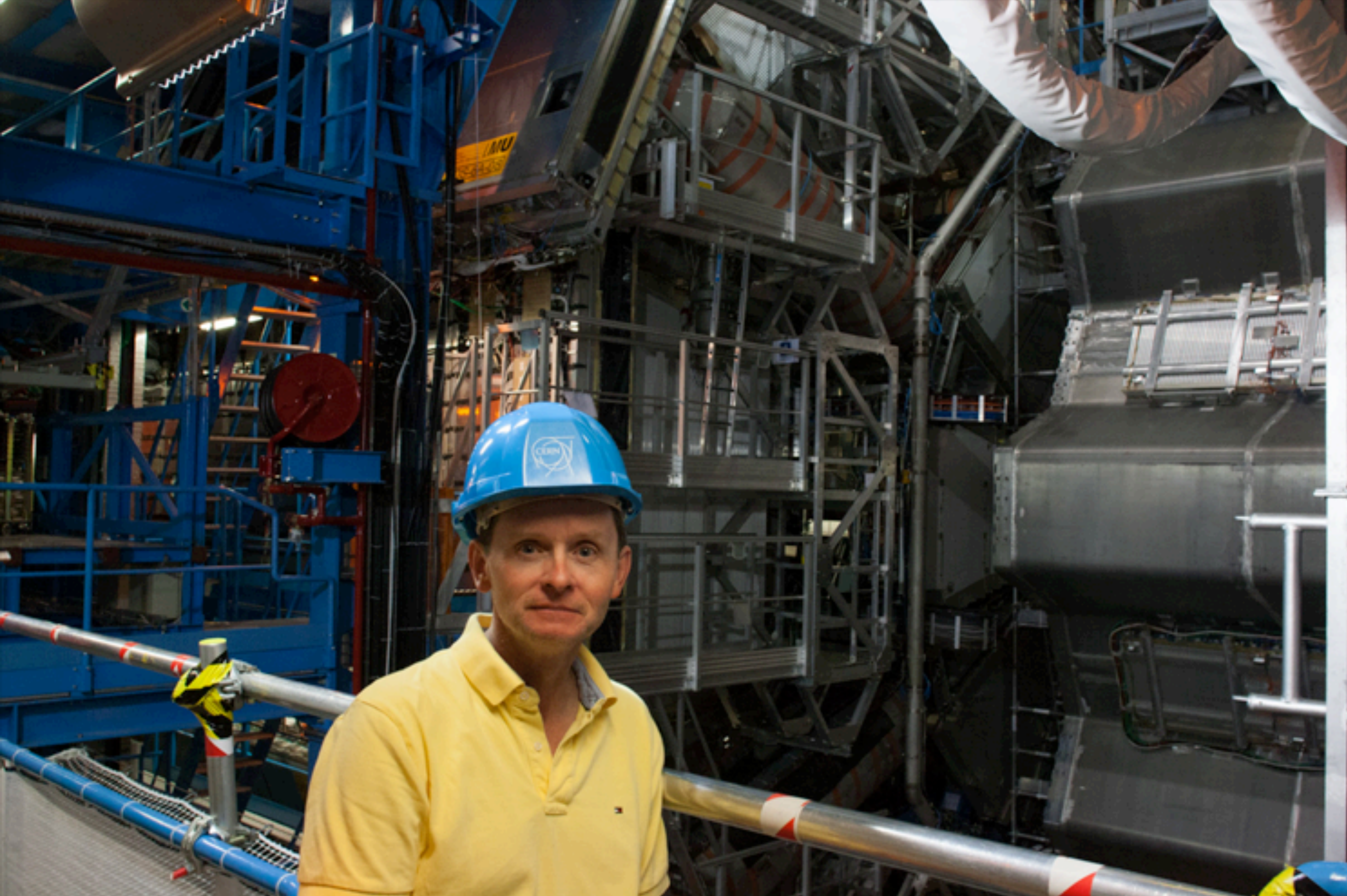
ATLAS Detector

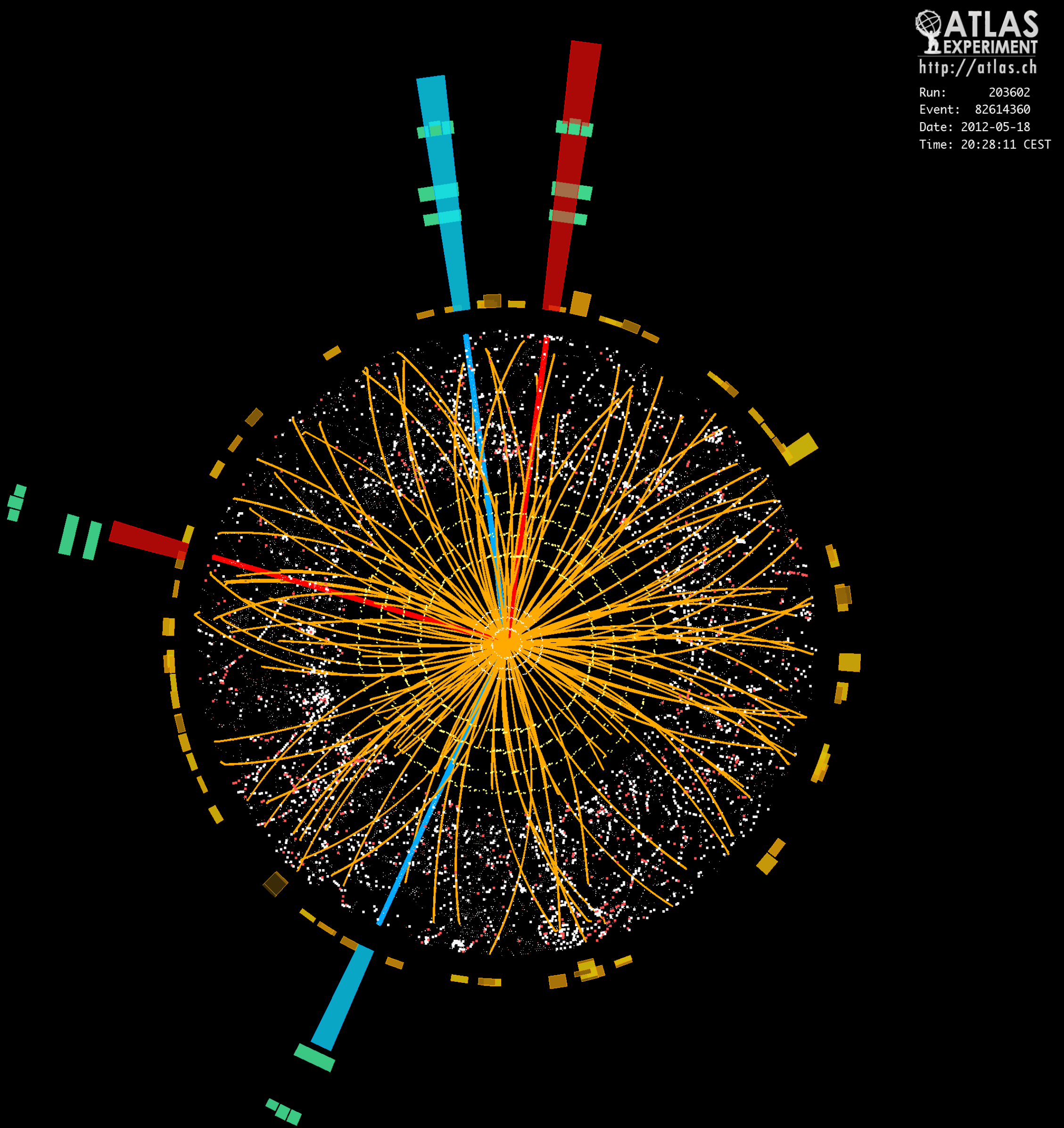


44 m wide
25 m diameter



ATLAS detector during construction in 2005




ATLAS EXPERIMENT
http://atlas.ch
Run: 203602
Event: 82614360
Date: 2012-05-18
Time: 20:28:11 CEST

The New York Times

Science

July 4, 2012

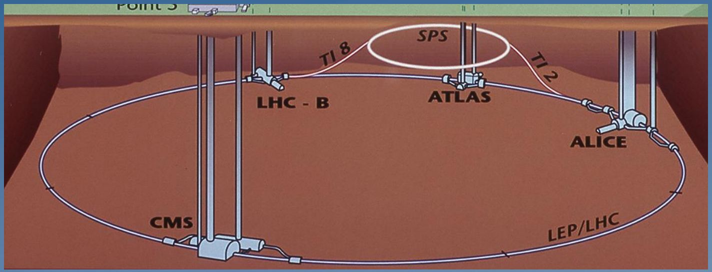
WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

ENVIRONMENT SPACE & COSMOS

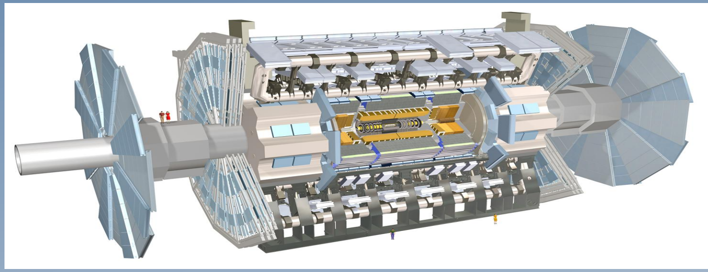
Physicists Find Elusive Particle Seen as Key to Universe

Pool photo by Denis Balibouse

Scientists in Geneva on Wednesday applauded the discovery of a subatomic particle that looks like the Higgs boson.

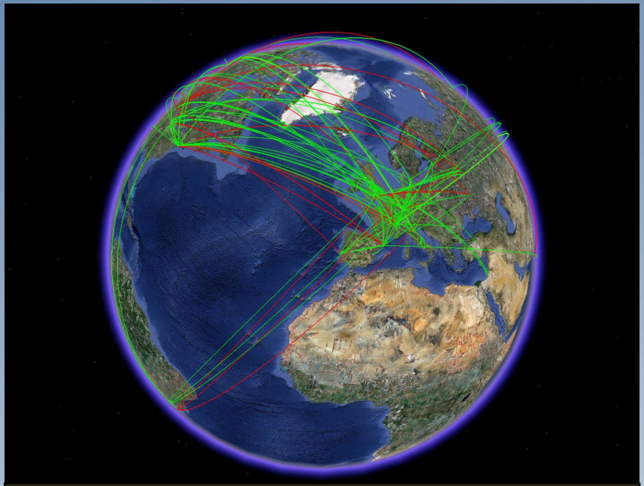


40 million collisions per second



100,000 collisions selected

200 events per second



WLCG Computing Grid
CERN Tier0
10 Tier1 sites
60+ Tier2 sites

150 PB data

Clouds in High Energy Physics

Long-term preservation of software and data of HEP experiments

Distributed cloud computing using HEP and non-HEP clouds

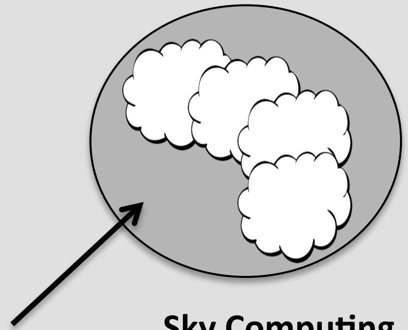
Utilize special computing resources attached to the detectors

Use commercial clouds for exceptional computing demands

Simplify the management of heterogeneous in-house resources

Distributed cloud computing

Grid of Clouds



Sky Computing
K. Keahey etal

Use multiple, heterogeneous IaaS clouds for batch workloads

Design Goals

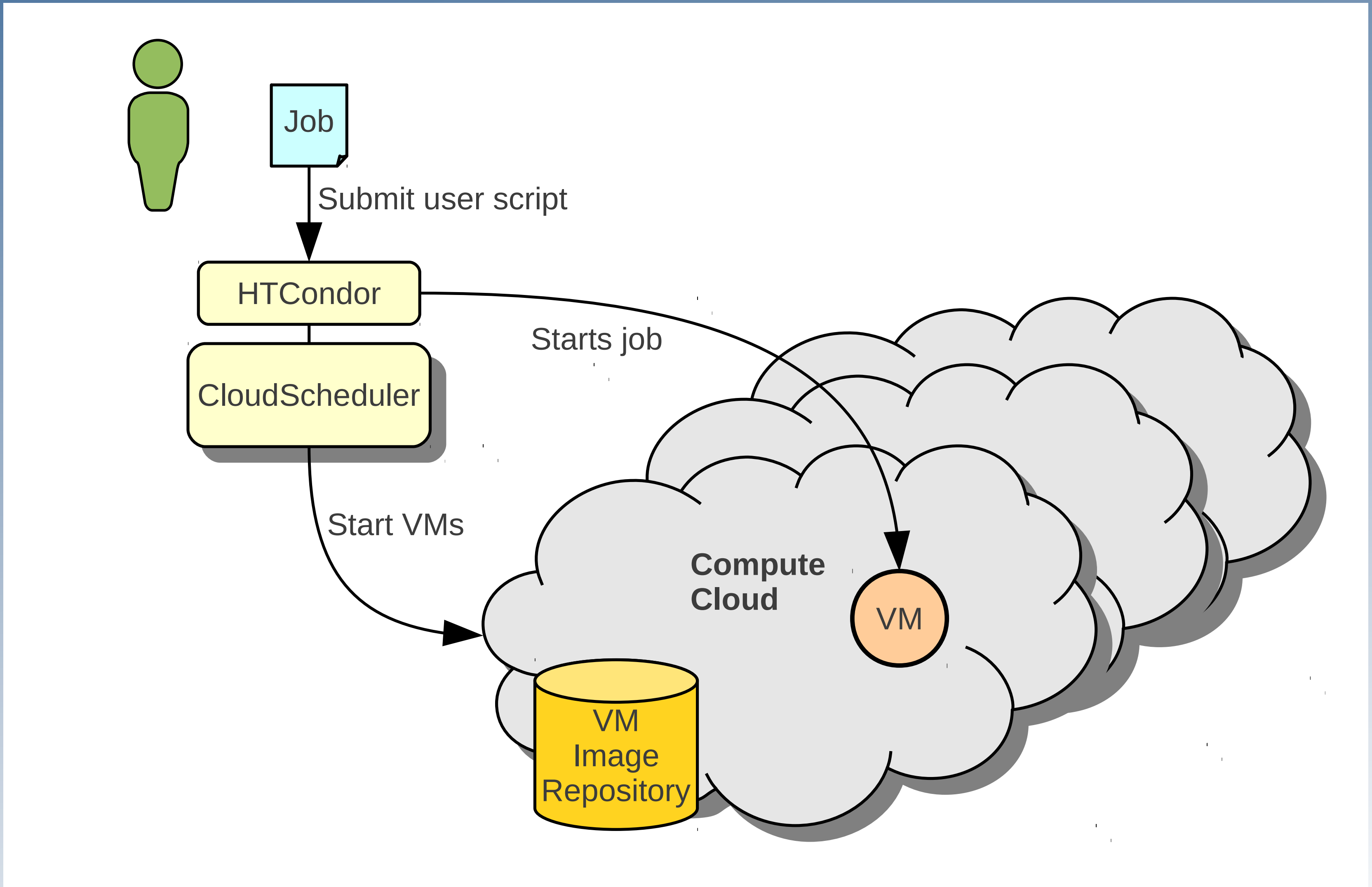
Independent of the IaaS cloud type or hypervisor

No application software on any cloud
(No HEP specific software)

Use dedicated and opportunistic resources

Leverage existing cloud or batch computing software

Distributed Batch Cloud Computing



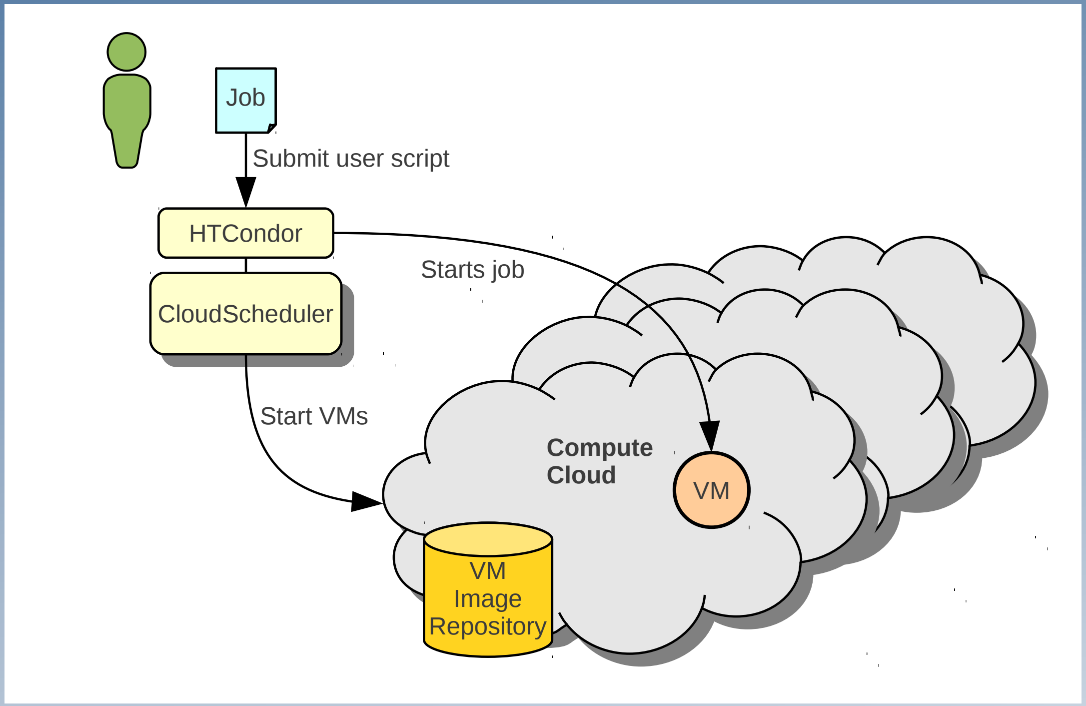
HTCondor

Designed as a cycle scavenger

Ideal as a job scheduler for a dynamic environment

Job description file

VM image
RAM
Number cores
Storage space



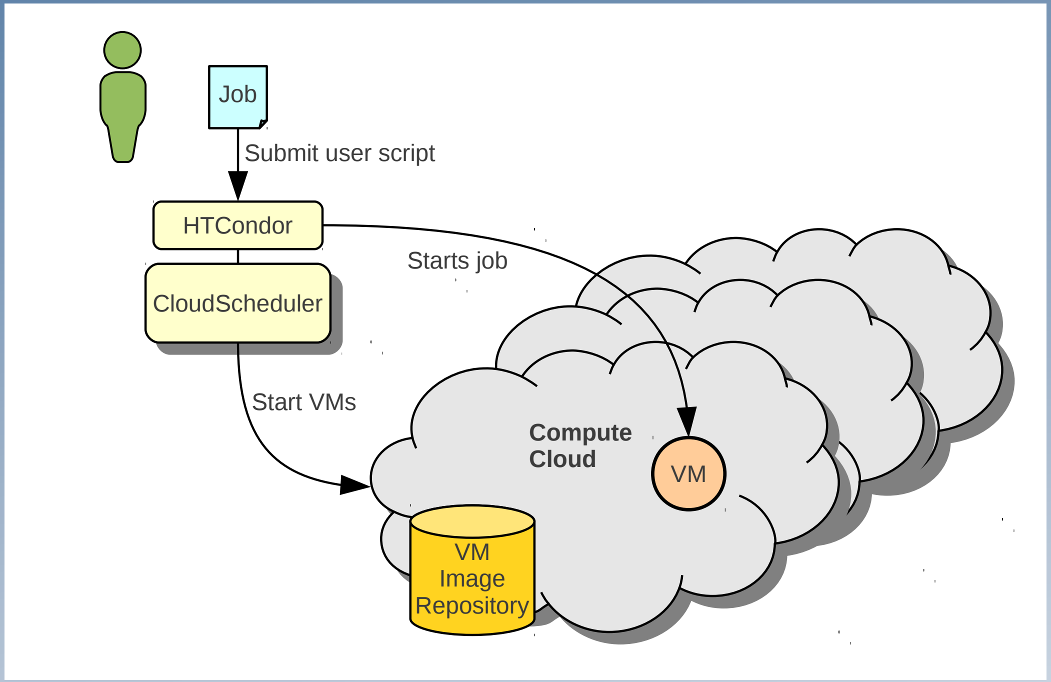
Cloud Scheduler

Custom component

Reviews the requirements of jobs in the HTCondor queue

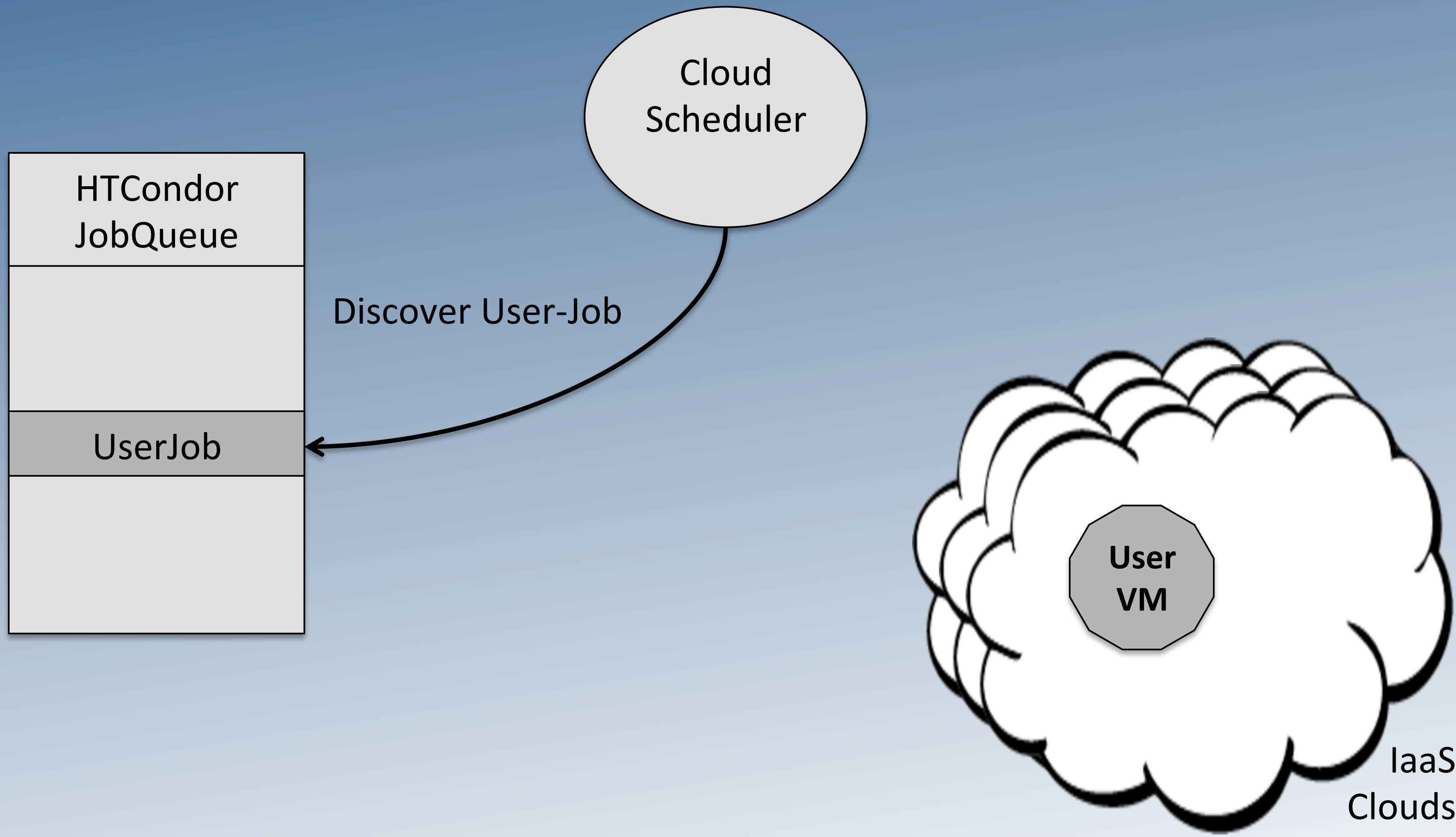
Requests the boot of user images

Monitors the VMs



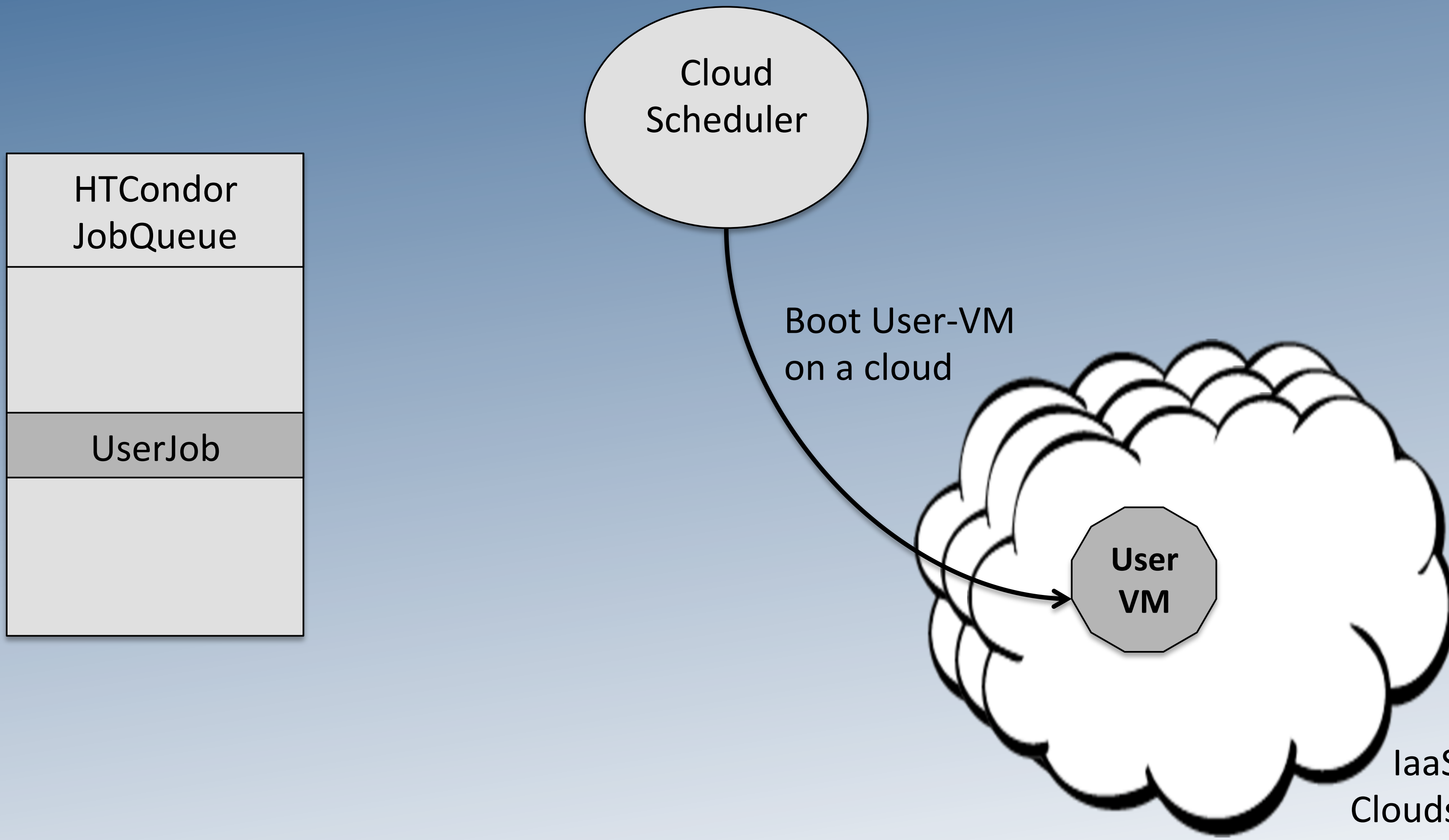
We run multi-core VMs
VM remains active if jobs in queue

Work flow

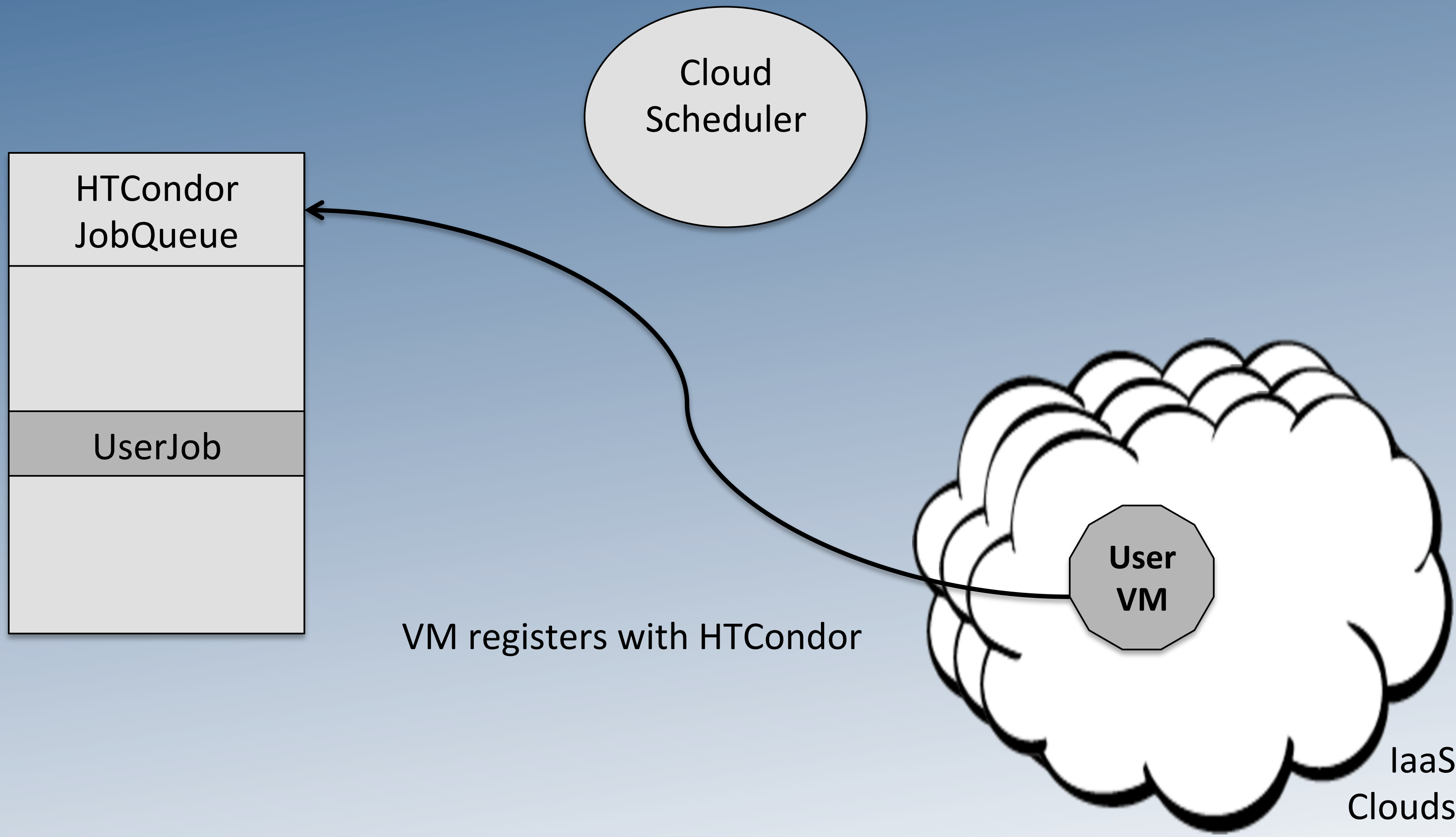


Cloud Scheduler
HTCondor JobQueue
Discover User-Job
UserJob
User VM
IaaS Clouds

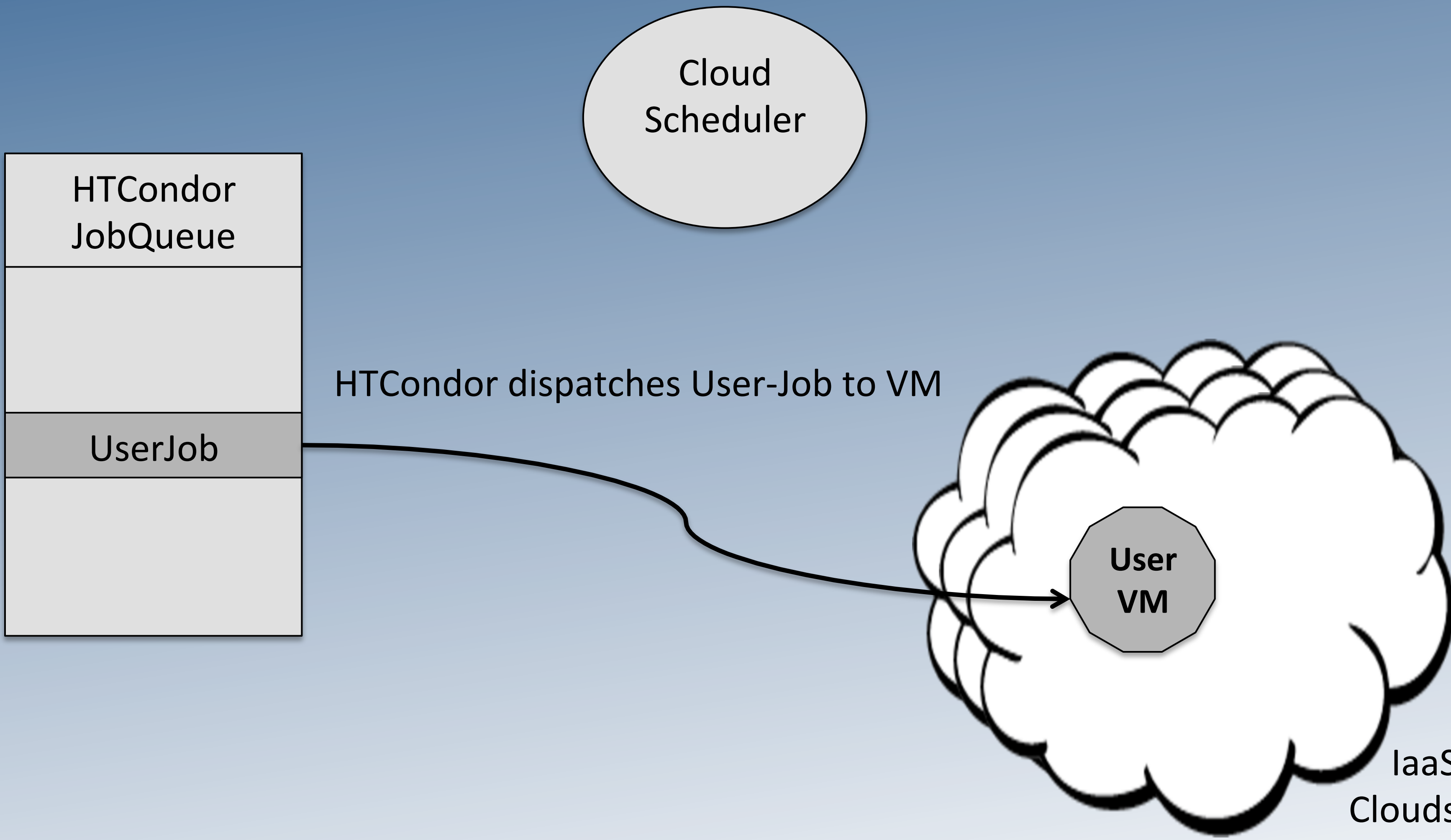
Work flow


HTCondor
JobQueue
UserJob
Cloud
Scheduler
Boot User-VM
on a cloud
User
VM
IaaS
Clouds

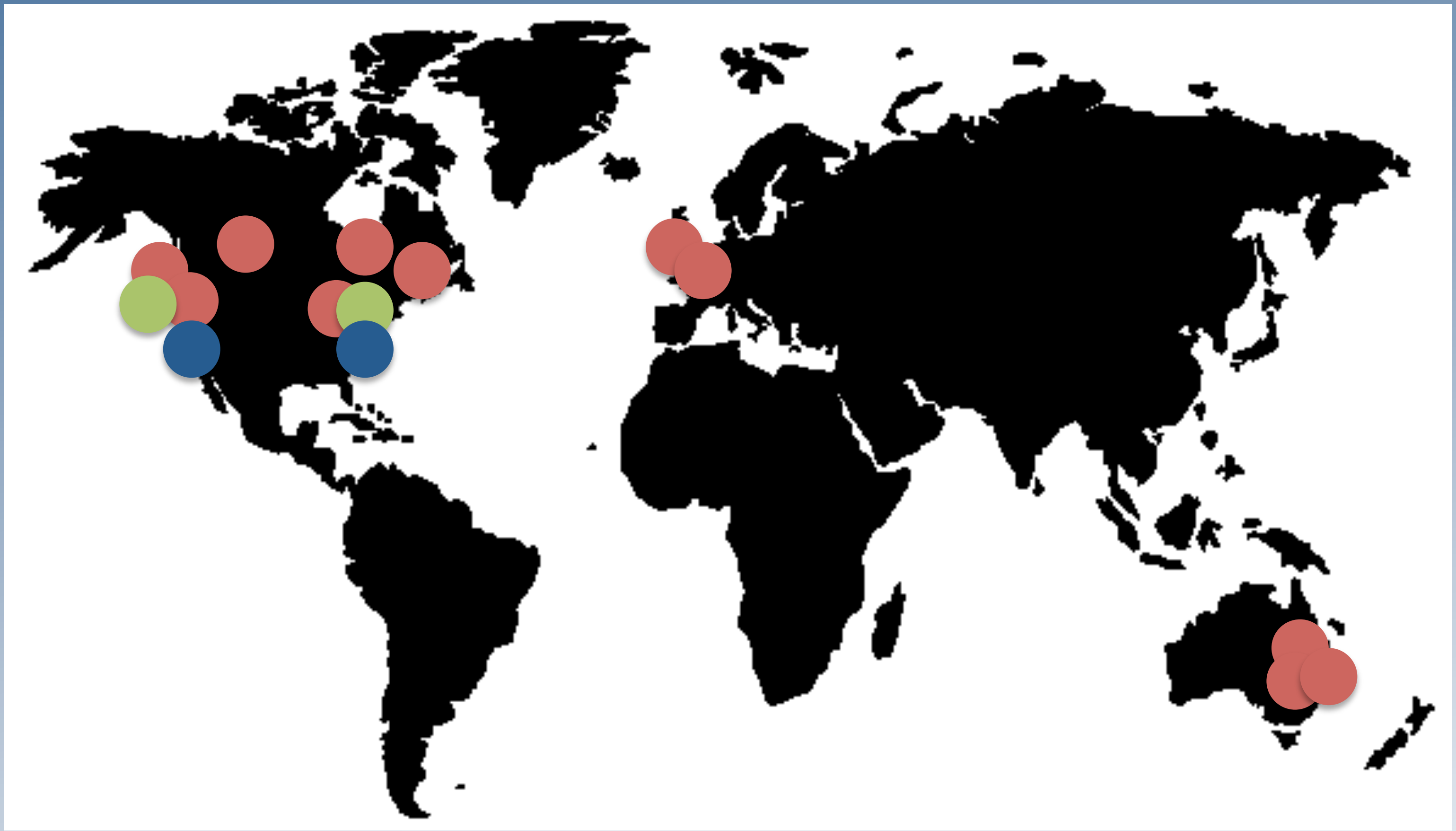
Work flow


Cloud
Scheduler
HTCondor
JobQueue
UserJob
User
VM
VM registers with HTCondor
IaaS
Clouds

Work flow


Cloud
Scheduler
HTCondor
JobQueue
UserJob
HTCondor dispatches User-Job to VM
User
VM
IaaS
Clouds

Cloud resources

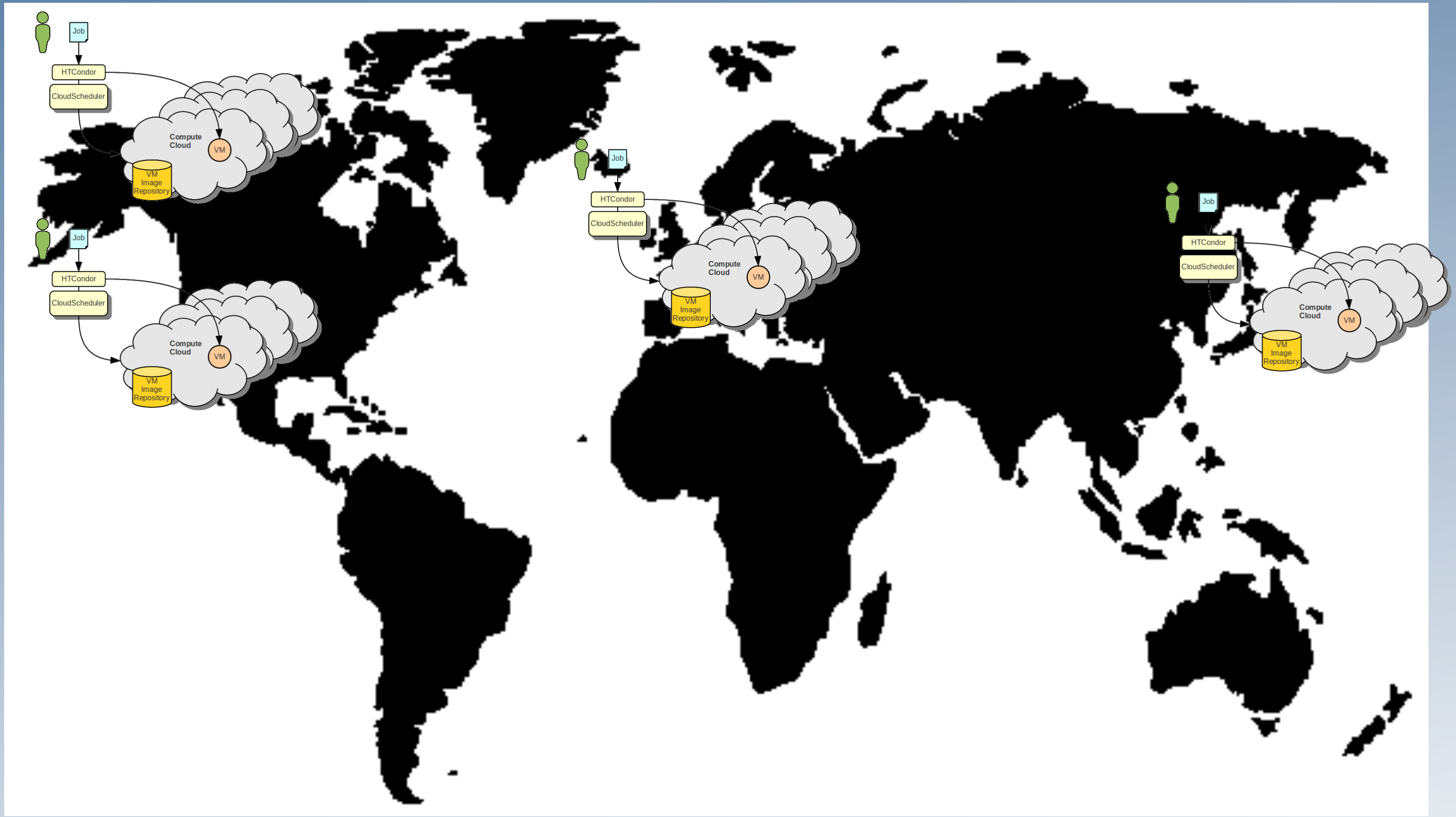


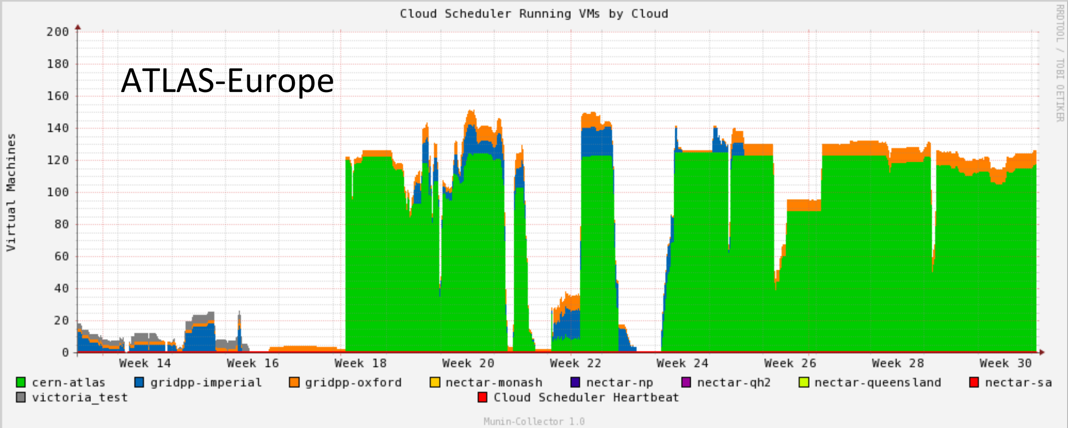
OpenStack   Nimbus   EC2/GCE

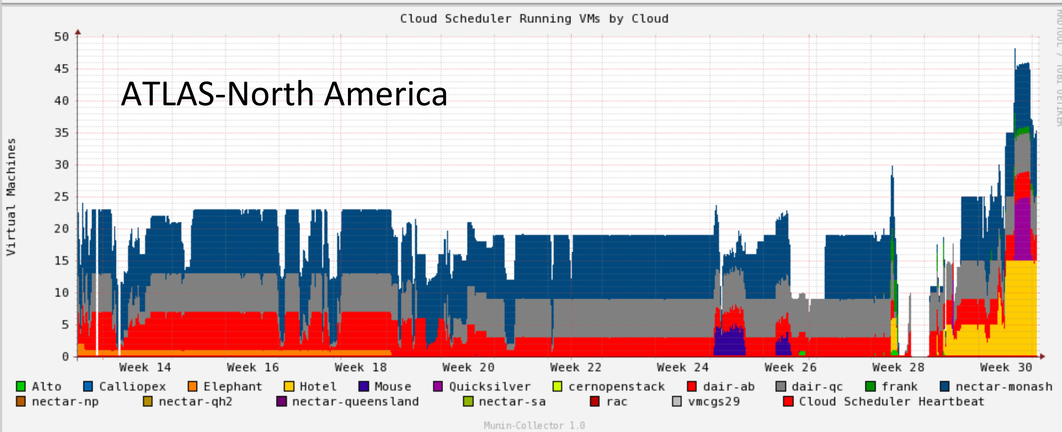
There are four HTCondor/CloudScheduler instances in production

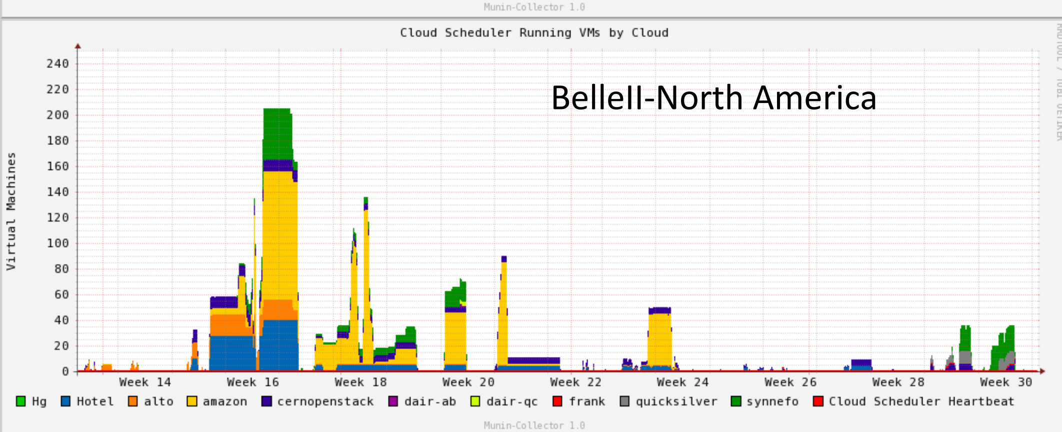
Three HEP (ATLAS-Europe, ATLAS-North America and BelleII), and astronomy

Each has its own set of clouds









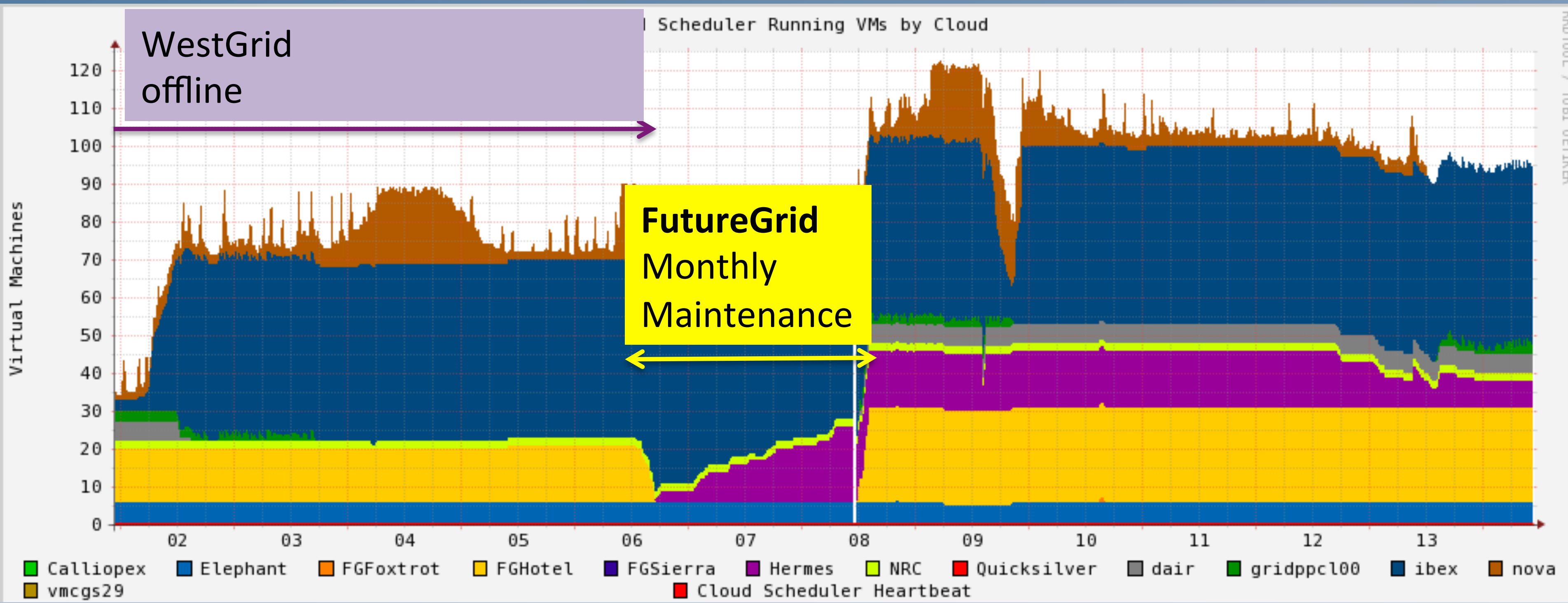
Running VMs

Number of 8-core VMs (15 weeks)

Each colour is a different cloud

Each HTCondor/CloudScheduler instance has different clouds

Cloud evaporation and condensation



Number of 8-core VMs

Operate three HEP HTCondor/CS Instances

ATLAS

North America (2.5 years)

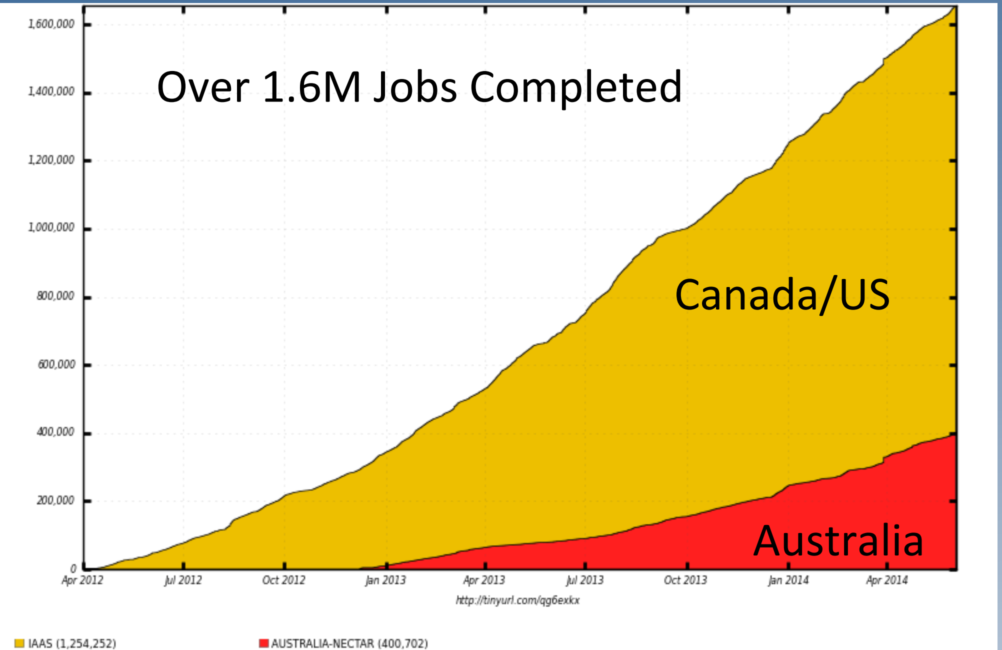
Europe (8 months)

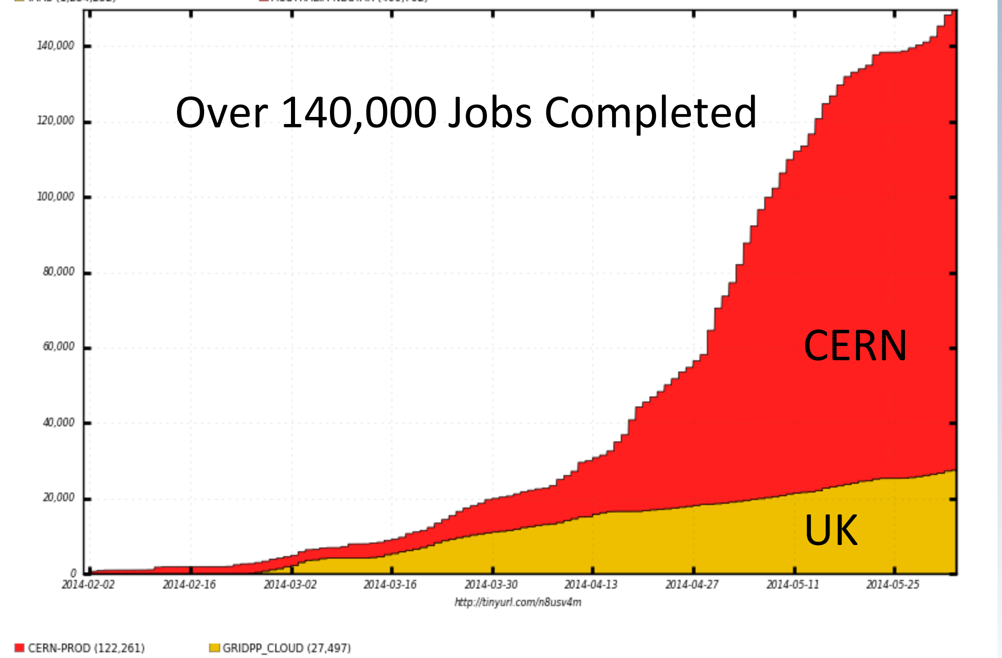
BelleII

North America (8 months)

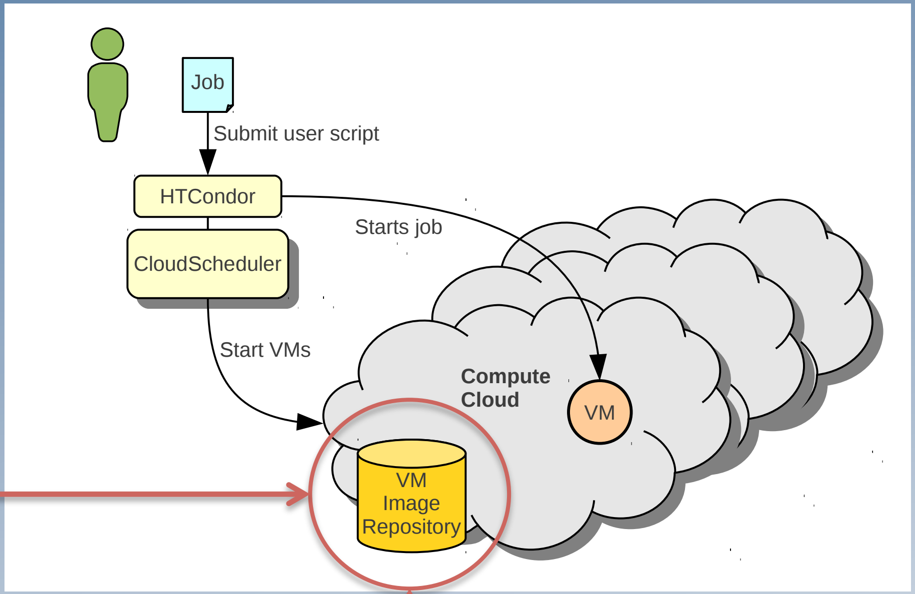
Average work load 1500 cores

Peak work 2500 cores





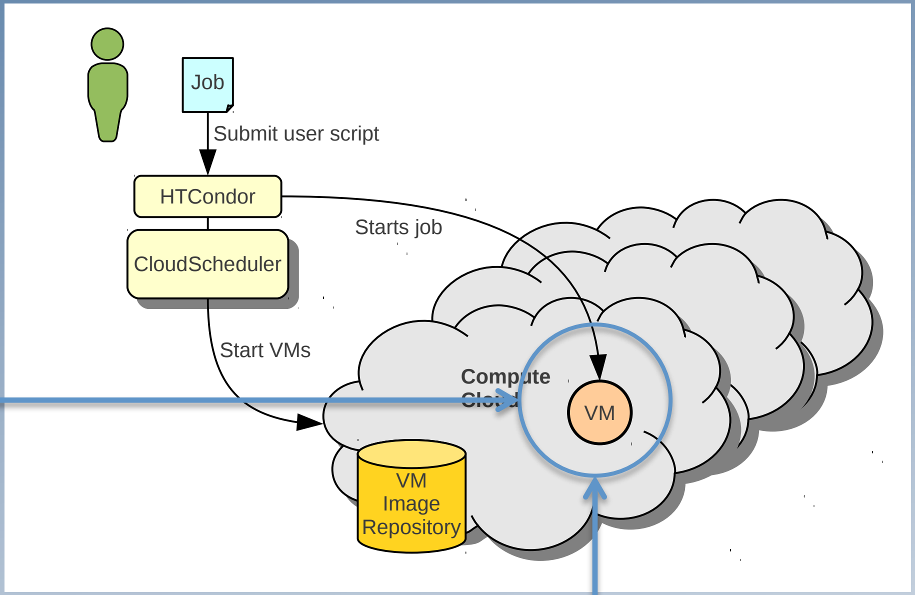
Evolution of the distributed cloud



Managing VM images over multiple clouds

Micro VM images

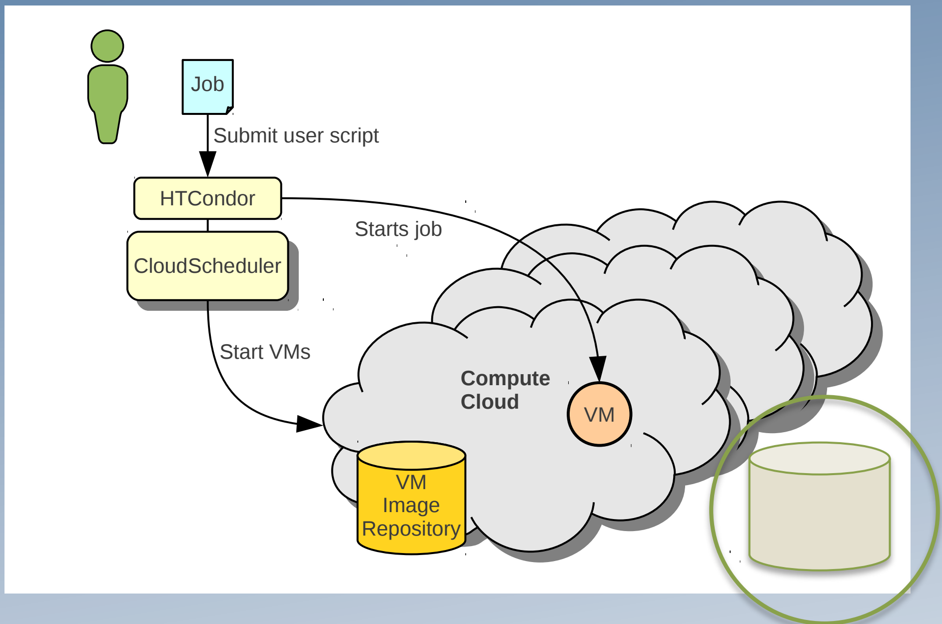
Evolution of the distributed cloud



Remote access to application software

Site-specific contextualization

Evolution of the distributed cloud

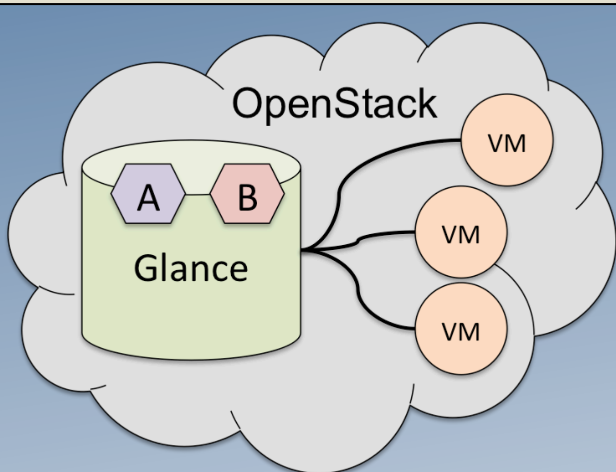


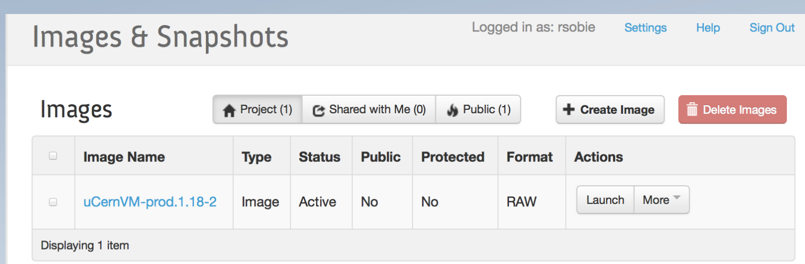
Application data federation
Accessing distributed data repositories

VM management in a distributed cloud

We need to manually upload our images to each site

Each site has different credentials

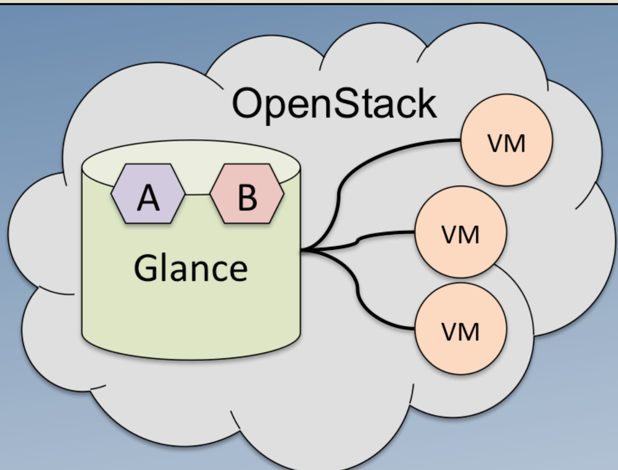


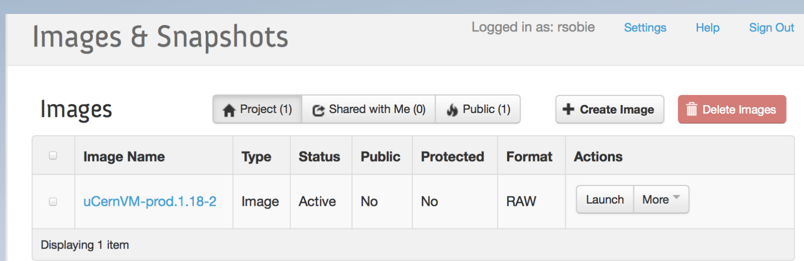


VM management in a distributed cloud

We need to manually upload our images to each site

Each site has different credentials





Images & Snapshots

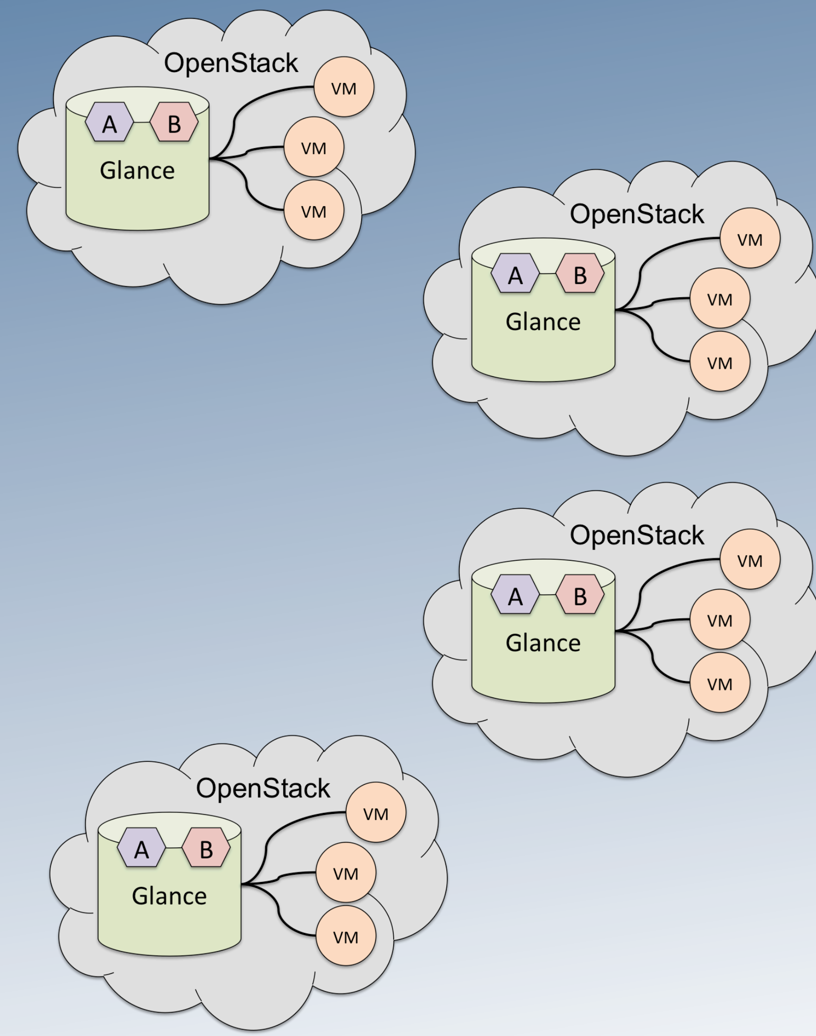
Logged in as: rsobie Settings Help Sign Out

Images

Project (1) | Shared with Me (0) | Public (1) | Create Image | Delete Images

	Image Name	Type	Status	Public	Protected	Format	Actions
	uCernVM-prod.1.18-2	Image	Active	No	No	RAW	Launch More

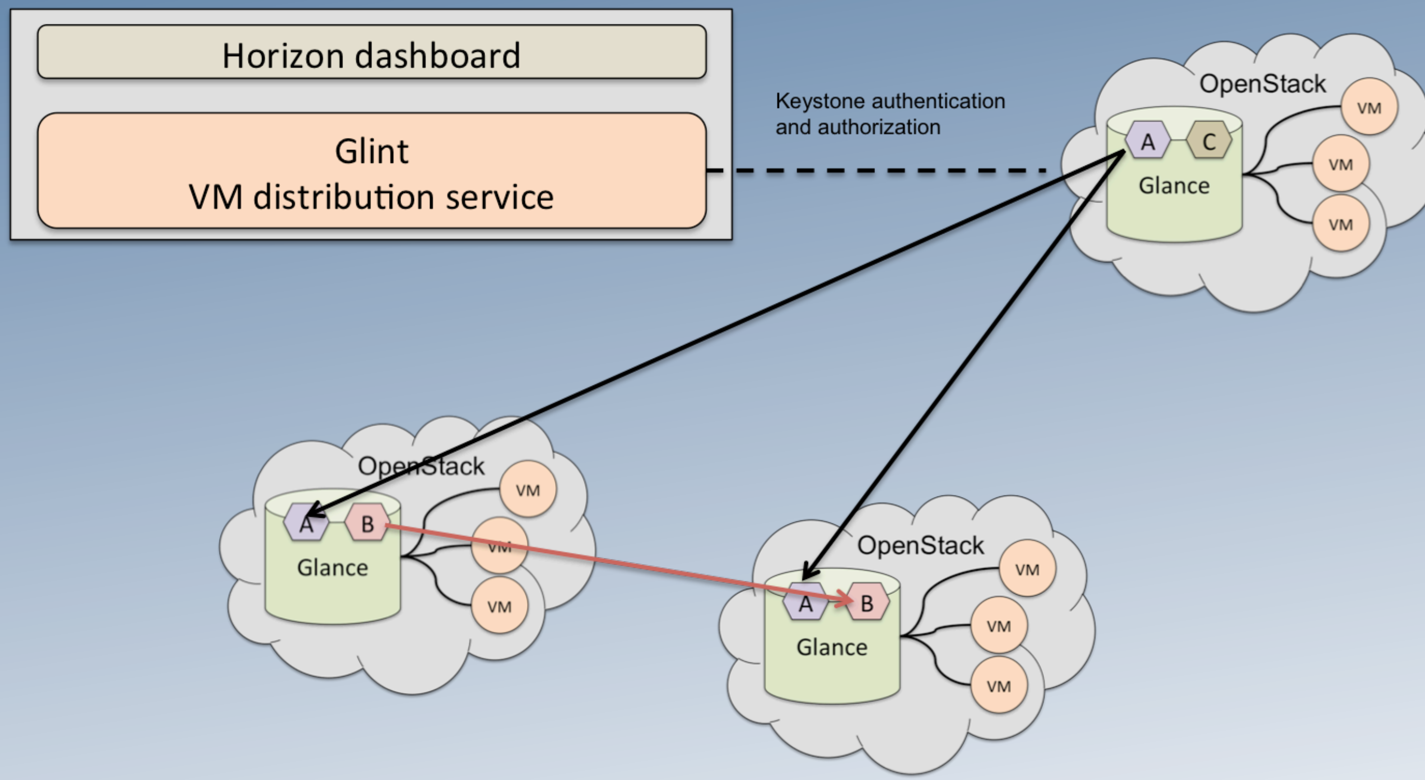
Displaying 1 item



Glint – a VM distribution service

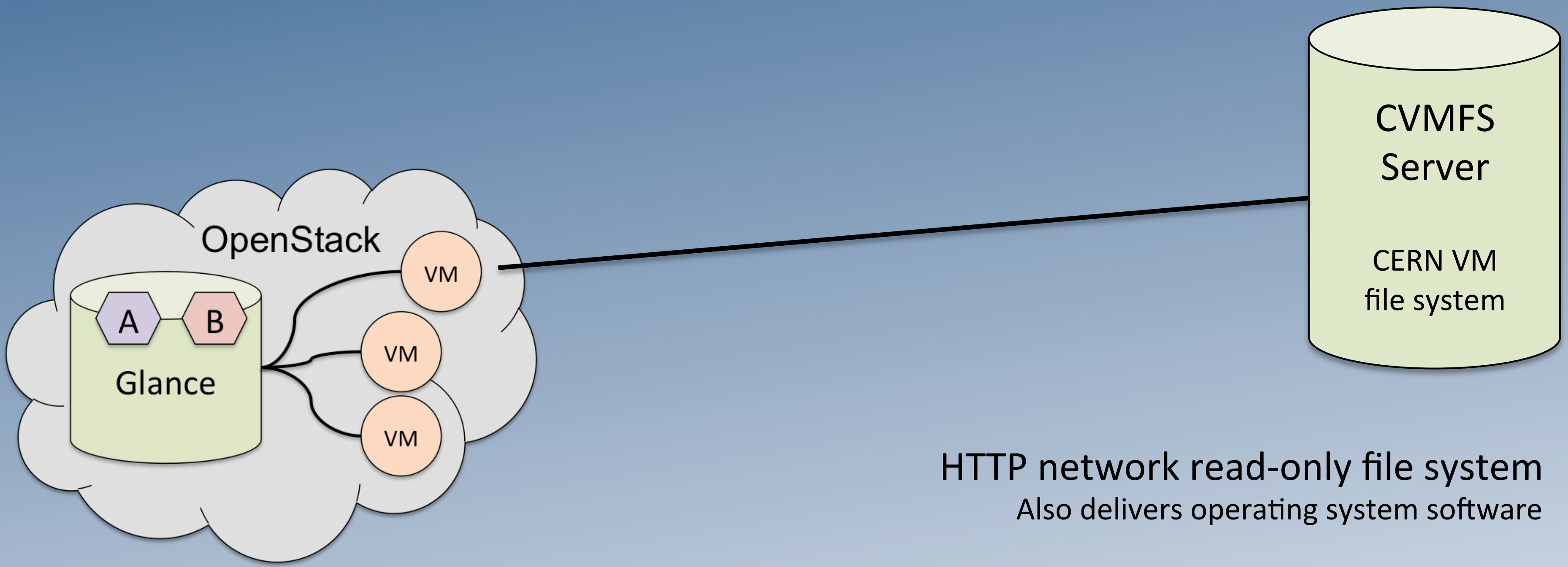
Glint provides the user the ability to manage their images on multiple OpenStack clouds

Integrated in the OpenStack Glance GUI



Exploring how to make it work with other clouds such as EC2 and GCE

Application software distribution I

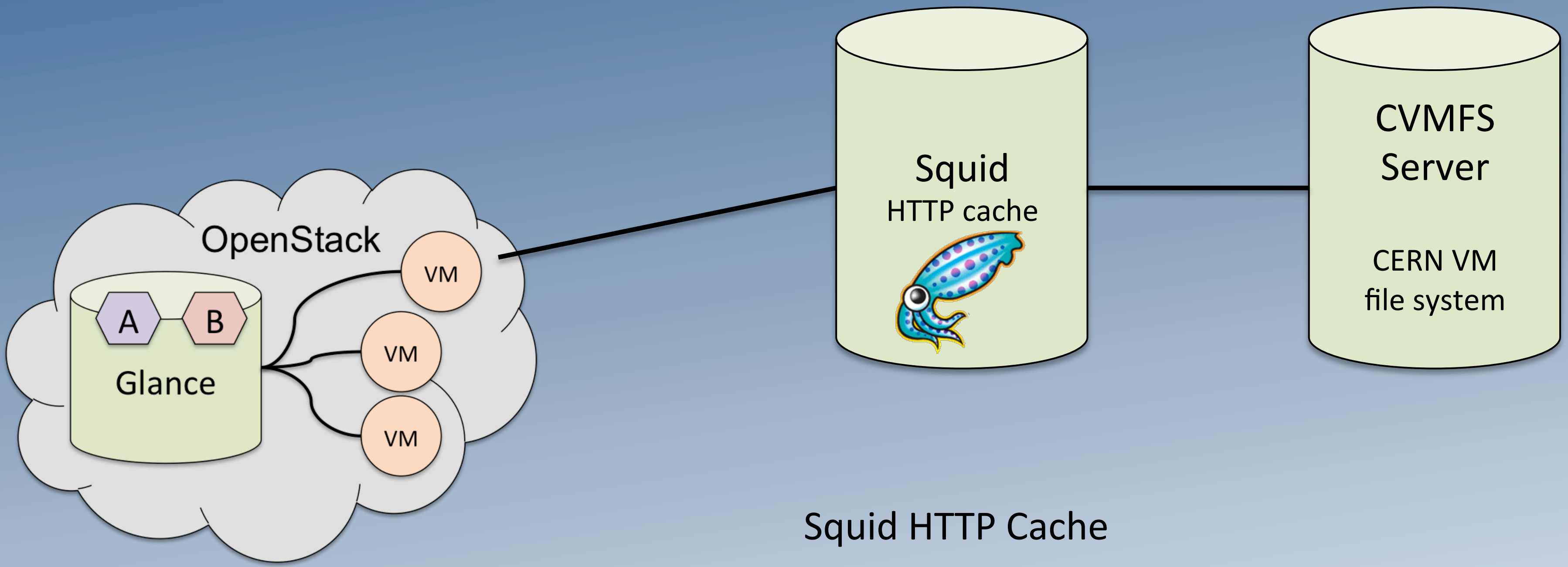


HTTP network read-only file system
Also delivers operating system software

Highly scalable
Multiple levels of caching

Reduces the size of the VM to few MB
Micro-VM images

Application software distribution II

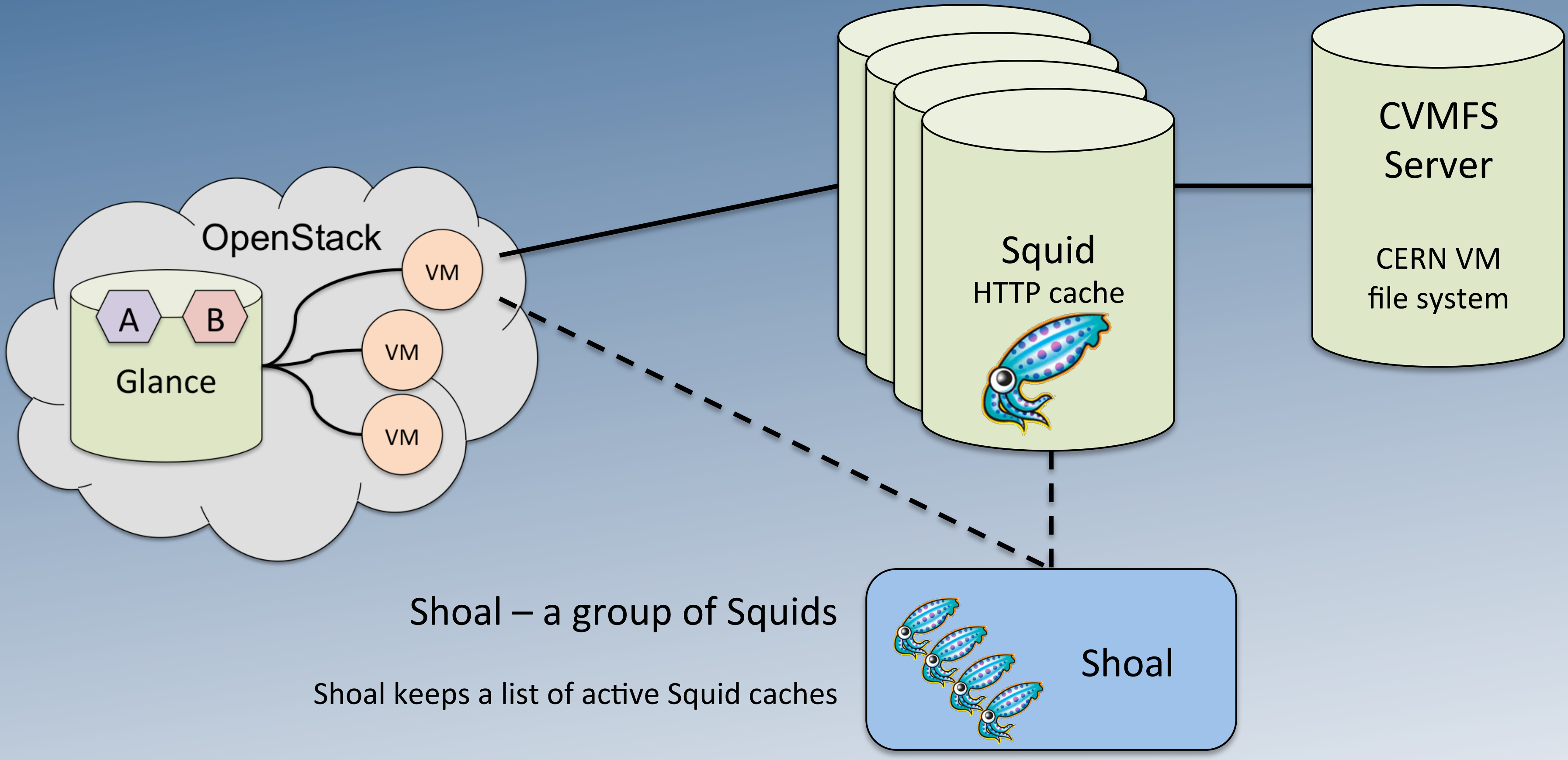


Squid HTTP Cache

Rather than query the CVMFS master repository we use Squid HTTP caches
Approximately 100 Squid caches distributed around the world

Each VM uses a statically assigned Squid cache

Application software distribution III



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

Shoal: http://arxiv.org/pdf/1311.0058v1.pdf
github.com/hep-gc/shoal

Micro CERN VM

Micro-CERN VM, Bloomer etal
http://iopscience.iop.org/1742-6596/513/3/032009/pdf/1742-6596_513_3_032009.pdf

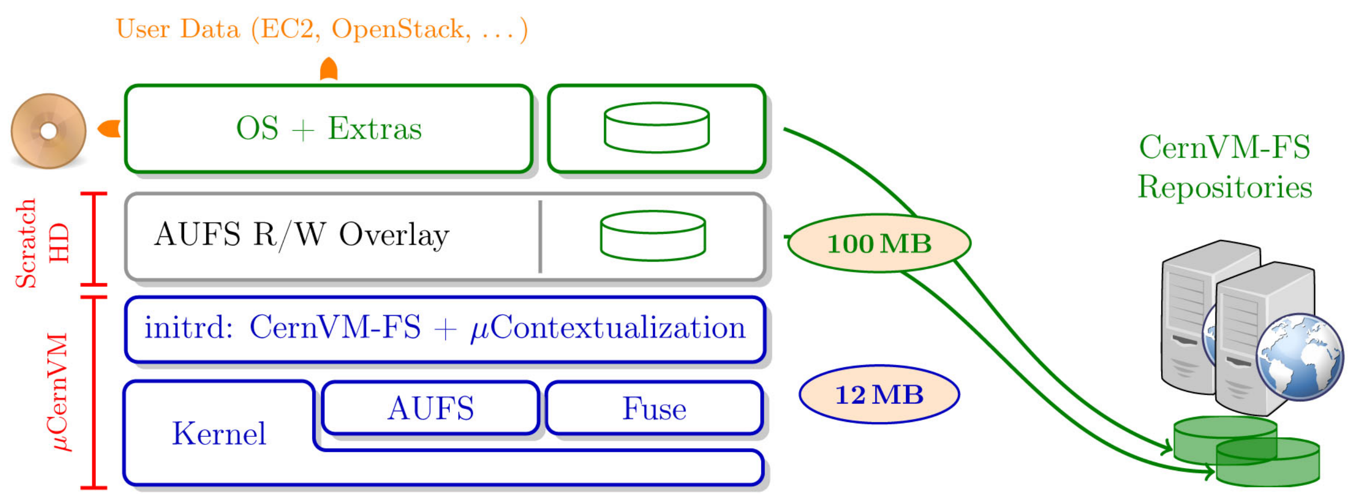
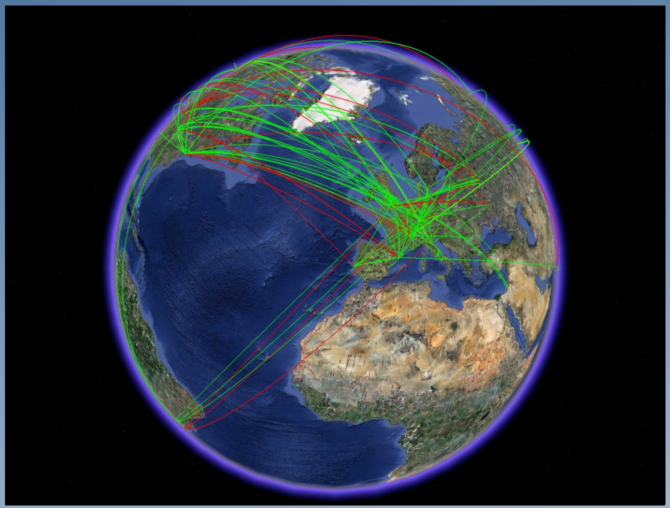


Figure 1. A μCernVM based virtual machine is twofold. The μCernVM image contains a Linux kernel, the AUFS union file system, and a CernVM-FS client. The CernVM-FS client connects to a special repository containing the actual operating system. The two CernVM-FS repositories contain the operating system and the experiment software.

VM (12 MB) with "just enough operating system" and a CVMFS client
Used in production operation starting July 2014

Data Management



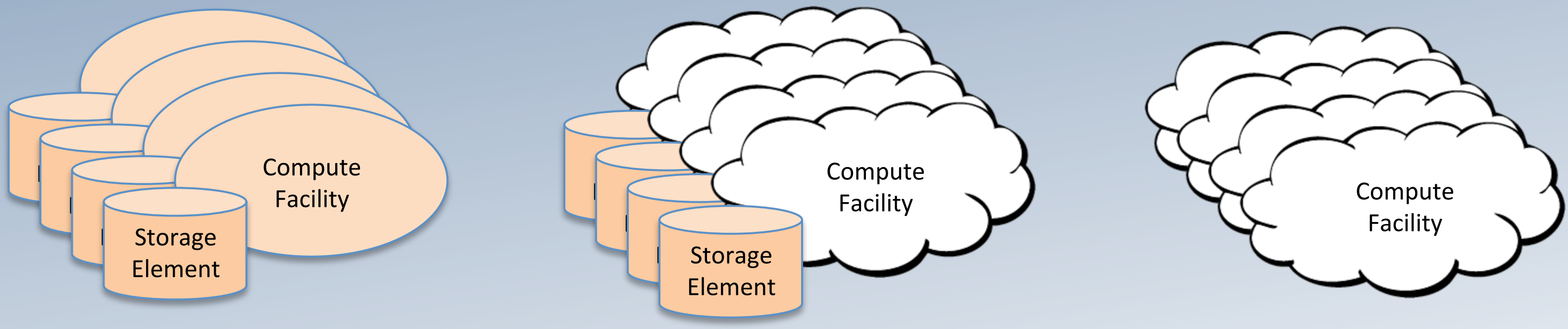
Approximately 100 computing sites

Each HEP site has CPU and storage

10-100G network

We want to use the data at any SE on all clouds

Storage Element (SE)

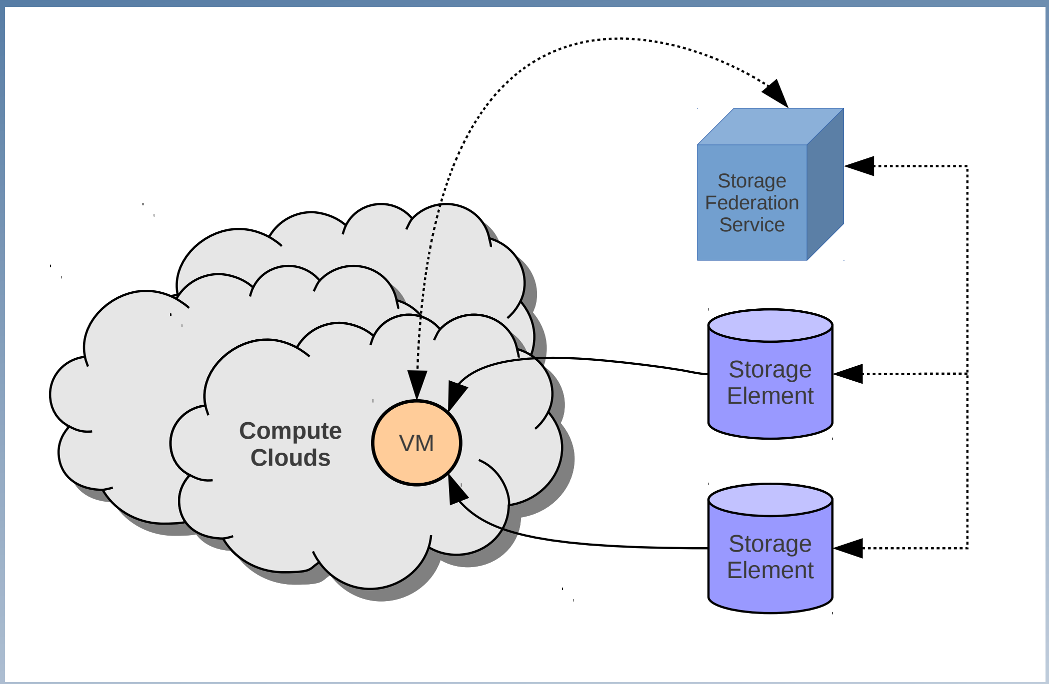


Non-cloud HEP

Cloud HEP

Non-HEP cloud

Data Federation



Generic Data Federator

Developed at CERN

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

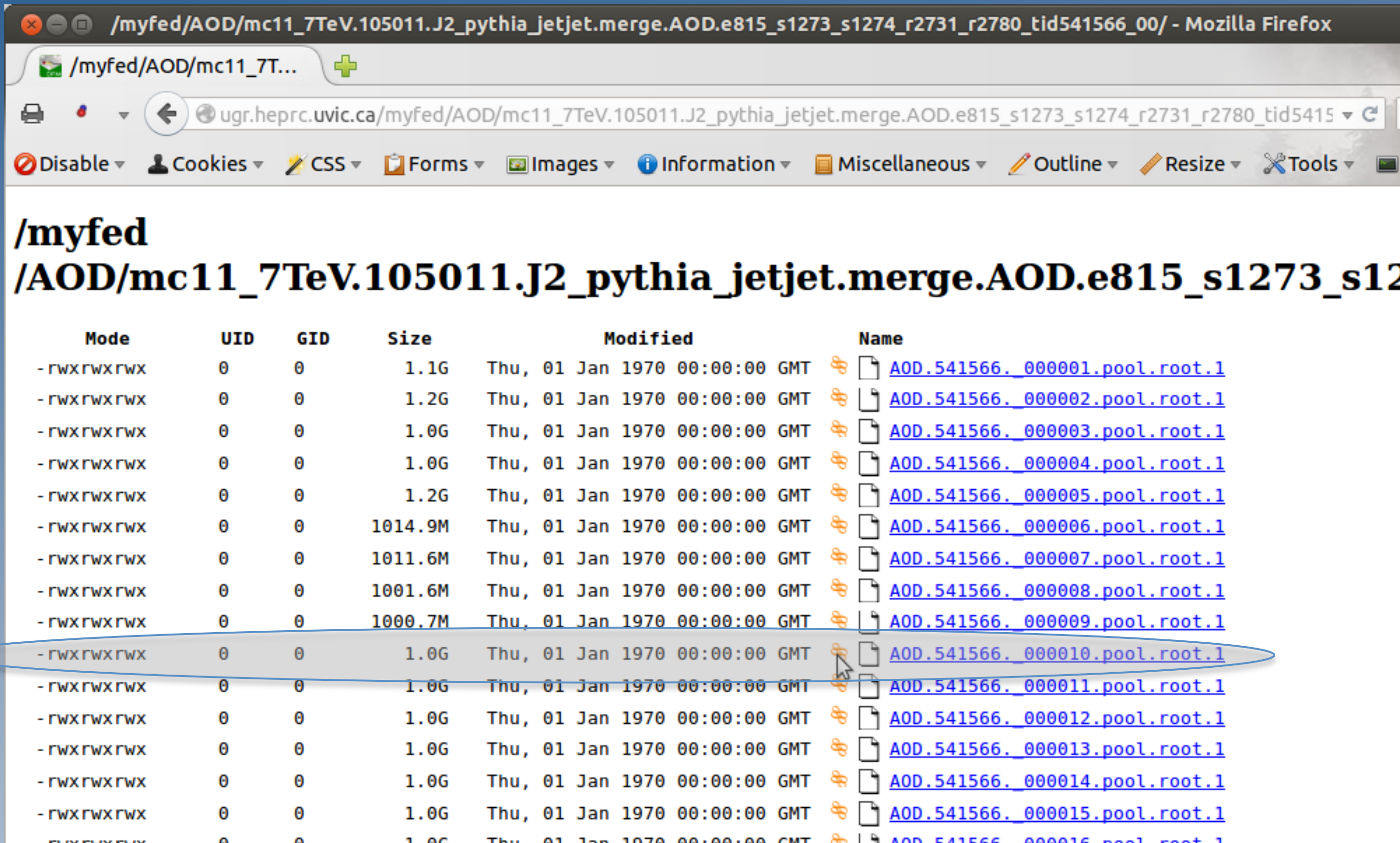
Dynamic discovery

Directs the VM to the nearest Storage Element with the input data

GeoIP information to pick the closest site

We use WebDAV for file access

Open-source HTTP protocol standard


/myfed/AOD/mc11_7TeV.105011.J2_pythia_jetjet.merge.AOD.e815_s1273_s1274_r2731_r2780_tid541566_00/ - Mozilla Firefox

/myfed/AOD/mc11_7T...

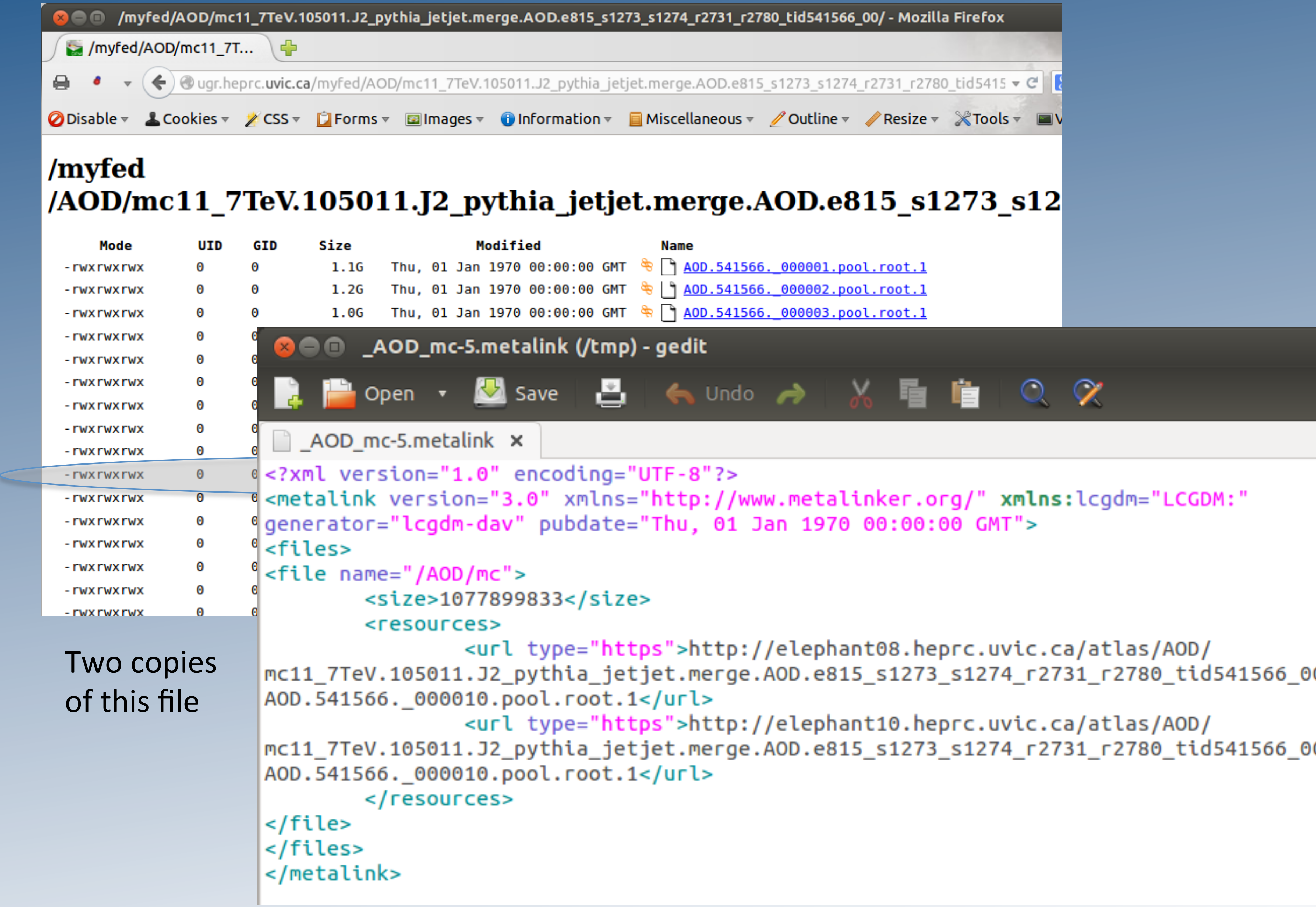
ugr.heprc.uvic.ca/myfed/AOD/mc11_7TeV.105011.J2_pythia_jetjet.merge.AOD.e815_s1273_s1274_r2731_r2780_tid5415

Disable Cookies CSS Forms Images Information Miscellaneous Outline Resize Tools

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

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



Two copies of this file

Networks are critical to distributed clouds

Data federations are only becoming feasible with 100G research networks
http://supercomputing.uvic.ca

BBC | News | Sport | Weather | Capital | Futu

NEWS TECHNOLOGY

Home | US & Canada | Latin America | UK | Africa | Asia | Europe | Mid-East | Business | Health

14 December 2011 Last updated at 10:16 ET

Share

Scientists break world record for data transfer speeds

BCNET
Shared IT Services for Higher Education & Research in British Columbia

Home | Members | Services | Community | Network | News & Events | About

Home › News › 2013 › December › 19 › International team demonstrates intercontinental OpenFlow network at Super Co

International team demonstrates intercontinental OpenFlow network at Super Computing 2013

Caltech

About Caltech | News & Events | Research

Back to All News

11/21/2012

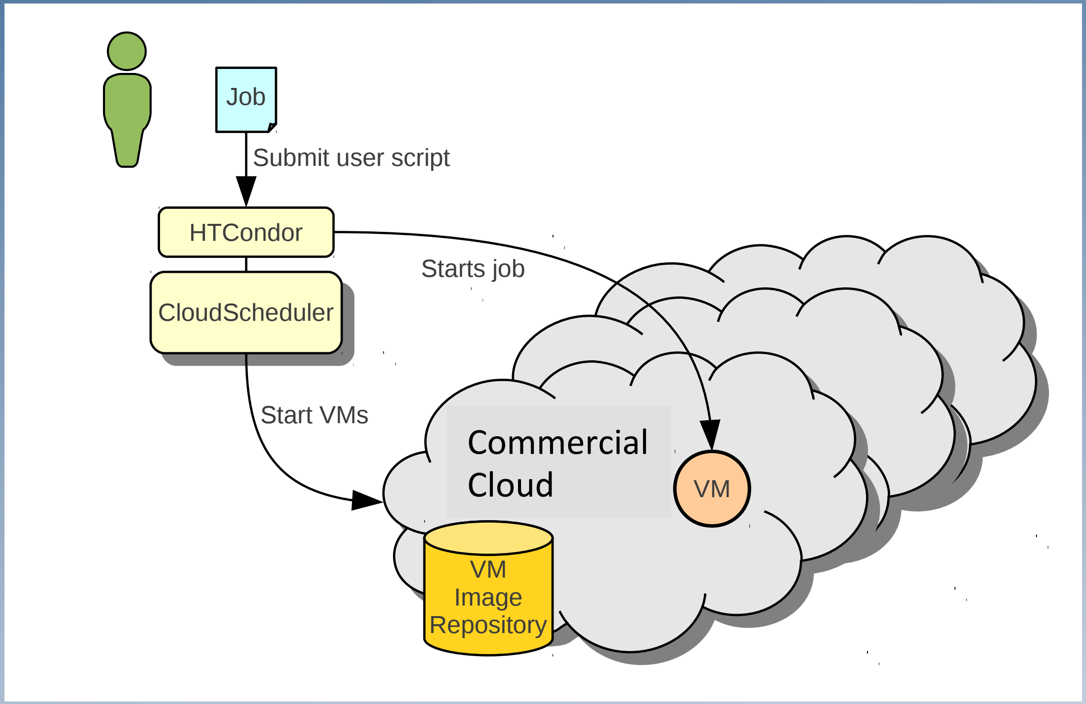
High-Energy Physicists Smash Records for Network Data Transfer

New methods for efficient use of long-range networks will support cutting-edge science

canarie

Canadian physicists achieve 100 gigabit/second transatlantic transmission, enabled by CANARIE and its global partners

Cloud bursting



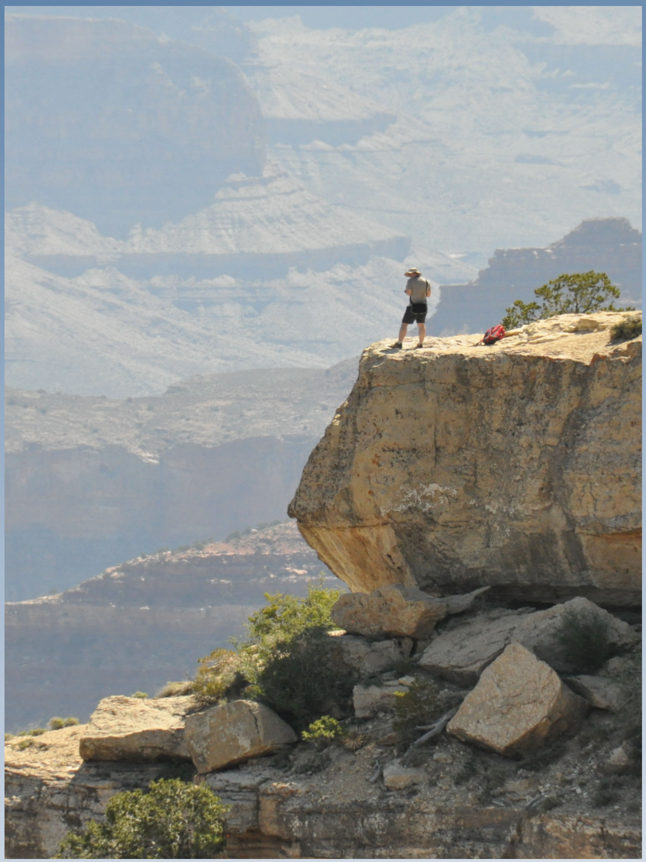
Cloud Bursting
Using commercial clouds when our existing facilities are fully utilized

Amazon EC2
1000 cores
Spot price tracker

Google Compute Engine
Under test

Each commercial cloud has its own configuration and complexity
Not all sites are connected to the research network

Issues and Challenges



We are often the first user of a new cloud

Clouds are not configured in the same way
(meta-data is often inconsistent)

Cloud software is rapidly changing
(OpenStack did not exist a few years ago)

ID management and authentication

HEP applications stress the non-HEP clouds
(many clouds oversubscribe their resources)

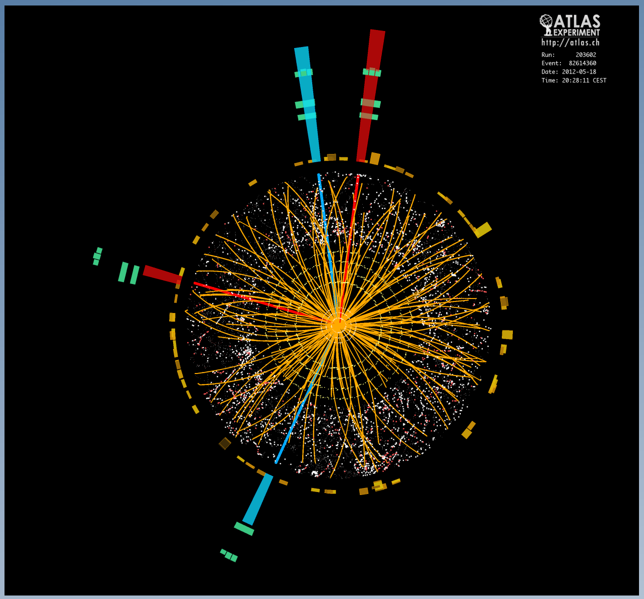
Future

Distributed services
VM images
Software caches
Data federations

Cloud management
Automate integration and testing of new clouds
Standardized cloud configuration
Common authentication and authorization
Improved monitoring

Commercial clouds
Viable solution
Cost trend

Summary



We have designed and constructed a distributed cloud computing system

Currently used for HEP and astronomy applications

Uses IaaS clouds distributed around the world

We are adding new IaaS clouds to the system

We are enhancing its dynamic capabilities (context-aware cloud computing)

Exploring the use of commercial resources for cloud-bursting

Acknowledgements

Contact information: rsobie@uvic.ca

Web sites: http://rjs.phys.uvic.ca/

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

Site-specific contextualization

VM contextualization process:

CloudScheduler boots the user-VM and includes the user-data (specified in the Job Description file)

During the VM boot sequence:
cloud-init interprets the user-data and configures the Puppet-agent on the VM

The Puppet-agent on the VM connects to our Puppet master to retrieve the configuration for the VM