# *Clouds for research computing*

Randall Sobie
Institute of Particle Physics
University of Victoria

Collaboration
UVIC, NRC (Ottawa), NRC-HIA (Victoria)

## *Outline*

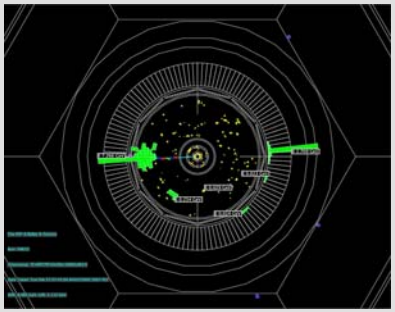
- Research computing
- Cloud computing
- Clouds for research computing
- Distributed research cloud for particle physics and astronomy
- Outlook

# *Research Computing*

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

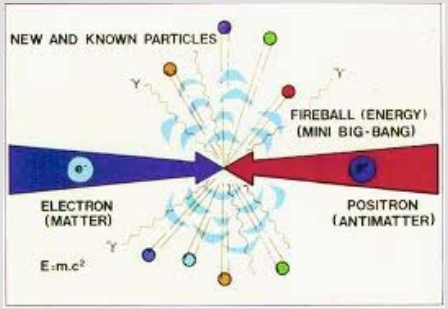
We build instruments large detectors to record the collisions of matter and antimatter

SLAC National Accelerator Lab



Record billions of particle collisions

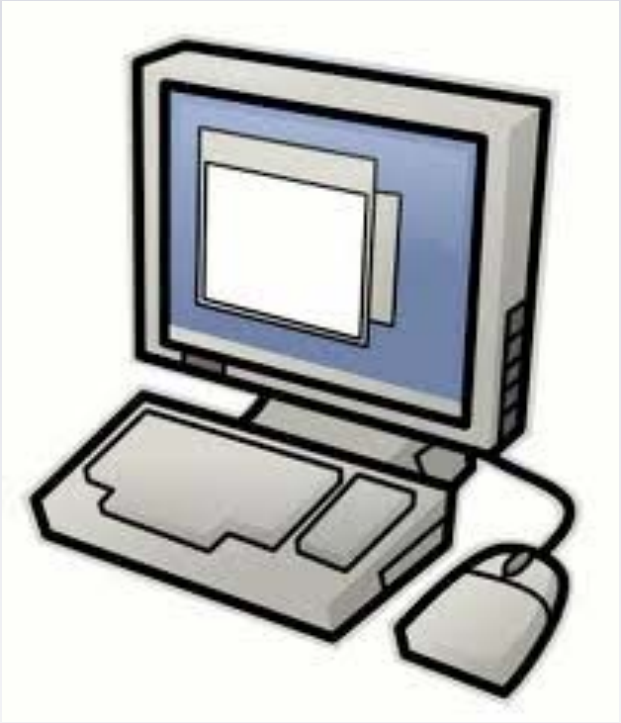
"Events"



# *Research* ***Computing***

**Desktop computing**

12+ processor cores
10+ TB storage
10G network cards



**Applications**

Interactive computing
Visualization
Remote sensor control

# *Research* ***Computing***

**Cluster computing**

300+ cores/rack
1000+ TB storage
Interconnectivity

**Application**

High Throughput HTC
Small scale parallel jobs
(physical sciences)

# *Research* ***Computing***

**Super Computers**

TOP500
100K cores

High-speed
interconnectivity

**Applications**

Climate modeling
Cosmology

# *Research* ***Computing***

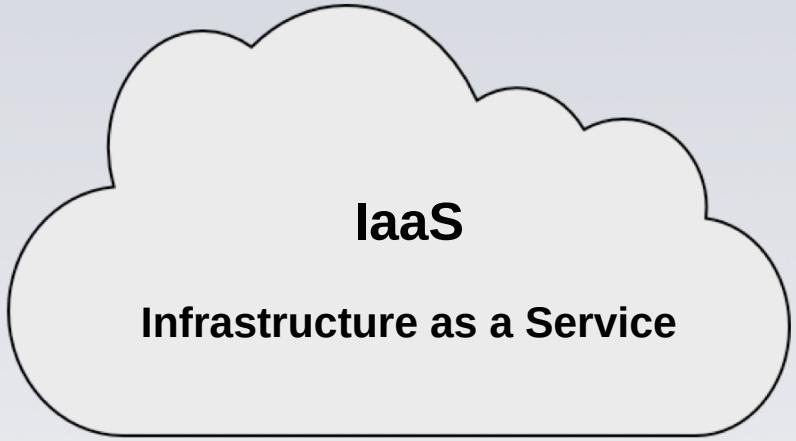
**Computing Grids**

Distributed
compute clusters
100K+ cores
1000TB ++ storage

**Applications**

Very High Throughput
Particle physics
Earth sciences
Health sciences

# *Cloud computing*

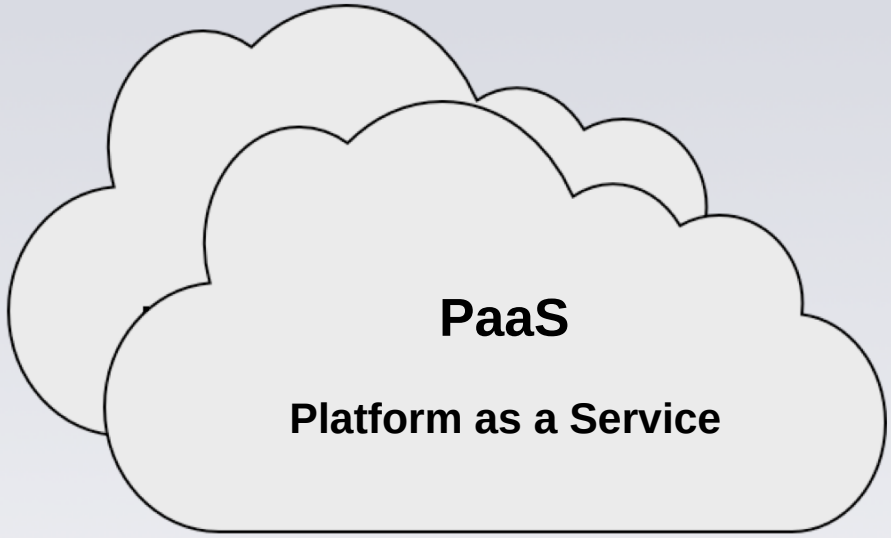



IaaS

***Delivers computing infrastructure as a service***
Science clouds
Commercial clouds (EC2)

HTC platform

# *Cloud computing*

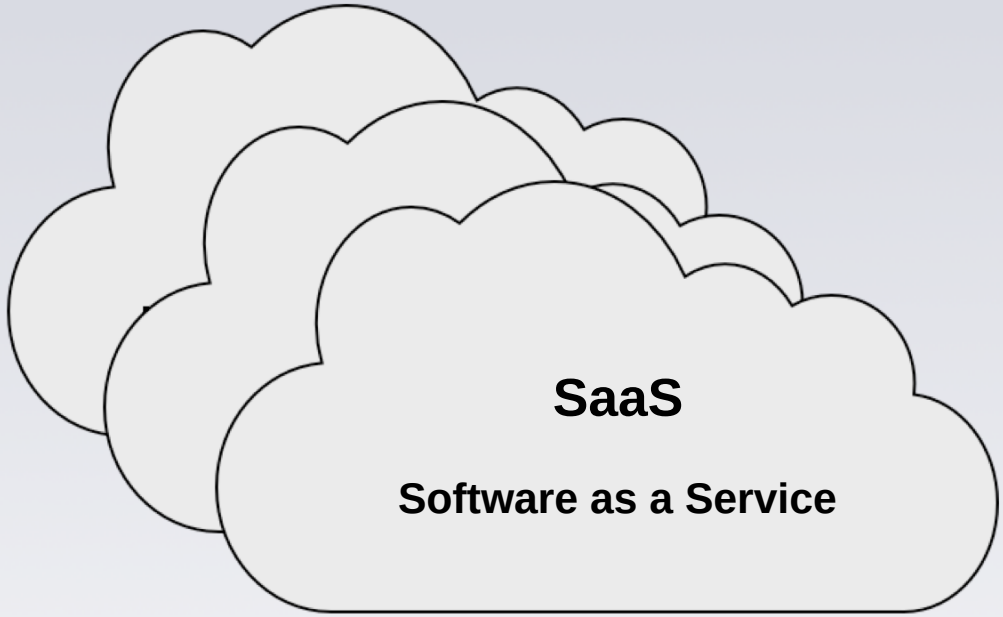



PaaS

***Delivers computing platform or software stack as a service***
Commercial clouds (EC2)

eg Instrumentation control

# *Cloud computing*




SaaS

***Delivers applications or software stack as a service***
Commercial clouds (EC2)

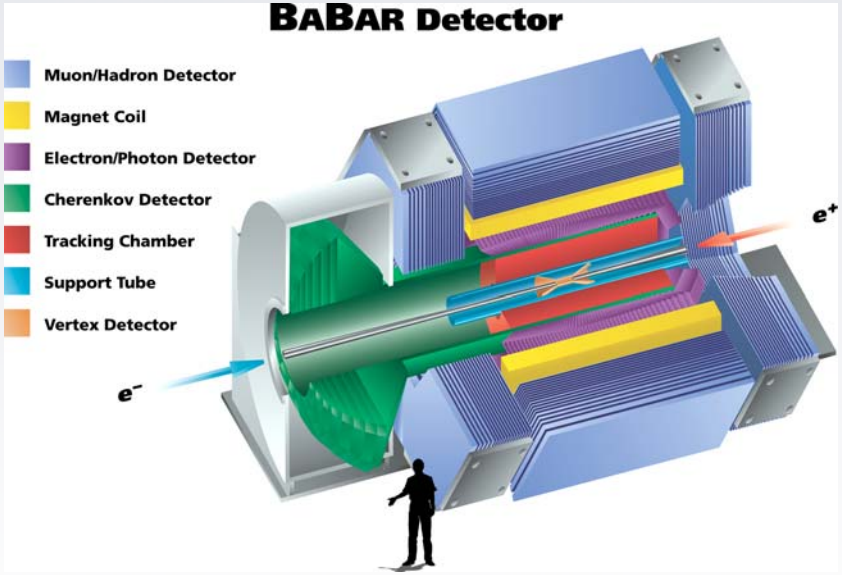
eg Mathematica or Matlab

# *Why use clouds?*


The BaBar project stopped recording data in 2008.

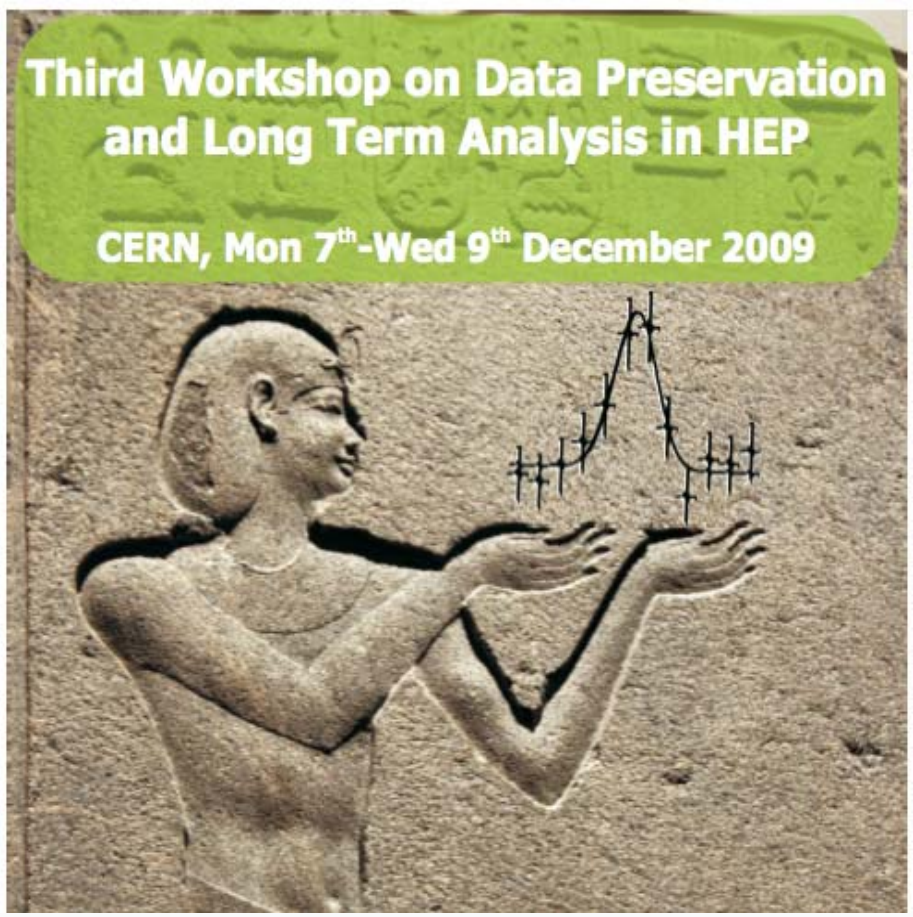
Complex code developed over 15 years
Limited to specific operating systems and libraries
Diminishing resources and few people.

Virtualization is the only solution for preserving the software environment



# *Data preservation*

***Topical issue for many research fields***




Particle physics community has largely ignored this issue often because new facilities make the old data obsolete.

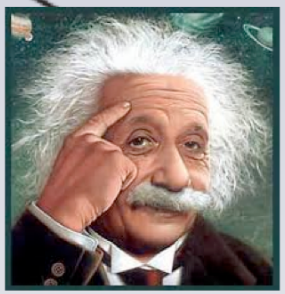
Projects are now decades long and unlikely to be repeated for many years.

We need to ensure the data is accessible for the long term.

**Challenge**

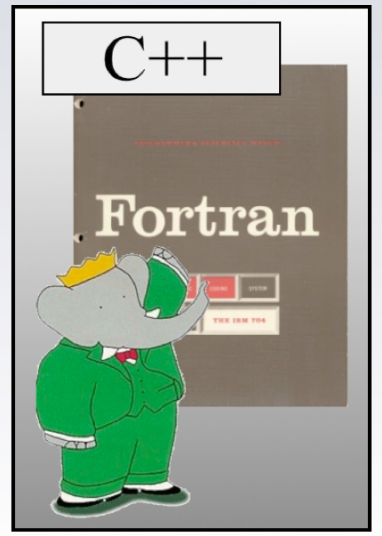
*We need to preserve the software environment*

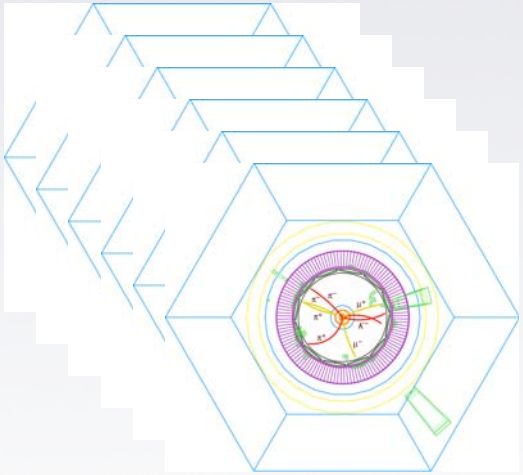
# Requirements



**Sophisticated user communities**
Non-GUI users
Batch computing

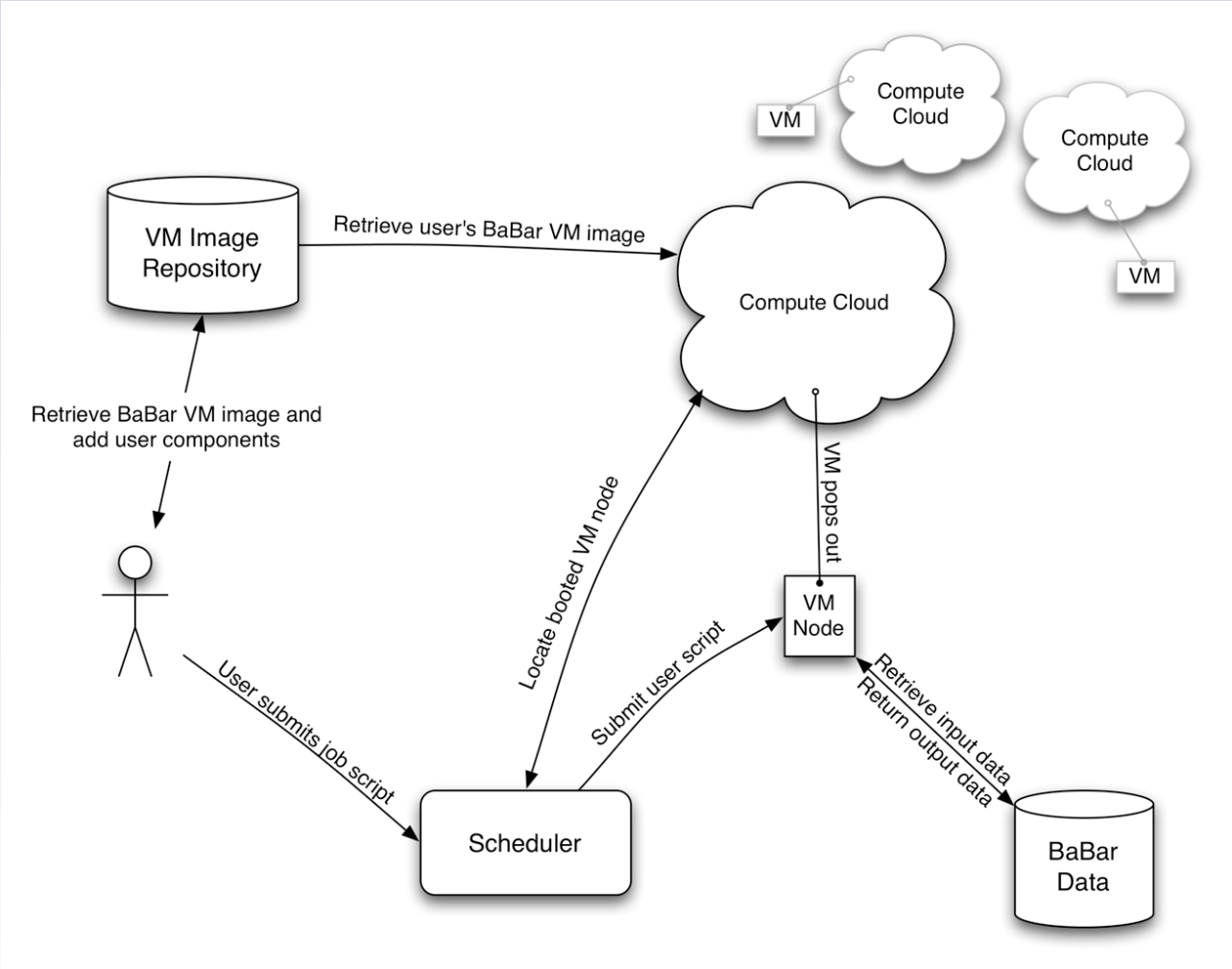
**Complex software packages and demanding system requirements**
Specific OS system
Specific application libraries and compilers



**Medium-scale data sets (100s TBs)**
Data accessed (on-demand) from remote repositories

# *Conceptual design of a distributed cloud*




*Sky Computing* or *Grid of Clouds*

**Design goals:**

Leverage existing work in grid computing (authentication, data management, networks)

Use existing research computing facilities and get access to new resources

Boot user-customized VMs in a familiar batch computing environment

Simplify systems configuration by removing the application dependence

Use the network to move data to the clouds

# Components

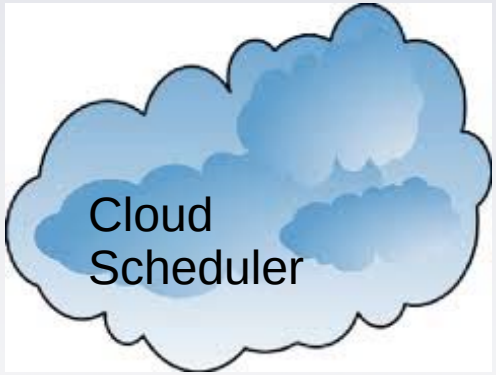
**Application encapsulation**
**Image replication**
**eg Xen, KVM**



**Clouds**
**WS interface**
**eg Nimbus, OpenStack, EC2**

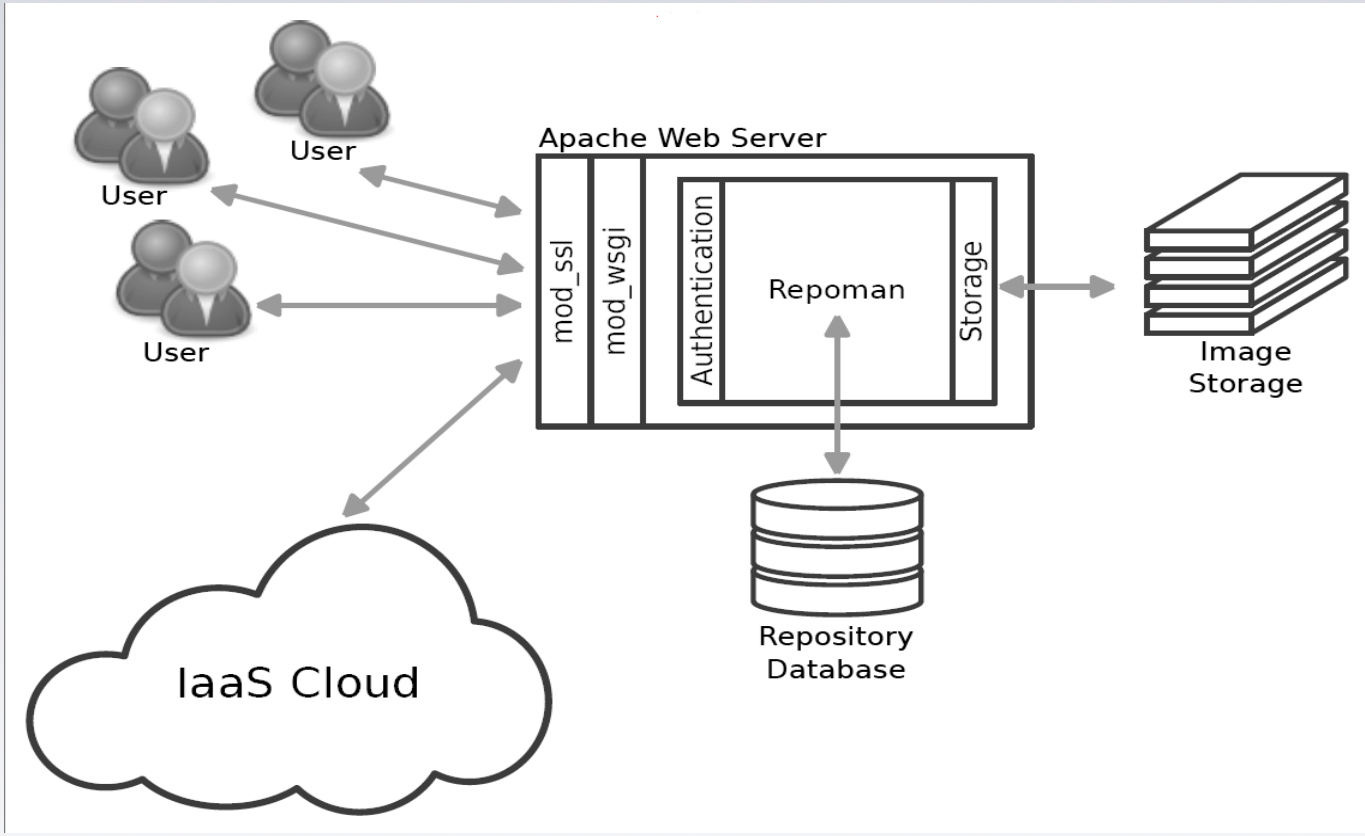
**Dynamic resources**
**eg Condor, SGE**

Job Scheduler




**Managing multiple clouds**
**eg Cloud Scheduler**

# VM management : Repoman



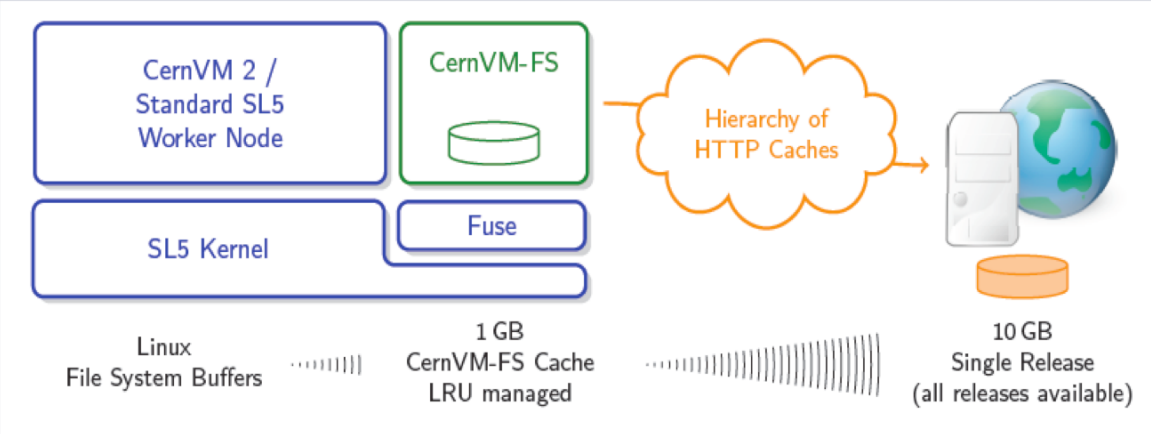

***Repoman***

Is an image repository for storing standard or base VMs.

It is also used by the users to store and manage their images.

The user adds the http link to the VM into their job requirements file.

# CERN-VM Filesystem (CVMFS)




**CVMFS** is a virtual software appliance designed to provide dynamic installation of application software

CVMFS retrieves only the files required by the user

Implemented as a FUSE module (File System in User Space)

User requires an https/http connection to a read-only file system

**Advantages:**

- BaBar software suite is 30G
- User analysis reduces size to 4 G
- Boot/save times significantly decreased
- Reduced load on network transfer of VM images

# *Authentication*

We use X509 certificates for authentication (except Amazon EC2)

Used in particle physics (LHC/CERN) and also by Westgrid (Compute Canada)

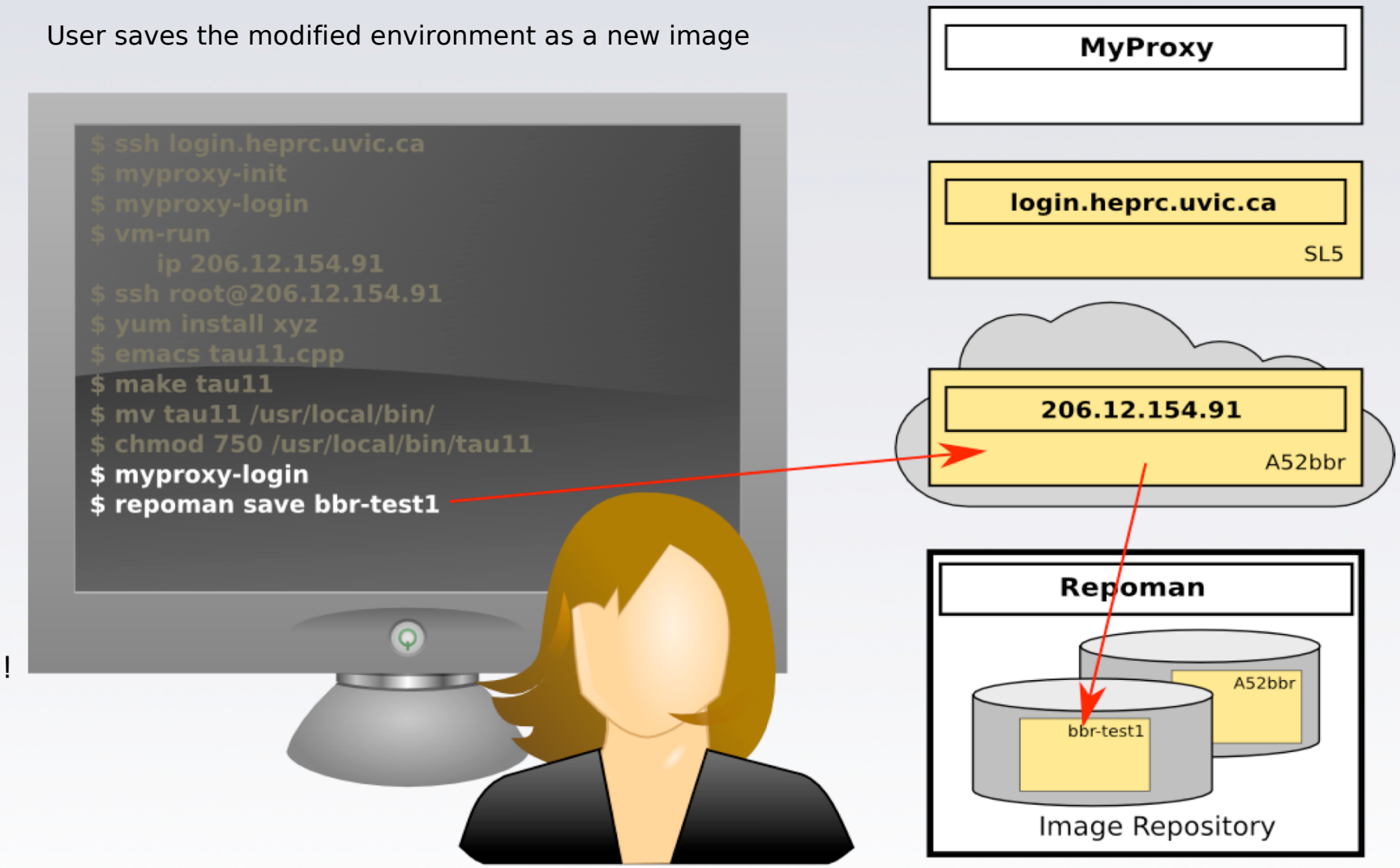
Certificates issued by Grid Canada

**X509** is an **ITU Telecommunication** standard for a public key infrastructure (PKI) for single sign-on

We use it to X509 certificates for user job management, access to Repoman and access to the data storage

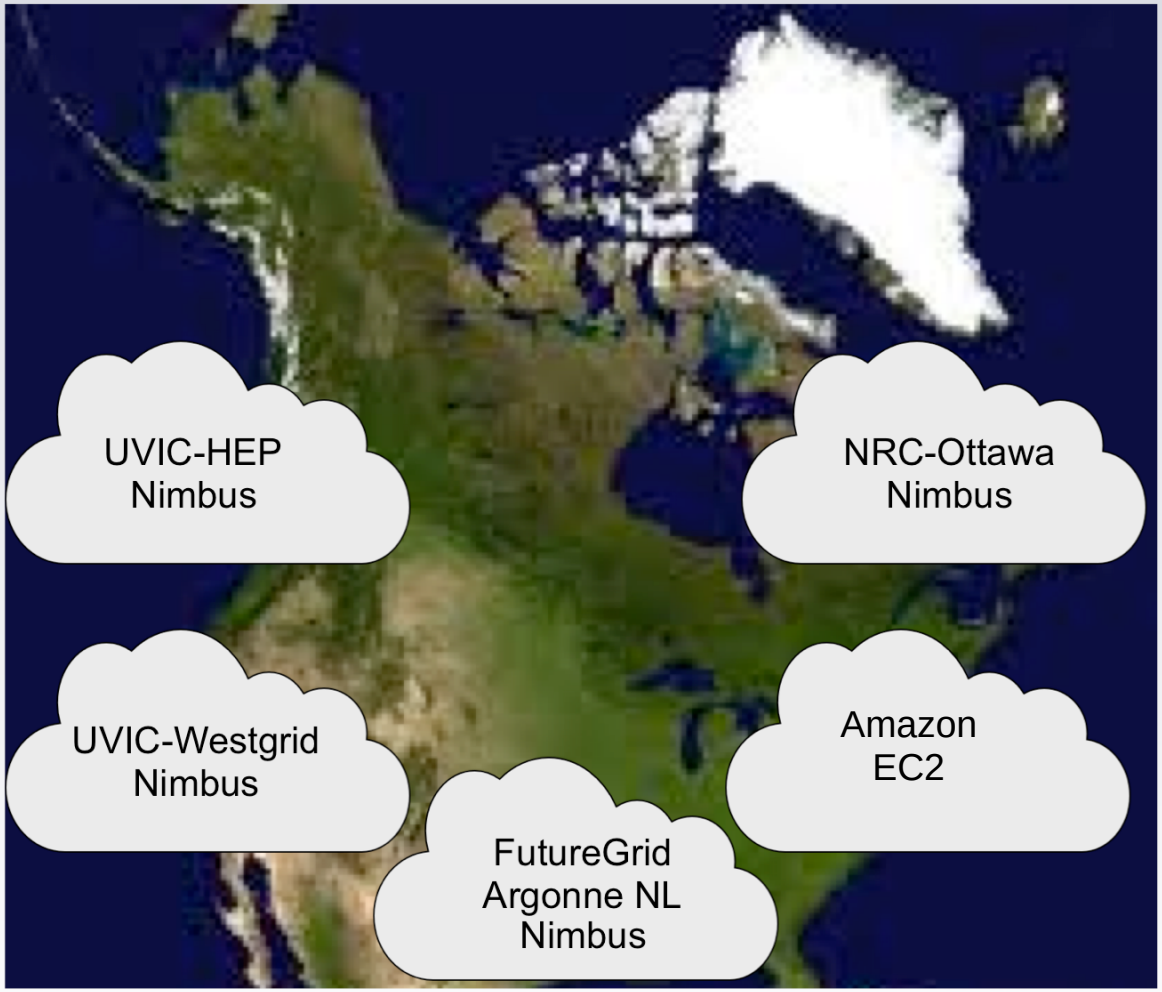
# The Interactive System

User saves the modified environment as a new image




!

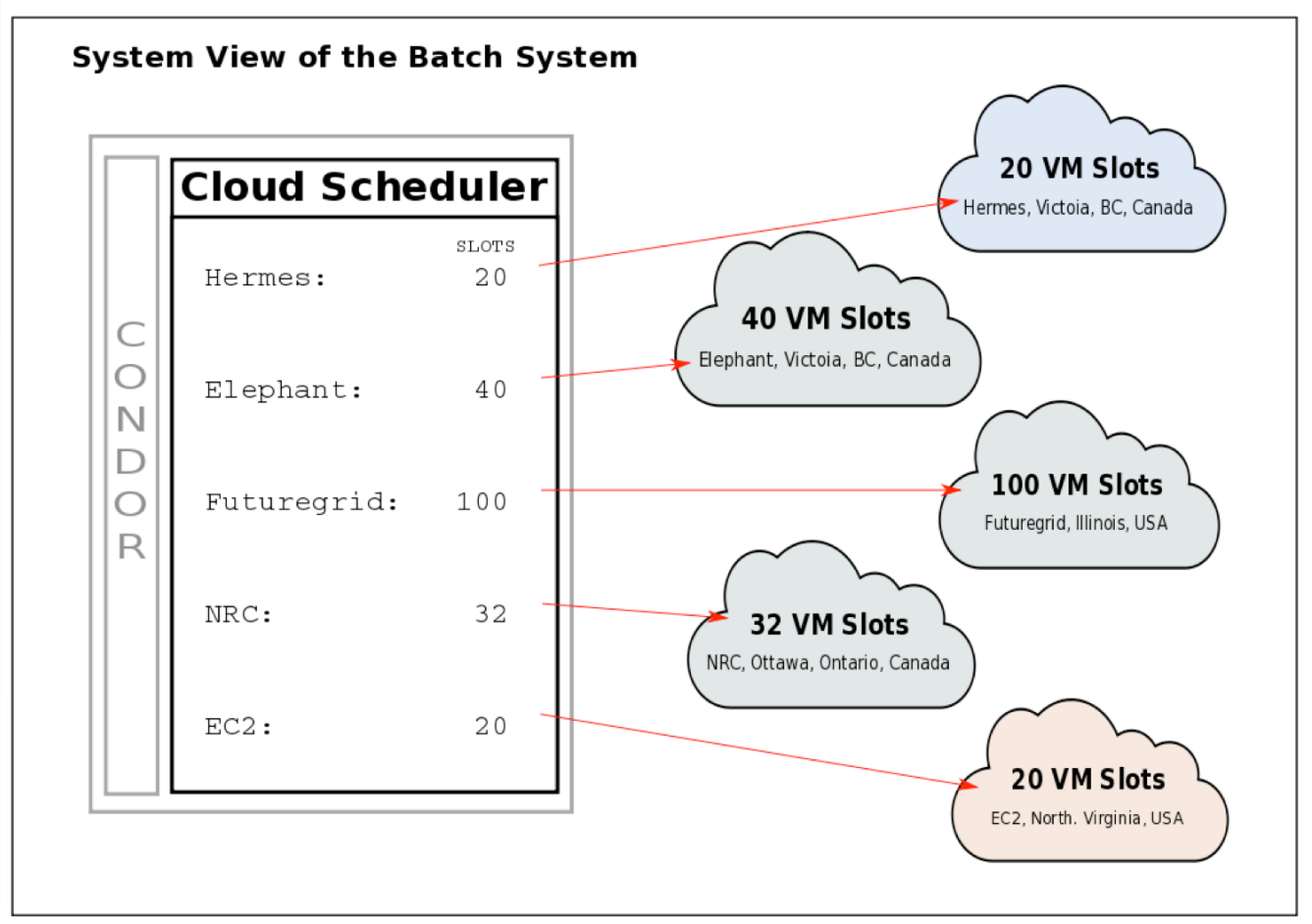
# IaaS cloud resources




Access to few-100s cores on science clouds in Canada and the United States

All sites (except EC2) on the research network

Integrating other FutureGrid sites in the distributed cloud

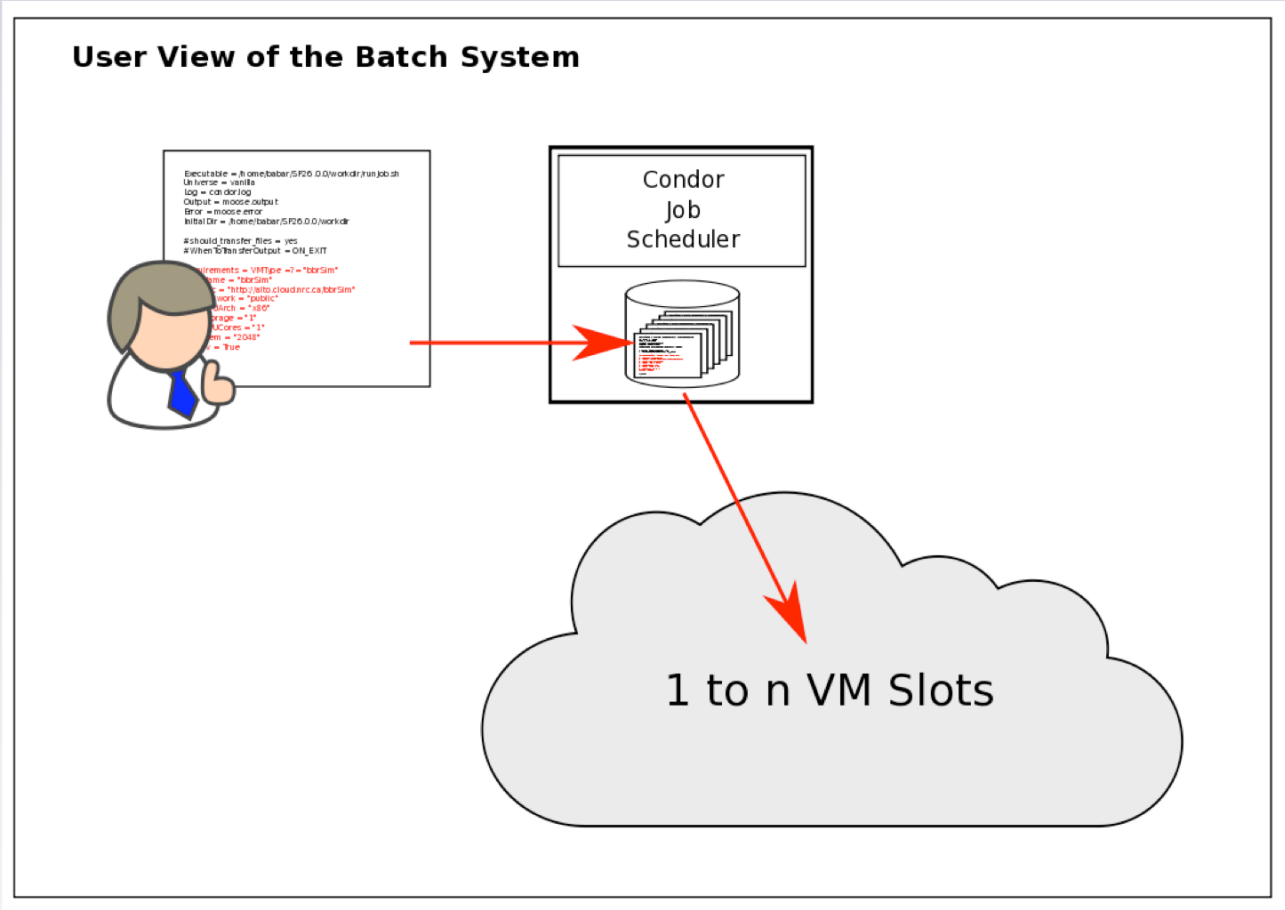



Access to 8 clouds
(Nimbus, Eucalyptus, EC2)

VMs retrieved from two repositories (except EC2) (implementing a VM cache)

Data located at two sites

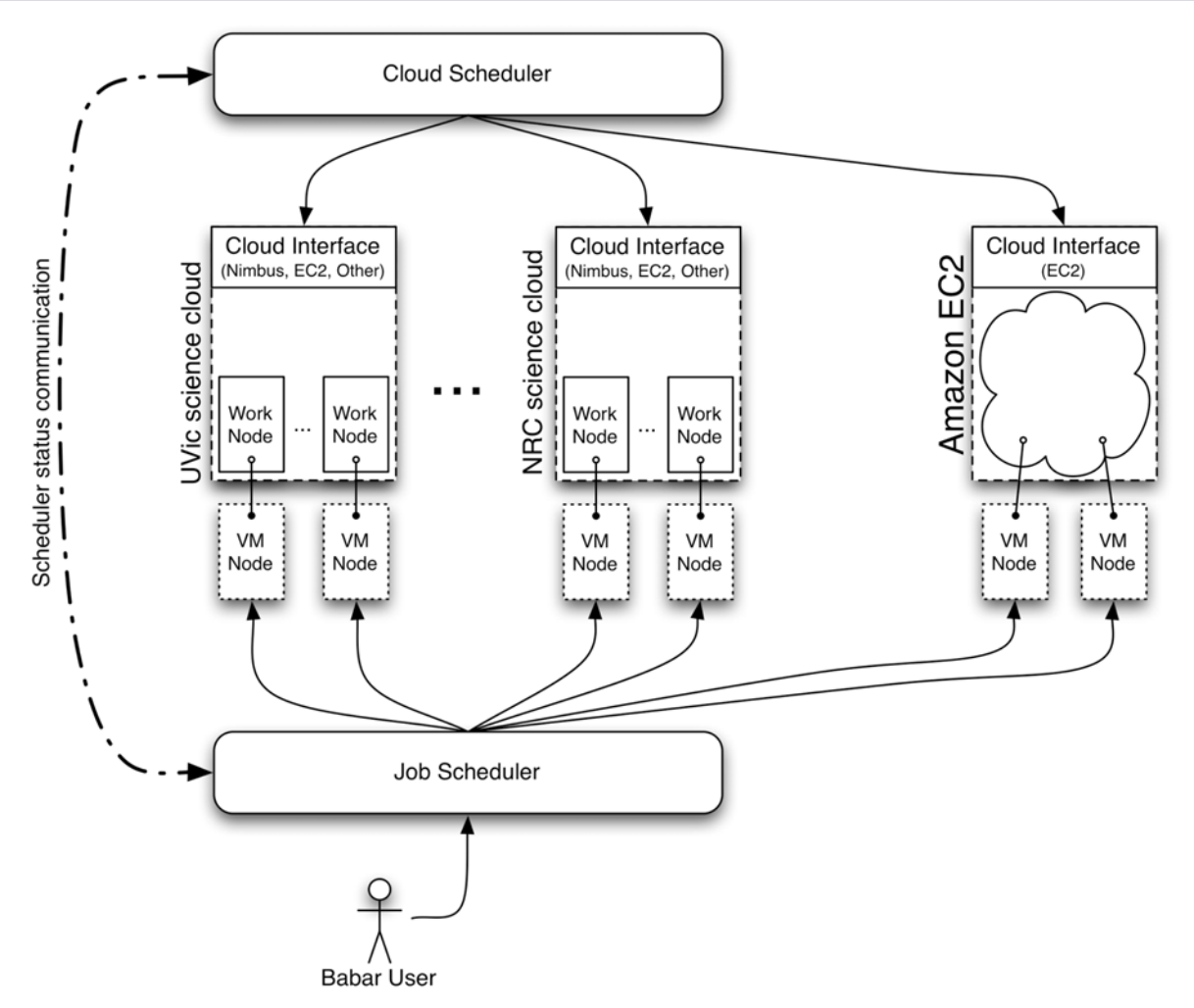
CS looks at the job queue and sends a request to the next available cloud to boot the User-VM




User view of the system is the same as a standard batch environment

Job script contains a link to the user's VM required for the job

# *Cloud Scheduler*




Cloud scheduler looks at the job queue

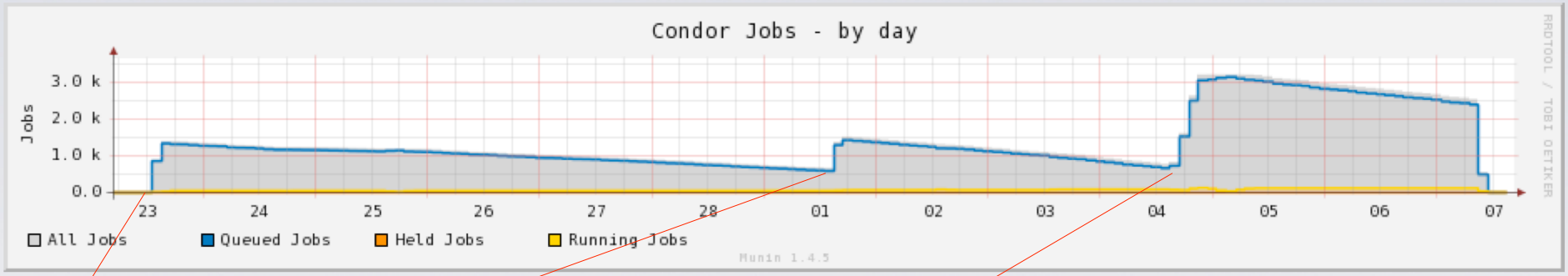
Makes a request to boot a user-VM on a cloud

The cloud retrieves the user-VM from the repository

The user-VM attaches itself to the Condor pool and Condor sends the user-job to the user-VM.

The user-VM stays active if there are more jobs that require it.

# Simulation Production

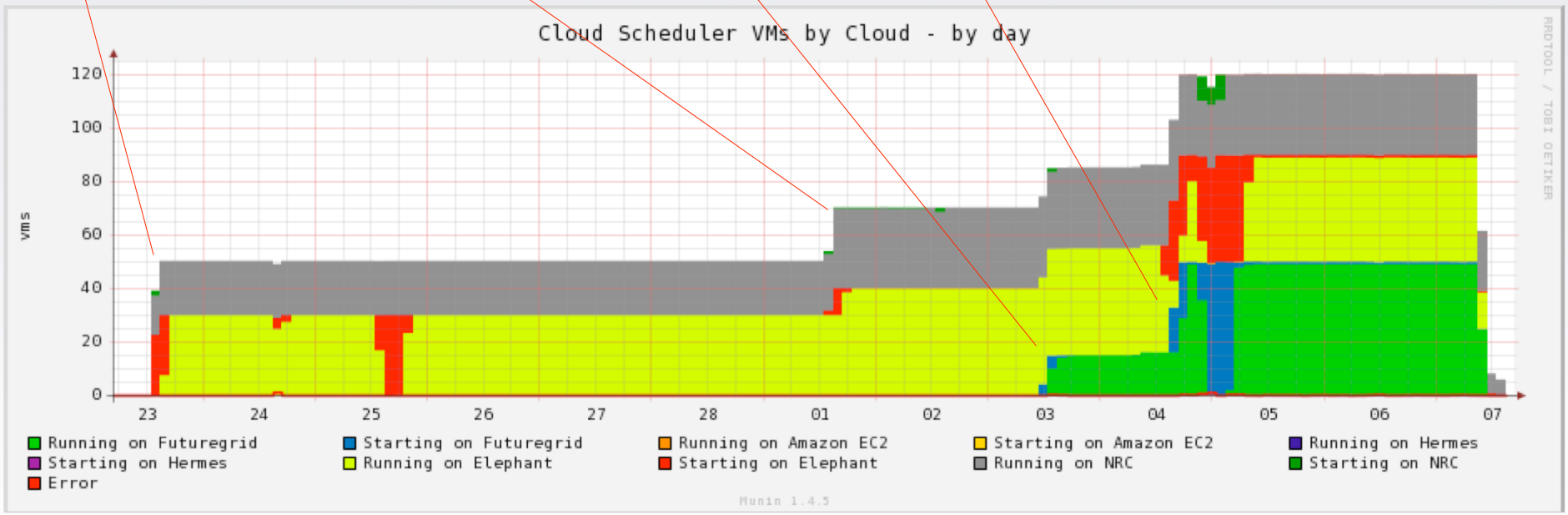



~1300 jobs
50 slots
2 time zones

~850 more jobs, 20 more slots

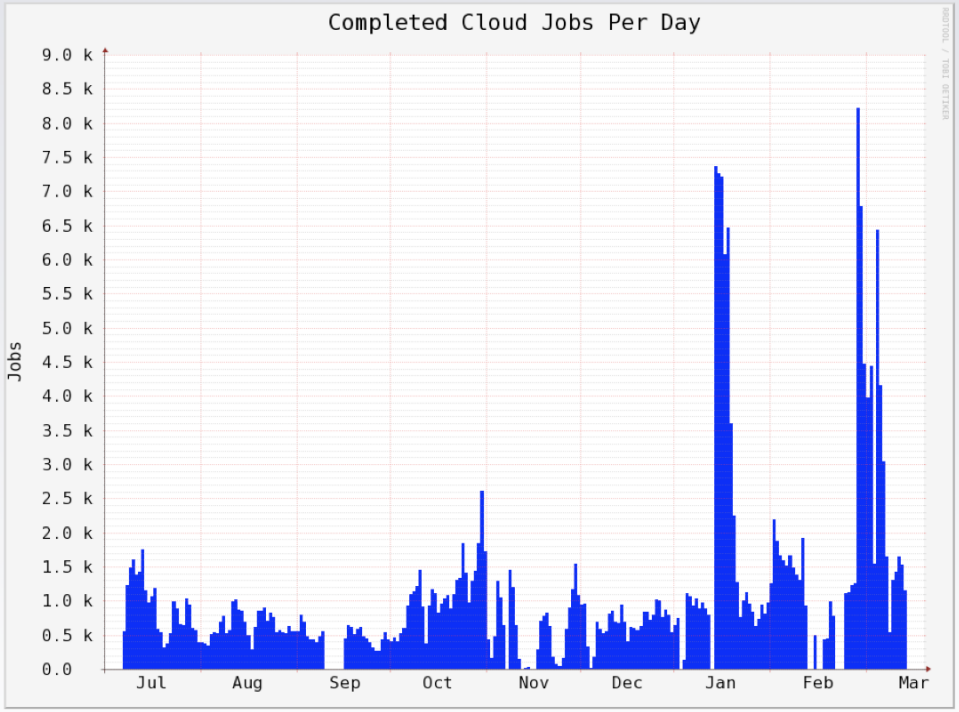
15 more slots, 3rd time zone

2500 more Jobs, 35 more slots



# *Astronomy applications*

**CANFAR Project**
***Canadian Advanced Network for Astronomical Research***
*UIVC, UBC, NRC-HIA*
*CANARIE-funded project*




Distributed cloud used to process survey data

In production for 8 months using different IaaS cloud resources

Compute Canada cloud site at UVIC

Enabling system for user analysis as well as production jobs

See Tuesday presentation

# *Current status*

- Run up to 500 simultaneous jobs over 8 clouds
  - > 100,000 successful jobs
  - Should scale for low-IO applications
- Testing user-analysis (chaotic, high IO)
  - Early test :10TB analysis in 2 days
- Caching of user-VMs (only transfer 1 16G VM)
  - Testing squid-cache
- Network challenges
  - Needed to work with CANARIE to resolve some issues
  - CANARIE will soon connect with Amazon EC2

# *Summary*

- Our distributed cloud focuses on applications in physical sciences with large high-throughput (HTC) workloads and a knowledgeable user community
- Fault-tolerant system using multiple-IaaS (commercial or science) cloud resources
- Based on open-source components with two new in-house elements
  - Cloud scheduler and Repoman (VM repository)
- Production use by astronomers and BaBar
- >100,000 VMs booted

Support provided by
CANARIE, NSERC, NRC, Amazon, Google, FutureGrid (NSF)

# Components & References

- Open Source code developed by University of Victoria:
  - Cloud Scheduler >=0.11.1, https://github.com/hep-gc/cloud-scheduler
  - Repoman, https://github.com/hep-gc/repoman
- Other Open Source components used:
  - Scientific Linux 5.x (Xen, KVM), http://www.scientificlinux.org
  - Nimbus >=2.5, http://www.nimbusproject.org
  - Condor >=7.4, http://www.cs.wisc.edu/condor
  - MyProxy, http://grid.ncsa.illinois.edu/myproxy
  - Xrootd, http://xrootd.slac.stanford.edu
  - Lustre >=1.8.3, http://wiki.lustre.org/index.php/Main_Page
  - Squid 2.7.STABLE8, http://www.squid-cache.org
  - Munin 1.4.5 (epel repository), http://munin-monitoring.org