HEP cloud production using the CloudScheduler/HTCondor Architecture

Ian Gable
University of Victoria
on behalf of many people and groups

CHEP 2015 Okinawa, 2015-04-15

Outline

Overview of the Architecture CloudScheduler/HTCondor

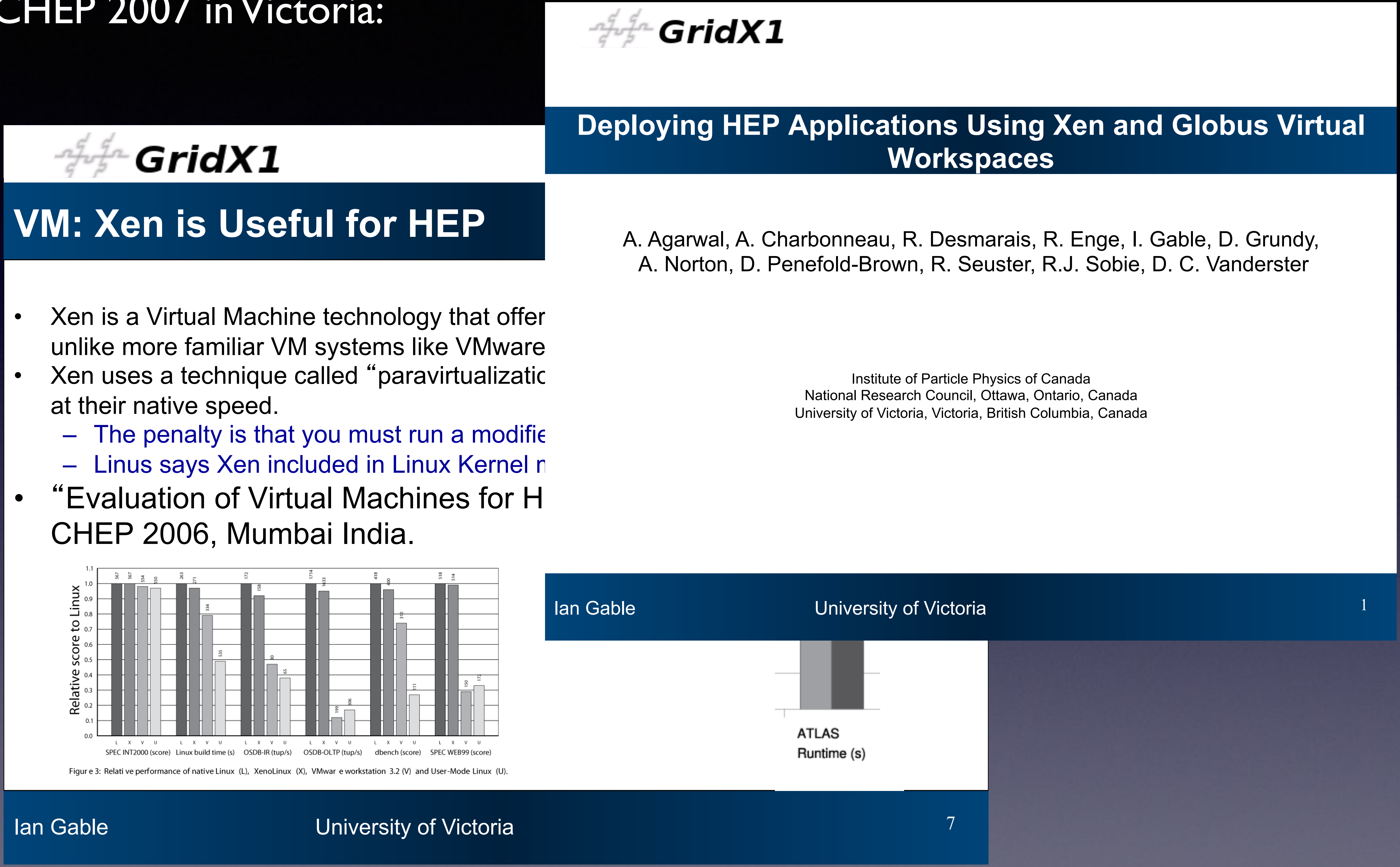
Key enabling technologies:

- CVMFS + µCernVM 3
- Shoal (web cache discovery)
- Glint (OpenStack plugin)

Current cloud resources, and pointers to other talks and posters.

Historical Perspective

CHEP 2007 in Victoria:



GridX1

Deploying HEP Applications Using Xen and Globus Virtual Workspaces

A. Agarwal, A. Charbonneau, R. Desmarais, R. Enge, I. Gable, D. Grundy, A. Norton, D. Penefold-Brown, R. Seuster, R.J. Sobie, D. C. Vanderster

Institute of Particle Physics of Canada
National Research Council, Ottawa, Ontario, Canada
University of Victoria, Victoria, British Columbia, Canada

Ian Gable University of Victoria 1

GridX1

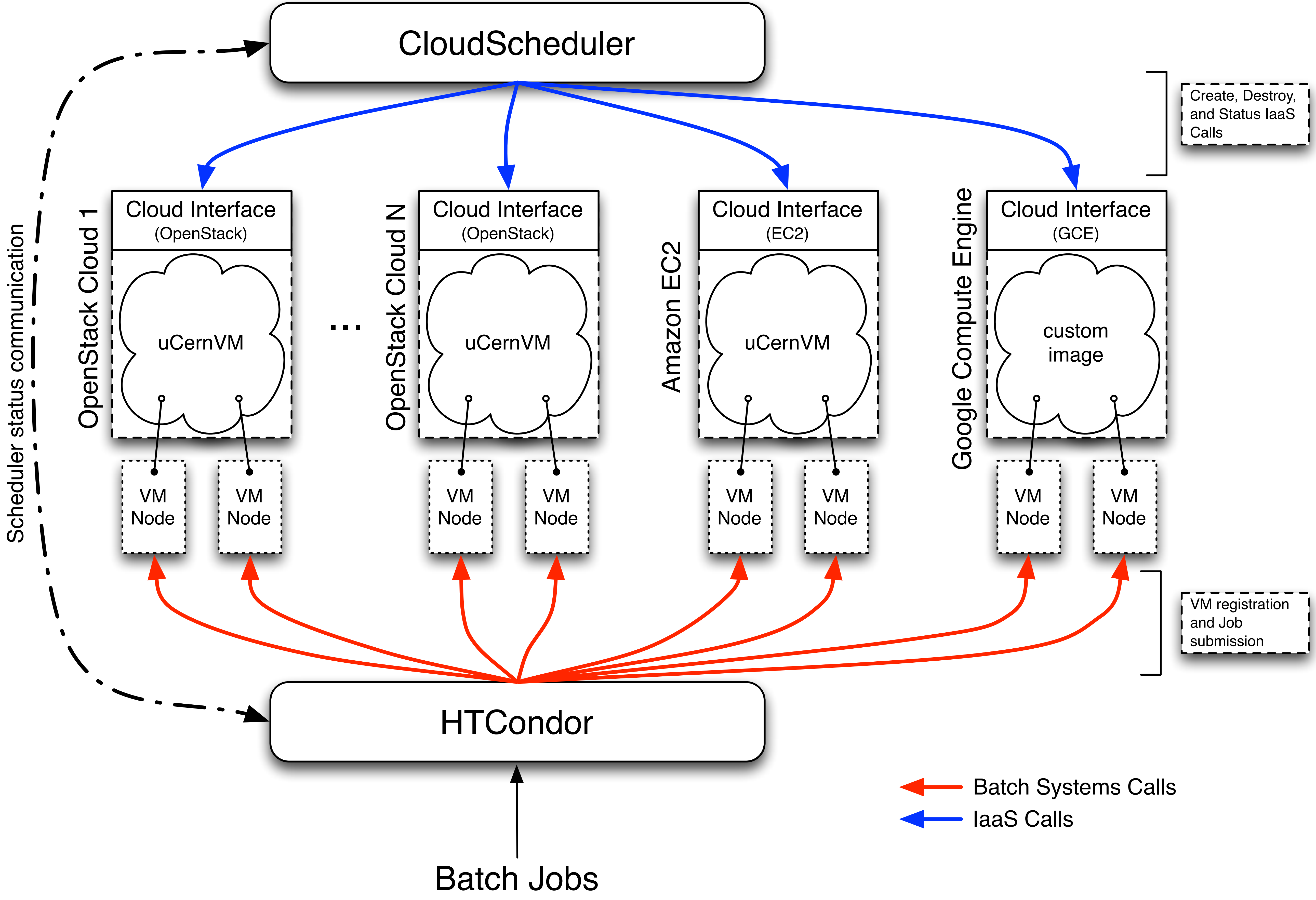
VM: Xen is Useful for HEP

- Xen is a Virtual Machine technology that offer unlike more familiar VM systems like VMware
- Xen uses a technique called "paravirtualizatic at their native speed.
  - The penalty is that you must run a modifie
  - Linus says Xen included in Linux Kernel n
- "Evaluation of Virtual Machines for H CHEP 2006, Mumbai India.

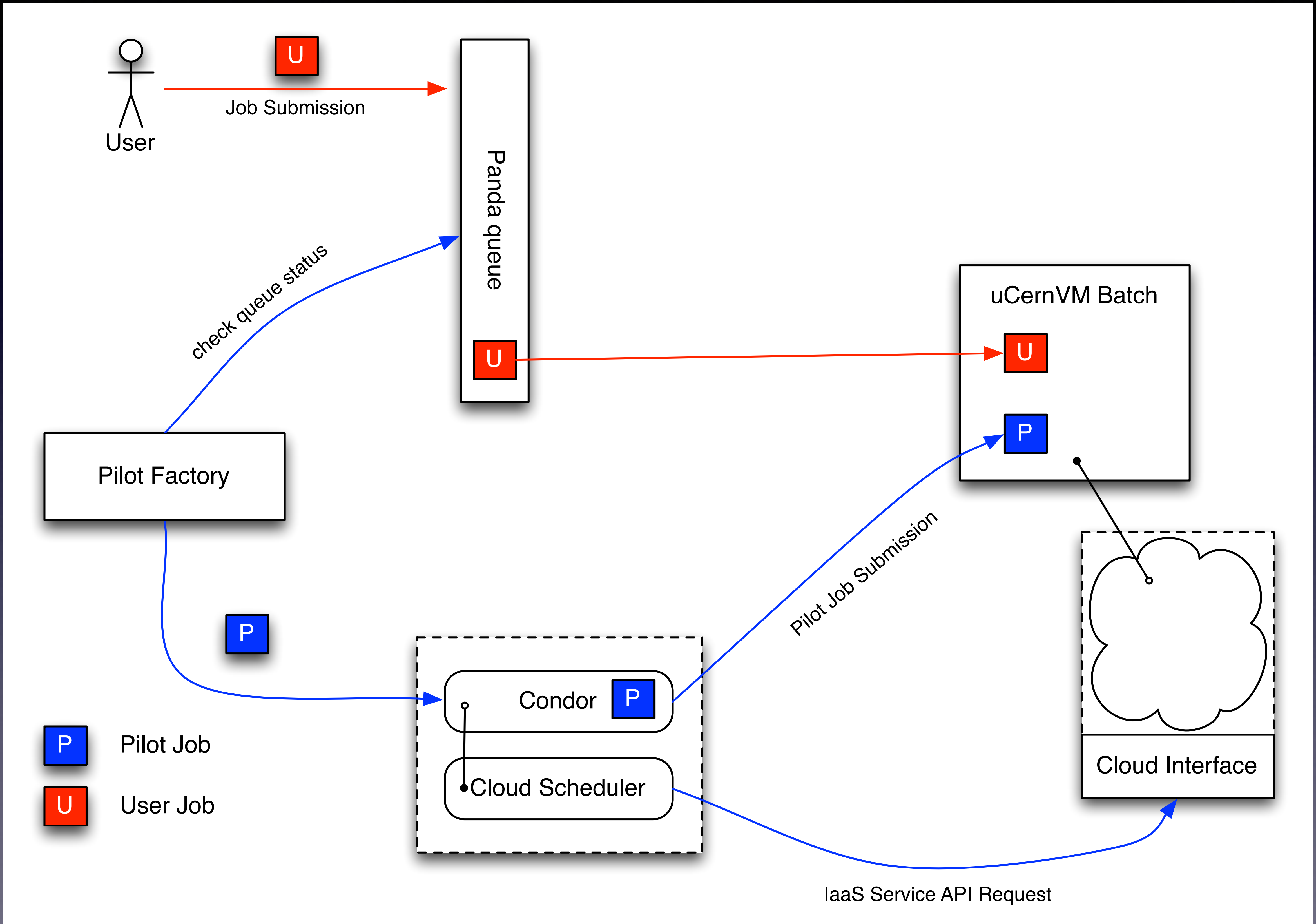
Figur e 3: Relati ve performance of native Linux (L), XenoLinux (X), VMwar e workstation 3.2 (V) and User-Mode Linux (U).

Ian Gable University of Victoria 7

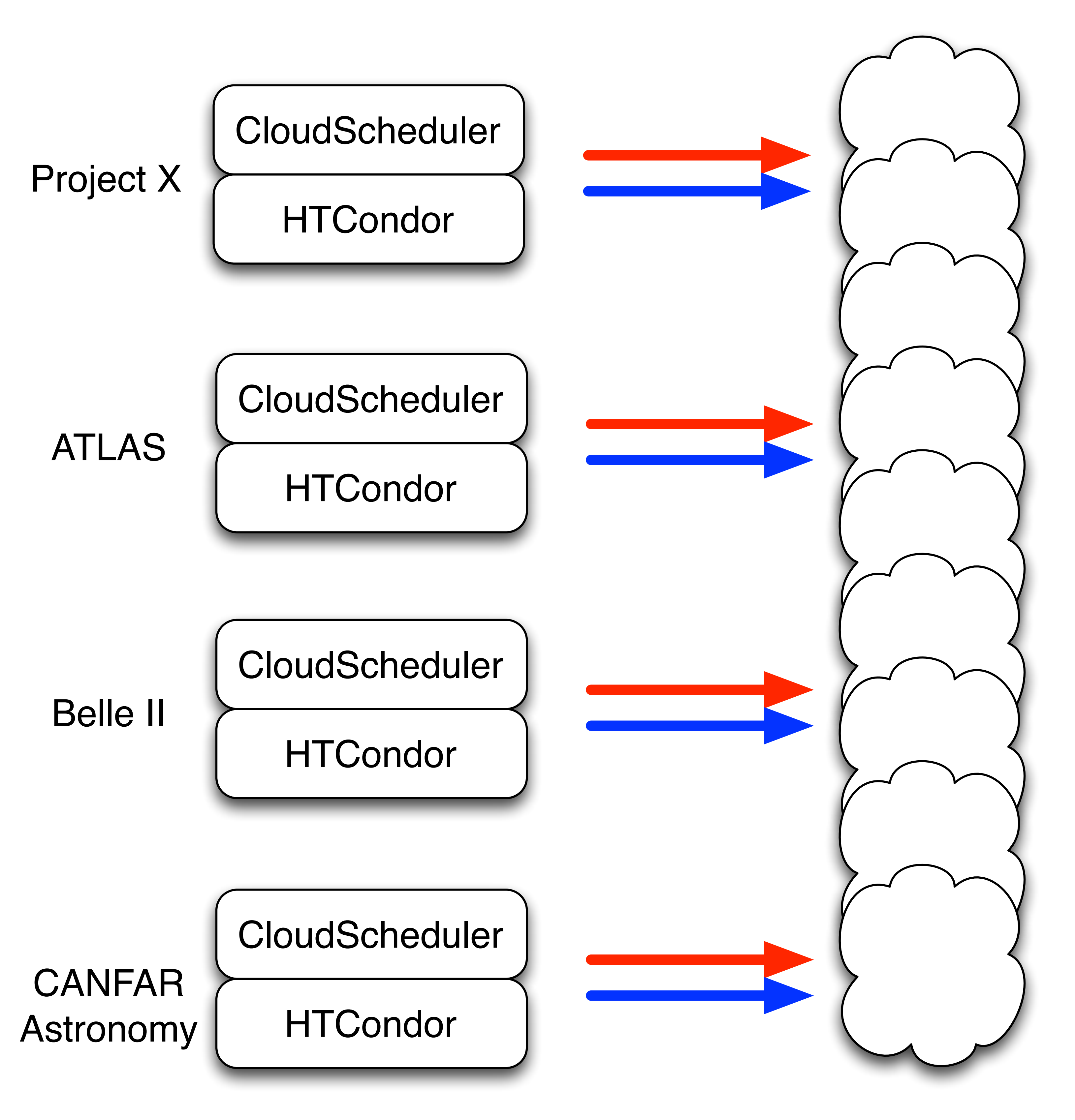
We had already started down the path to IaaS


CloudScheduler
Create, Destroy, and Status IaaS Calls
Scheduler status communication
OpenStack Cloud 1
Cloud Interface (OpenStack)
uCernVM
VM Node
VM Node
...
OpenStack Cloud N
Cloud Interface (OpenStack)
uCernVM
VM Node
VM Node
Amazon EC2
Cloud Interface (EC2)
uCernVM
VM Node
VM Node
Google Compute Engine
Cloud Interface (GCE)
custom image
VM Node
VM Node
VM registration and Job submission
HTCondor
Batch Jobs
Batch Systems Calls
IaaS Calls

Integrating with Pilot Frameworks



Using multiple instances



Support System: CVMFS + µCernVM 3

Without CernVM, completely unmanageable number of VMs image types.

All the well known CVMFS advantages

µCernVM brings the image size down to 20 MB, trivial to deploy everywhere

Complete operating system is staged in, as well as experiment software at Run time.

Requires good squid caching

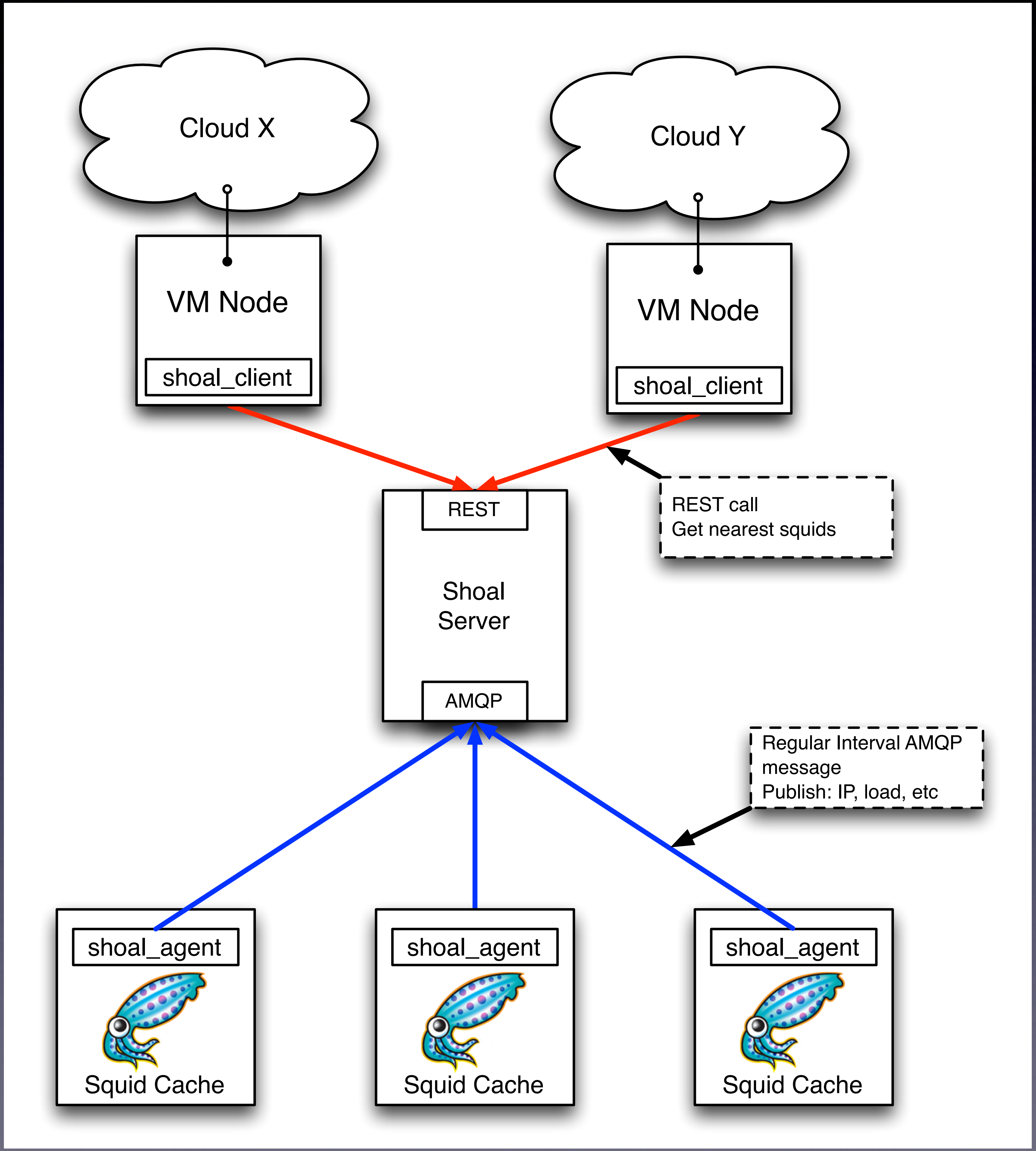
The caching challenge on IaaS cloud

When booting VMs on arbitrary clouds they don't know which squid they should use

In order to work well VMs need to able to access a local web cache (squid) to be able to efficiently download all the experiment software and now OS libraries they need to run

If a VM is statically configured to access a particular cache it can be slow (Geneva → Vancouver for example) and it can overloaded

Shoal: web cache publishing



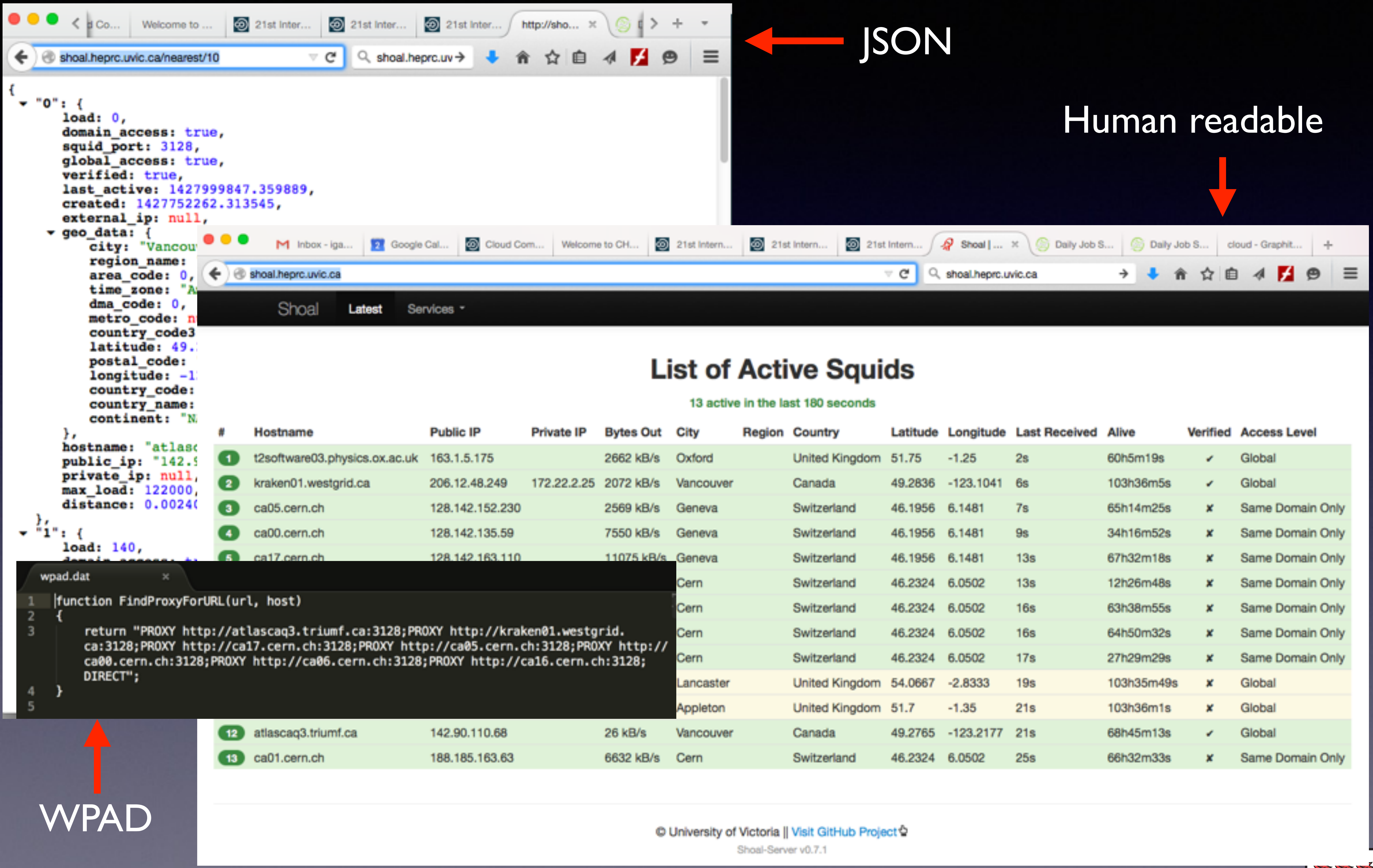
use the highly Scalable AMQP protocol to advertise Squid servers to shoal

use GeoIP information to determine which is the closest to each VM

Squids advertise every 30 seconds, server ‘verifies’ if the squid is functional

https://github.com/hep-gc/shoal

Available Shoal Interfaces



```
{
▾ "0": {
      load: 0,
      domain_access: true,
      squid_port: 3128,
      global_access: true,
      verified: true,
      last_active: 1427999847.359889,
      created: 1427752262.313545,
      external_ip: null,
    ▾ geo_data: {
          city: "Vancou
          region_name:
          area_code: 0,
          time_zone: "A
          dma_code: 0,
          metro_code: n
          country_code3
          latitude: 49.
          postal_code:
          longitude: -1
          country_code:
          country_name:
          continent: "N
      },
      hostname: "atlasc
      public_ip: "142.9
      private_ip: null,
      max_load: 122000,
      distance: 0.00240
  },
▾ "1": {
      load: 140,
```

List of Active Squids

13 active in the last 180 seconds

#	Hostname	Public IP	Private IP	Bytes Out	City	Region	Country	Latitude	Longitude	Last Received	Alive	Verified	Access Level
1	t2software03.physics.ox.ac.uk	163.1.5.175		2662 kB/s	Oxford		United Kingdom	51.75	-1.25	2s	60h5m19s	✓	Global
2	kraken01.westgrid.ca	206.12.48.249	172.22.2.25	2072 kB/s	Vancouver		Canada	49.2836	-123.1041	6s	103h36m5s	✓	Global
3	ca05.cern.ch	128.142.152.230		2569 kB/s	Geneva		Switzerland	46.1956	6.1481	7s	65h14m25s	✘	Same Domain Only
4	ca00.cern.ch	128.142.135.59		7550 kB/s	Geneva		Switzerland	46.1956	6.1481	9s	34h16m52s	✘	Same Domain Only
5	ca17.cern.ch	128.142.163.110		11075 kB/s	Geneva		Switzerland	46.1956	6.1481	13s	67h32m18s	✘	Same Domain Only
					Cern		Switzerland	46.2324	6.0502	13s	12h26m48s	✘	Same Domain Only
					Cern		Switzerland	46.2324	6.0502	16s	63h38m55s	✘	Same Domain Only
					Cern		Switzerland	46.2324	6.0502	16s	64h50m32s	✘	Same Domain Only
					Cern		Switzerland	46.2324	6.0502	17s	27h29m29s	✘	Same Domain Only
					Lancaster		United Kingdom	54.0667	-2.8333	19s	103h35m49s	✘	Global
					Appleton		United Kingdom	51.7	-1.35	21s	103h36m1s	✘	Global
12	atlascaq3.triumf.ca	142.90.110.68		26 kB/s	Vancouver		Canada	49.2765	-123.2177	21s	68h45m13s	✓	Global
13	ca01.cern.ch	188.185.163.63		6632 kB/s	Cern		Switzerland	46.2324	6.0502	25s	66h32m33s	✘	Same Domain Only

© University of Victoria || Visit GitHub Project

Shoal-Server v0.7.1

wpad.dat

```
function FindProxyForURL(url, host)
{
    return "PROXY http://atlascaq3.triumf.ca:3128;PROXY http://kraken01.westgrid.
    ca:3128;PROXY http://ca17.cern.ch:3128;PROXY http://ca05.cern.ch:3128;PROXY http://
    ca00.cern.ch:3128;PROXY http://ca06.cern.ch:3128;PROXY http://ca16.cern.ch:3128;
    DIRECT";
}
```

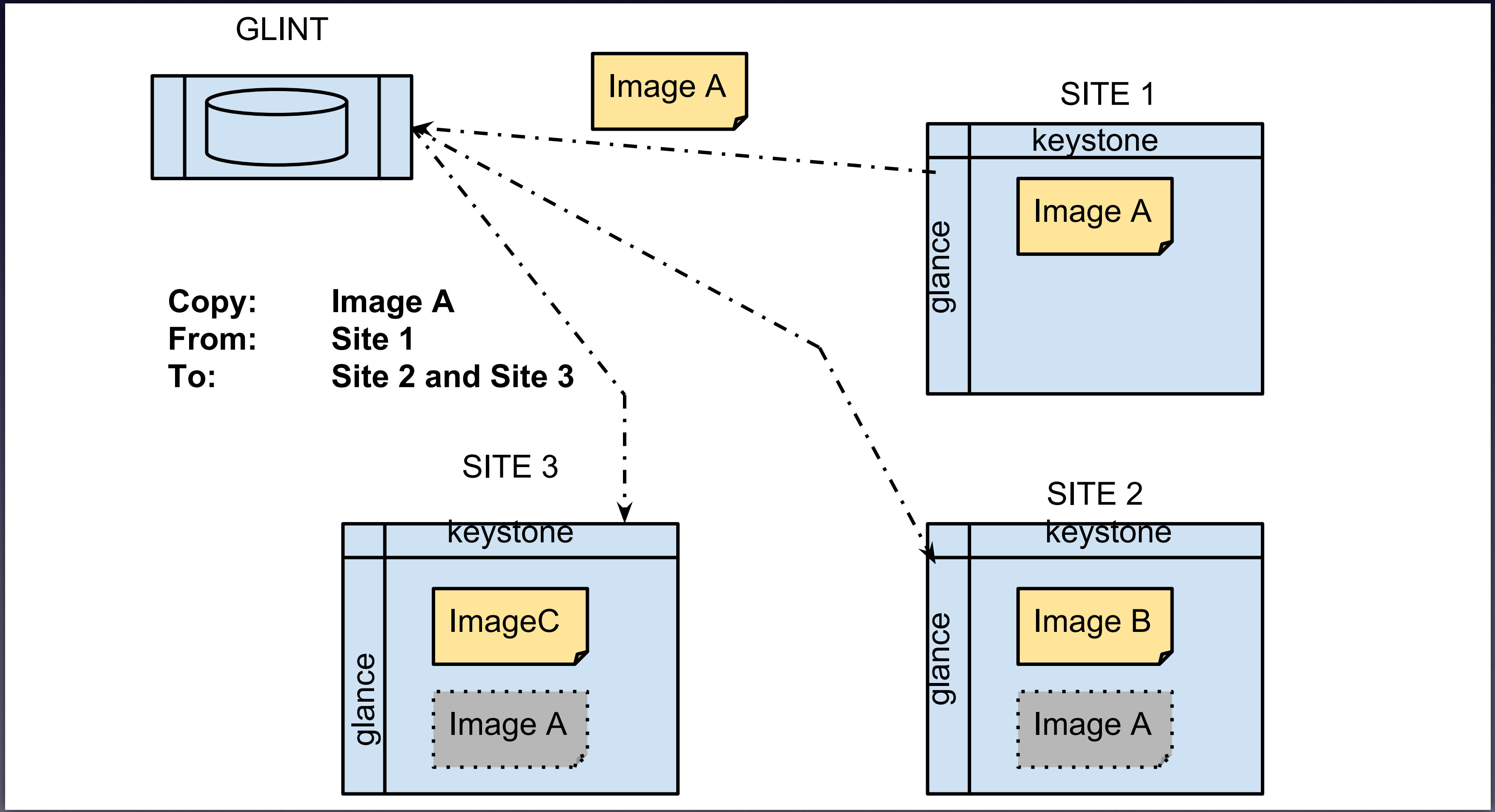
Glint: OpenStack image distribution

Distributing VM images by hand to 5 clouds is manageable but error prone

Distributing VM images by hand to 20+ clouds never happens right the first time and is hugely time consuming

Glint is the solution to this problem for OpenStack

Glint automates the deployment of images to multiple clouds



https://github.com/hep-gc/glint

Clouds in use

Currently Operating

Site	HEP Cloud?	ATLAS	BelleII
ComputeCanada-West	no	✔	✔
ComputeCanada-East	no	✔	✔
CERN	yes	✔	
GridPP-Datacentered	yes	✔	
CANARIE-West	no	✔	
CANARIE-East	no	✔	
Amazon EC2	no	✔	
ChameleonCloud-U Texas	no		✔
Google Compute Engine	no		✔

~ 4000 cores operating today

For comparison, Canadian T2+T1 running ~5000 today

Cloud can come and go with time.

Past clouds include:

NECTAR Australia,
National Research Council Ottawa,
FutureGrid - U Chicago,
Elephant Coud UVic

Summary

CloudScheduler/HTCondor flexible multi experiment way to run Batch Jobs on Clouds.

Key enabling technologies for this:

CVMFS + µCernVM 3

Shoal: dynamic Squid cache Publishing

Glint: VM Image Distribution

Current users ATLAS, Belle II, CANFAR, Compute Canada HPC consortium

"Evolution of Cloud Computing in ATLAS" Ryan Taylor, 17:00, this room

"Utilizing cloud computing resources for Belle II", Randall Sobie, this room

"Glint: VM image distribution in a multi-cloud environment" Poster Session B