



Experience using non conventional Grid resources

Federico Stagni, on behalf of the LHCb
distributed computing team

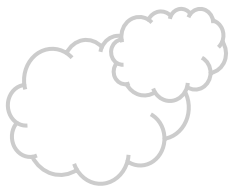


Credits

(in alphabetic order)

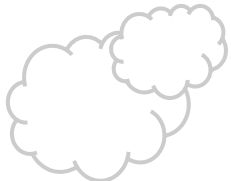
- Baptiste Cabarro
- Philippe Charpentier
- Joel Closier
- Víctor Mendez
- Andrew McNab
- Stefan Roiser
- Nathalie Rauschmayr
- Mario Úbeda García (from whom I stole many slides...)

LCHb Resources



what we have

- **LCG sites**
 - CREAM CE: direct pilot jobs submission
 - Integrating an ARC CE
- **One, large, DIRAC site**
 - (and a new, smaller one)
 - Submission of pilot jobs to torque, via SSH
- **HLT farm**
 - not virtualized, and no pilot jobs: pull model
- **Cloud, VAC, BOINC “sites”**
 - Running contextualized VMs



WLCG

...not for today



DIRAC sites

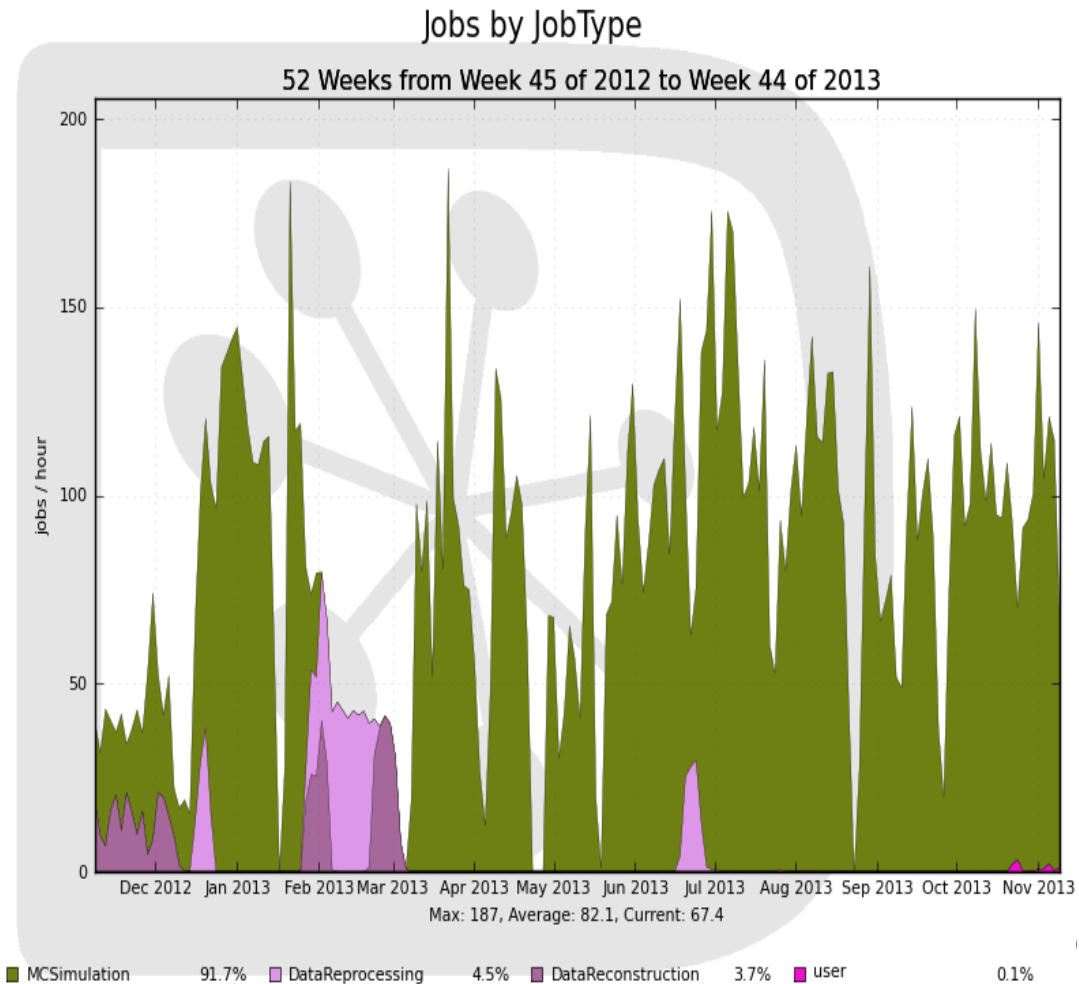
Using DIRAC SSH “CE”,
can have a batch system
behind.

Similar to BOSCO

Sending pilots directly,
same as for CREAM CEs

One large site with torque, stable
in production for 2+ years

Added a second lately



HLT farm

~1500 PCs and ~16000 cores

NOT virtualized

(no, for our case it was not needed)

Presented in CHEP 2012: <http://cds.cern.ch/record/1460607?ln=en>

but, with CVMFS!

Initialization via the **PVSS** Control Manager:

- connects/disconnects monitoring of the worker nodes
- manages startup of the dirac agents
- balances the load of the allocated farm
- monitors the connection with the internal PVSS services

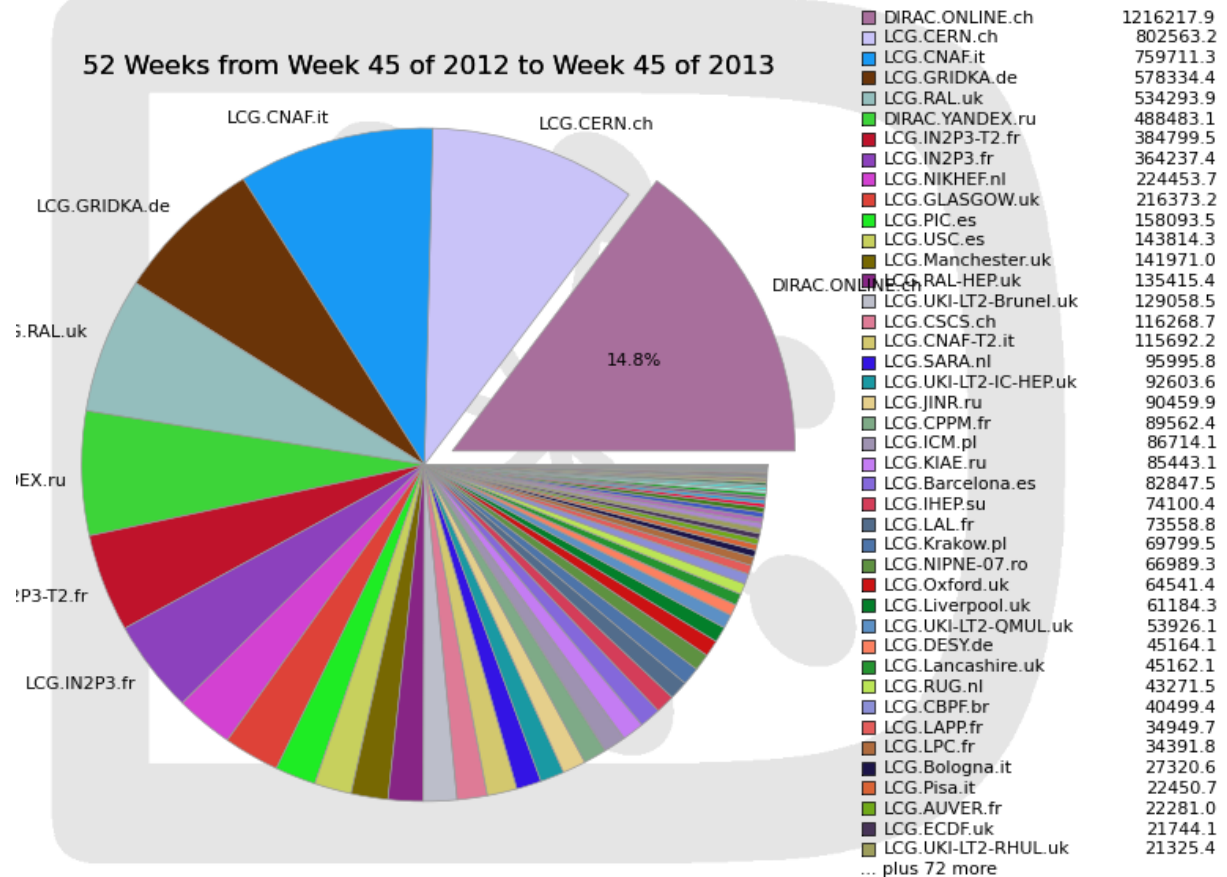
HLT farm: results

Stable in production since the beginning of 2013

top site since then

CPU days used by Site

52 Weeks from Week 45 of 2012 to Week 45 of 2013



Clouds

- OpenStack cluster @ CERN ("Grizzly")

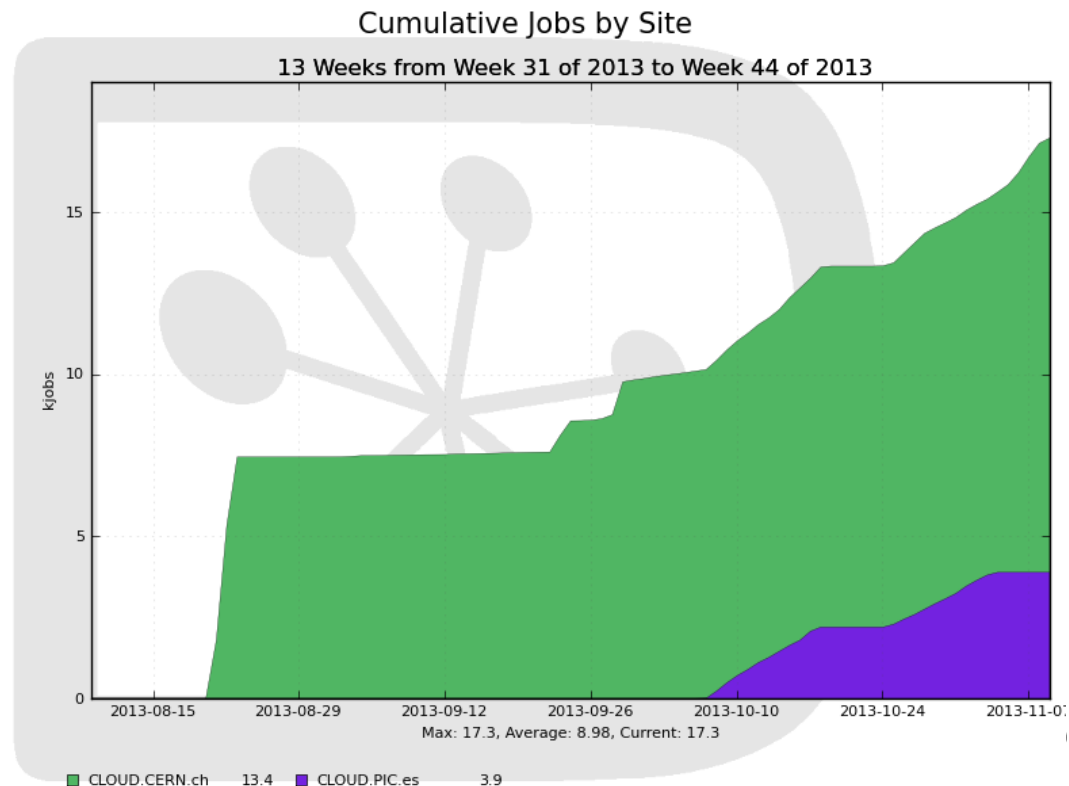
- 600 VCPU
 - CLOUD.CERN.ch
 - CLOUD.CERNMP.ch

- OpenNebula cluster @ PIC

- 75 VCPU
 - CLOUD.PIC.es

Soon (?) to be:

- StratusLab@RAL
- OpenStack@Imperial
- MeghaCloud (CESGA)



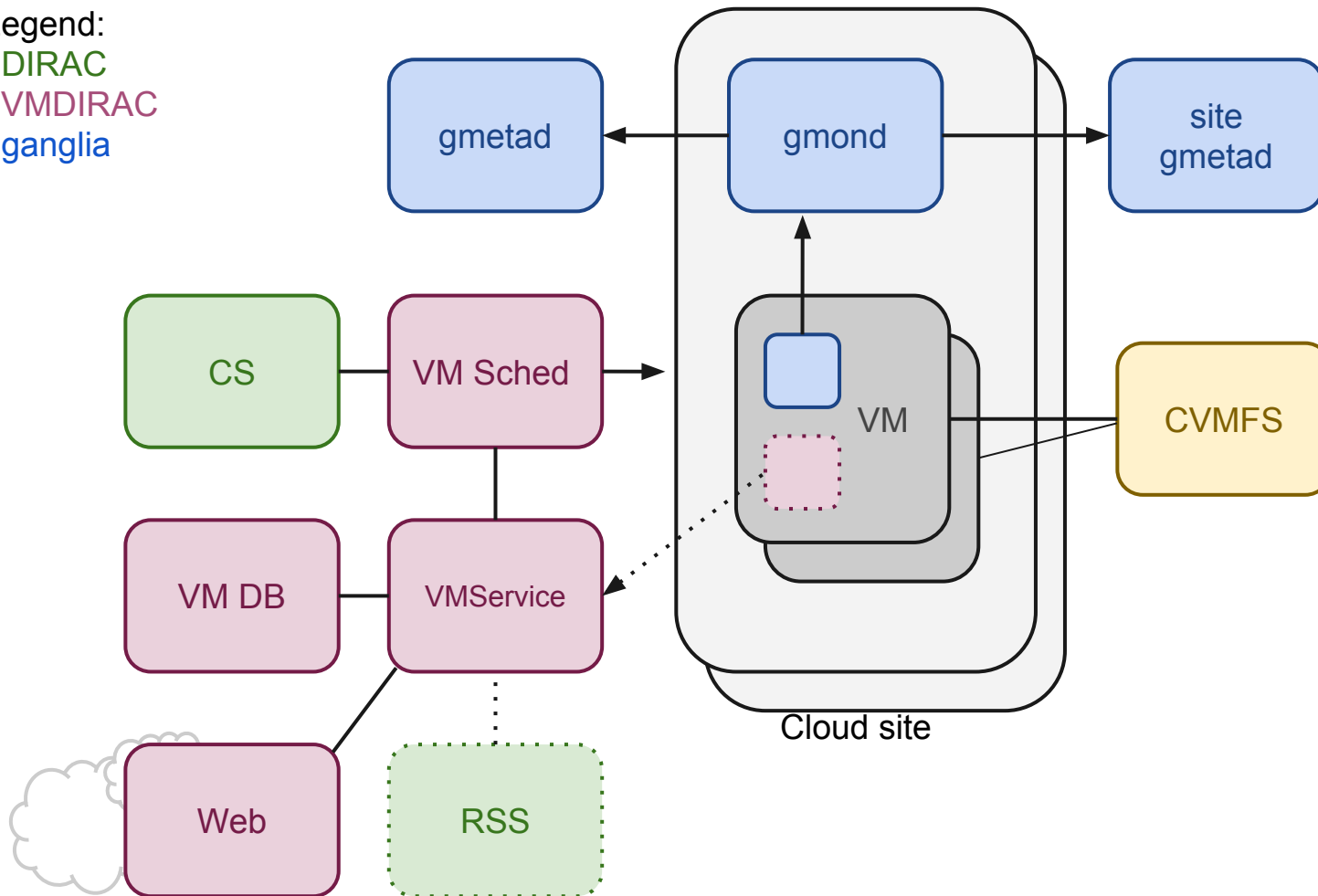
Architecture

Legend:

DIRAC

VMDIRAC

ganglia

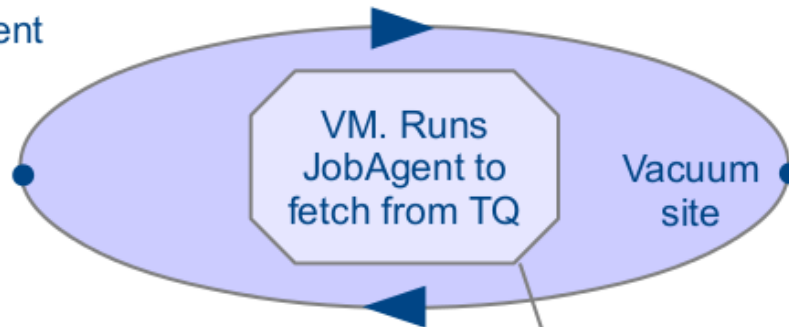


“The Vacuum”

Infrastructure-as-a-Client
(IaaS)

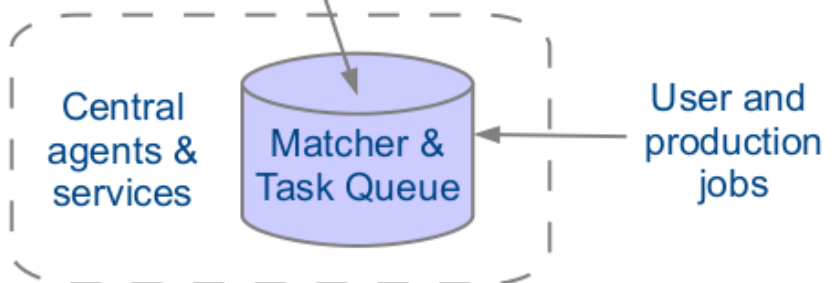
Instead of being created by VOs, the Virtual Machines appear spontaneously “out of the vacuum” at sites.

As with the other models, the JobAgent runs and requests real jobs from the Matcher and normal Task Queue.



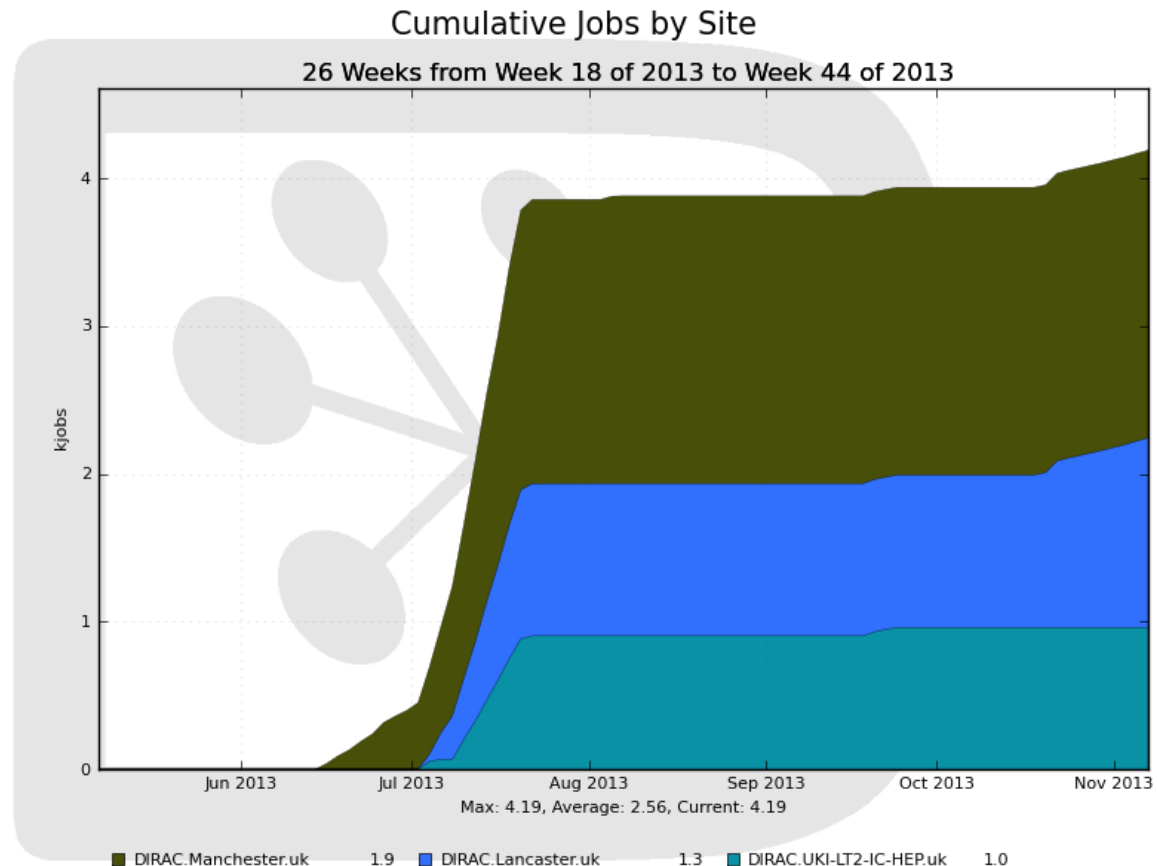
Hypervisors/hosts can run VMs for particular VOs depending on work available and target shares for each VO.

Requests
for real jobs



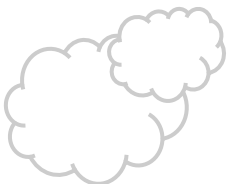
VAC: results

- Operates virtual machines at a site for a specific set of experiments, but using a considerably simpler software stack than at Cloud or Grid sites.
- Uses HEPiX MACHINEFEATURES for graceful termination of the VMs
- Tried during summer campaign



BOINC

- Server up and running: <http://lhcbathome.cern.ch/Beauty/>
 - The BOINC application is the VM itself
 - Got good suggestions from Test4Theory group
 - Configuration isn't easy, need to experiment more
 - Out of the box integration within LHCbDIRAC
 - The VM is low priority process and can be paused by the user or BOINC
- It works, but... with many “buts”...



So, we can run jobs

Great, but:

- what's the cost of it?
- what's NOT working?
- what will NOT work soon?
- what should we do as a community, and not as a VO?

Focus on clouds

Clouds: general considerations

- Cloud softwares didn't grow with institutions in mind
- Rather, seems they grew with credit card in mind:
 - You get what you pay
- “We” have access to few public clouds
- But what we need, as a community, are institutional clouds, and target shares
 - This is where I like VAC

Ulrik S, just ~2 hours ago

- Resource allocation and sharing
 - Fixed quota only, like dedicated resources
 - Quota management by the cloud admins. Changes require intervention by them
 - One tenant per sub-group, or even several tenants per service (eg batch) makes things hard to manage
- Resource usage
 - Unused quota risk to stay idle
 - Optimization difficult as idle resources are difficult to avoid ...
 - Lack of queues
 - Incoming requests need to be served immediately, but resources are limited
 - Maybe an economic model or spot market would help for a better usage of resources? Not aware of a concrete ongoing project

Clouds: more general considerations

- Clouds are certainly great for the sites
 - One tool for all their machines
 - Virtualization is flexibility
- For the VOs, we are gaining more flexibility
 - At the cost of a lot of work:
 - Before getting tailored VMs, there are a lot of details to set up
 - we shifted our role from being users to being, in many ways, sysadmins, and contact points

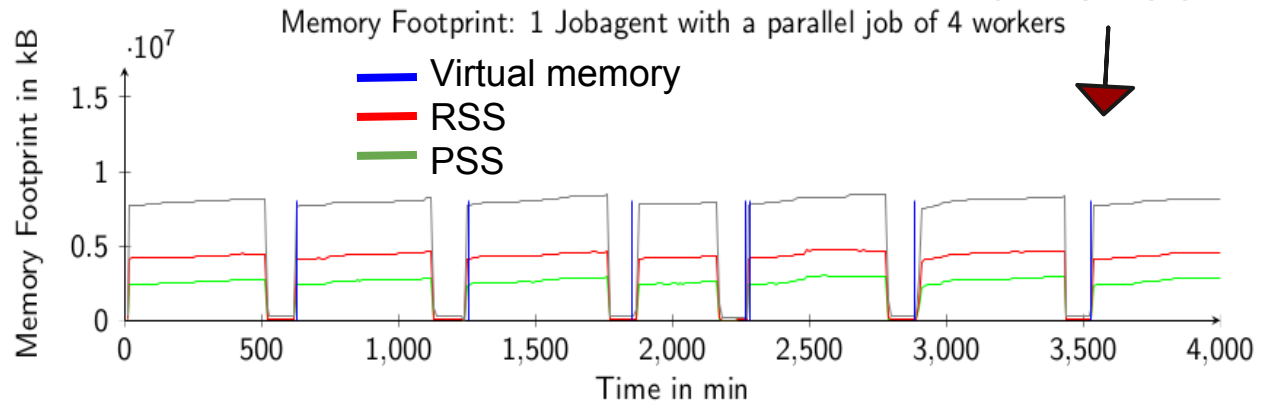
Flexibility in the Cloud: MP

- Different test configuration:
 - 1 JobAgent 4 Workers
 - 2 JobAgents 4 Workers each
 - 4 JobAgent 1 serial job each

GaudiMP

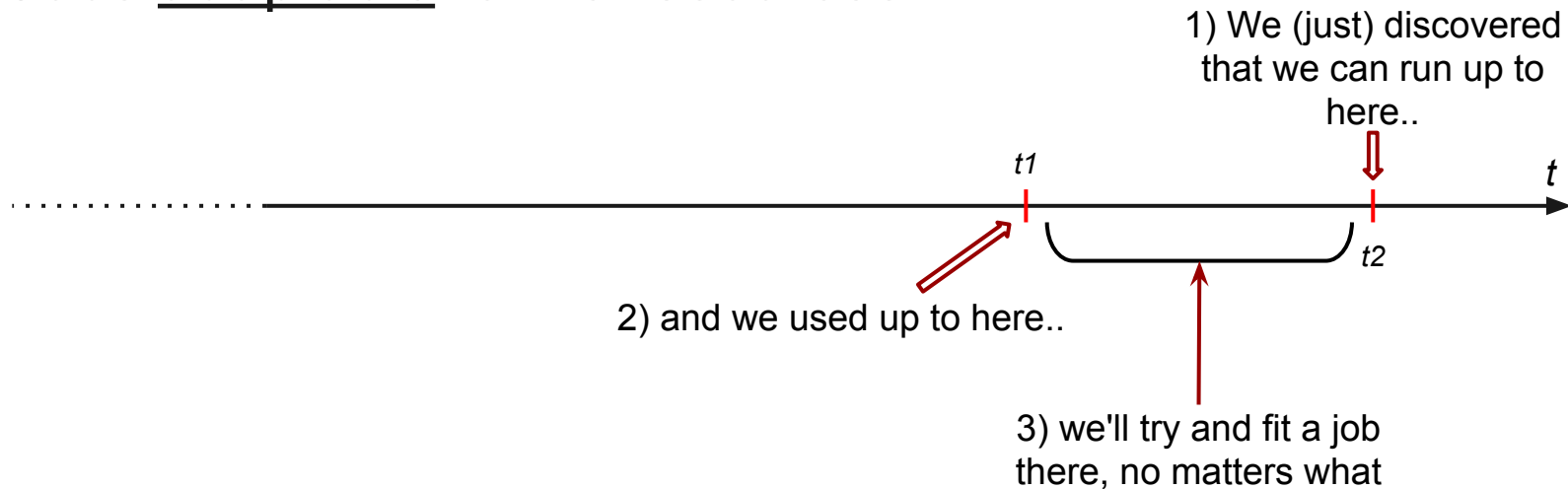
This has an impact
on sites

30%
reduction
in memory
usage



Flexibility (not only) in the Cloud

Jobs adaptable to the resources

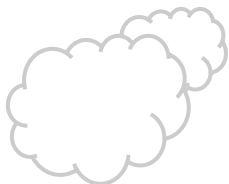


Running on all the "queues"

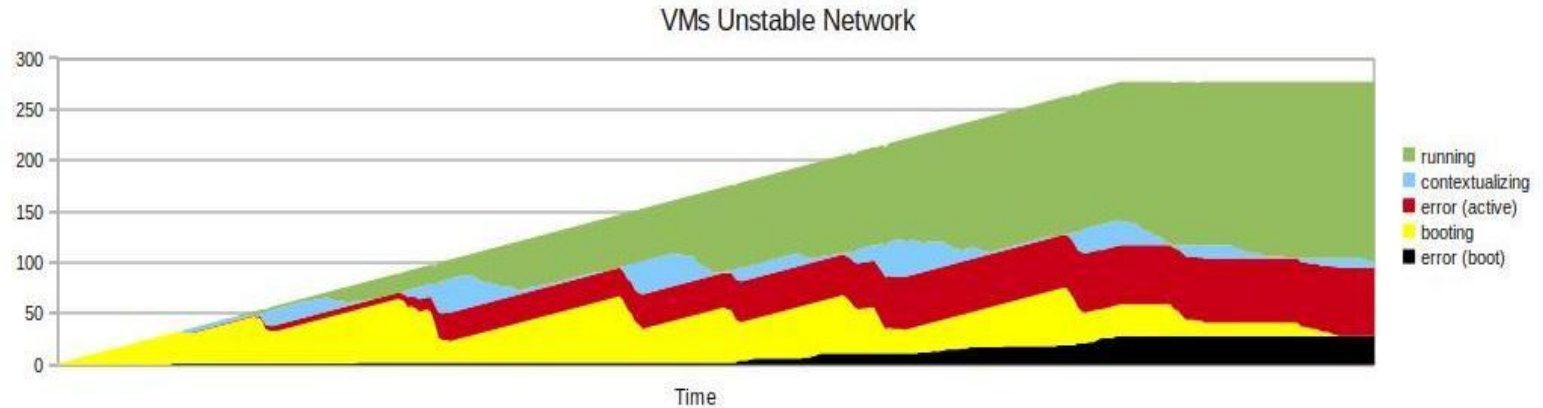
More resources

"short" jobs possible

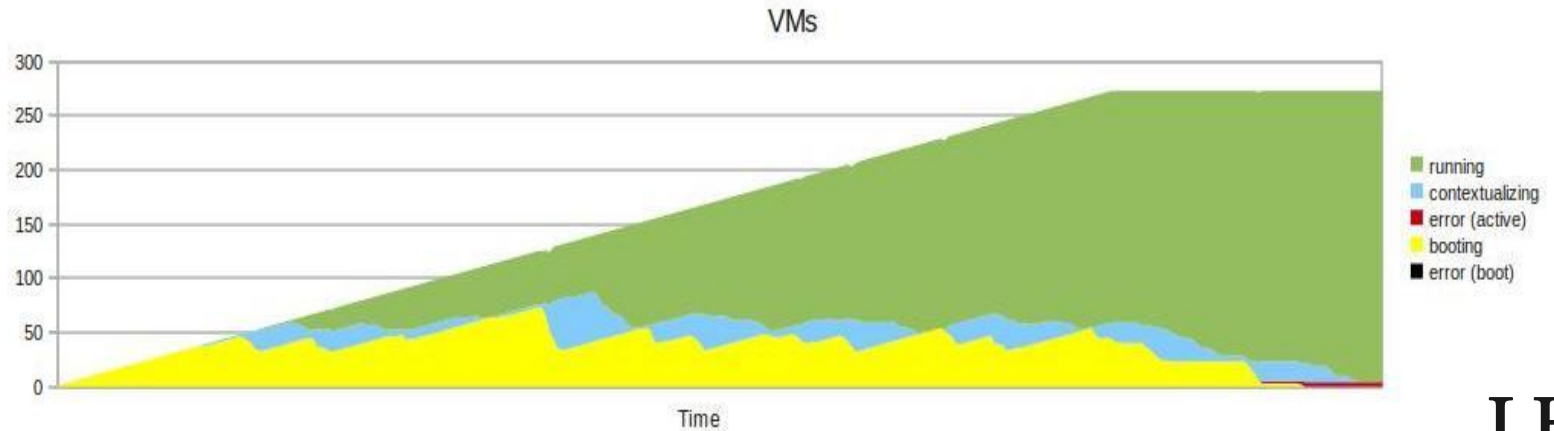
jobs masonry, for single and multicore jobs



Being the sysadmin



I don't recall having to monitor the network before...



A look into Monitoring /1

First of all, why do we need monitoring ?

to monitor the health of a CLOUD Site

← This, we always did

~ DIRAC.RSS

to monitor the health of each and every single VM

← This is new

~ VMDIRAC

Plots at

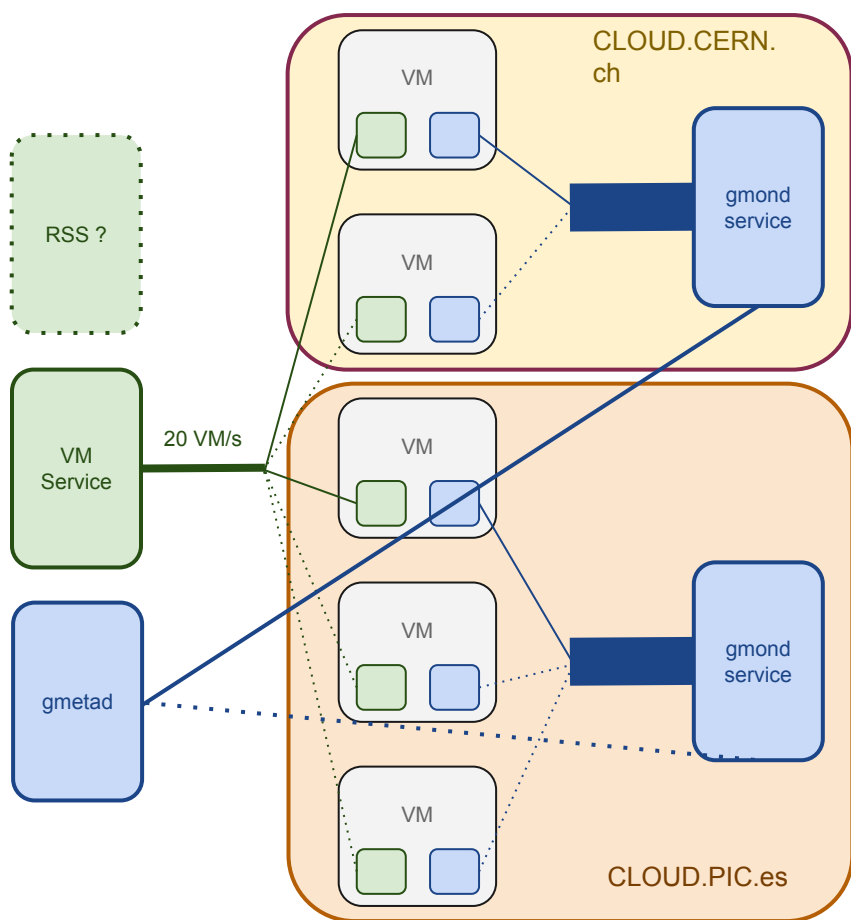
<http://lhcbmonweb.cern.ch/ganglia>

Easy to get swamped by metrics

which ones should be digested ?

boottime
bytes_in
bytes_out
cpu_active
cpu_idle
cpu_num
cpu_report
cpu_speed
cpu_system
cpu_user
cpu_wio
disk_free
disk_total
gexec
ip_address
last_reported
load_fifteen
load_five
load_one
load_report
location
mem_buffers
mem_cached
mem_free
mem_report
mem_total
mem_shared
network_report
os_release
packet_report
part_max_used
pkts_in
pkts_out
proc_run
proc_total
swap_free
swap_total

A look into Monitoring /2



ganglia

no connection with DIRAC,
data can be retrieved using a UDP socket,
rdd graphs,
out of the box on CernVM,
digested data can be read by third parties,
easy to write plugins.

VMMonitorAgent

connects with DIRAC (using TCP),
monitors JobAgent,
mimics some functionalities of XMPP,
google timeline graphs (being exact, it is
the VMDIRAC web view).

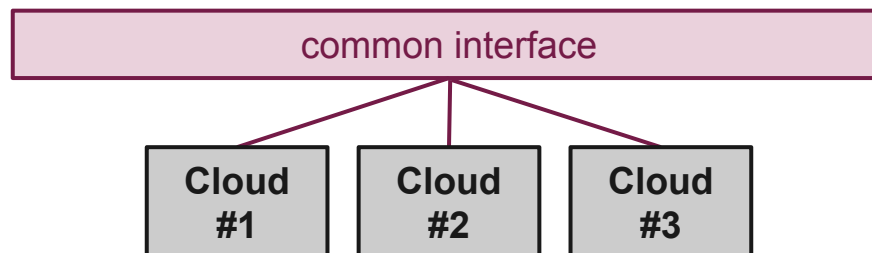
Monitoring roles

What is the role of the VO, and what the role of the provider?

Cloud softwares

OpenStack, OpenNebula, CloudStack,
StratusLab, Eucalyptus, ...

... shouldn't there be some restrictions, WLCG
wise?



Marketplace

- As of today, we upload the same image everywhere
- We dream of a common marketplace



And we hope this is doable

CernVM

Is (u)CernVM:

- the official WLCG image?
- yet another VM?

Miscellanea

- We are using amiconfig, not CloudInit
 - we won't participate in this religion war
 - atm we are not planning to change
- \$MACHINEFEAUTURES and \$JOBFEATURES
 - Great to see that it is moving forward
- It's time to plug in GocDB (and BDII?)
 - we are still at the gentlemen's agreement phase
- HEP06
 - Desirable to know
- AuthN: no username/password, please!
 - @Ulrik S.: yes, I believe we can learn from the Grid



To conclude

- The Grid isn't the Grid any more...
 - New resource types
 - There are not only clouds in the sky
- Virtualization is good but isn't free
 - use with cautions
- Virtualization often means adaptation
- As usual, flexibility is a key word



Comments and questions

?



Backup slides

Security

2 main threats:

VM access

credentials in user data

Security II - access

So far, we have an acceptable solution using PKI, but..

do we really need to leave a door opened on each and every single VM ?

if something goes wrong, we can terminate the VM and boot a new one in DEBUG mode.

pipe JobAgent logs to a messaging queue and store them at the CloudSite.

Security III - user_data

Again, two bullets:

- transport
- storage

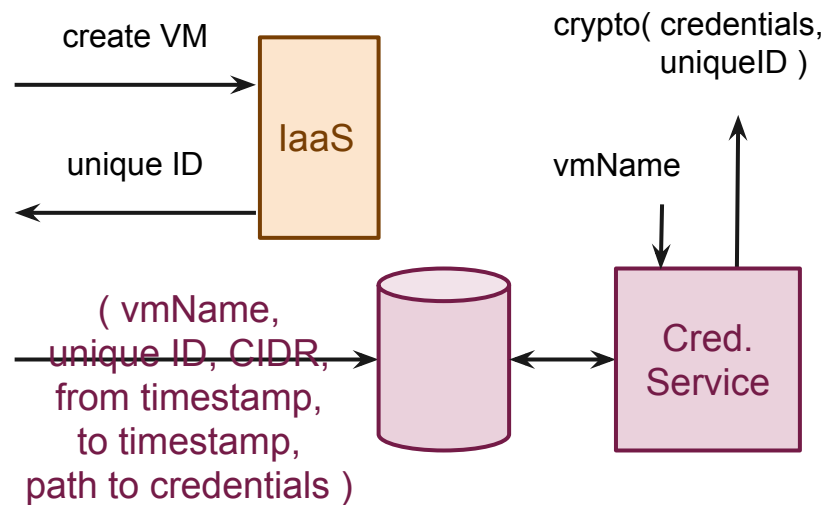
Transport adds TLS layer, so we are good.

However, there is who claims:

- what if someone gains access to the metadata server, he/she could get all our credentials...

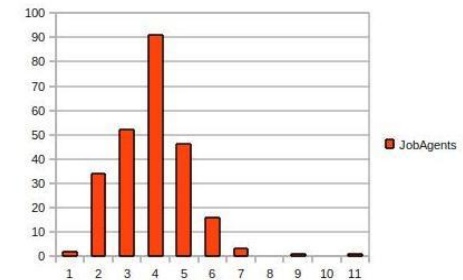
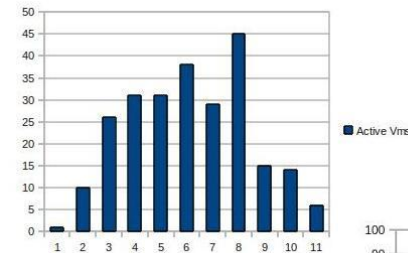
Security IV - user_data

Assuming a reasonable paranoia level, let's solve it.

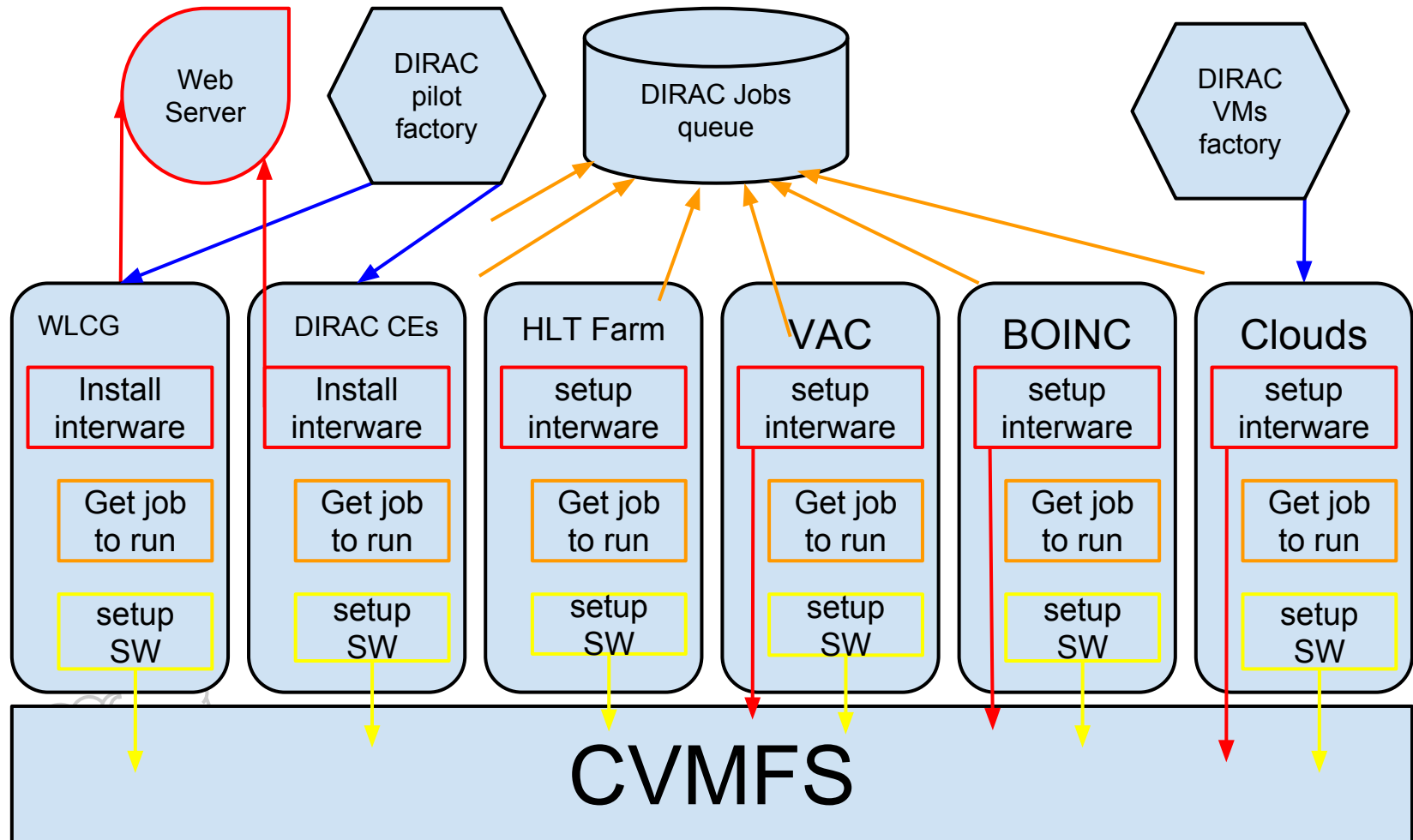


We know CIDR per Cloud Site.
Empirically we know the dispersion of the contextualization times.

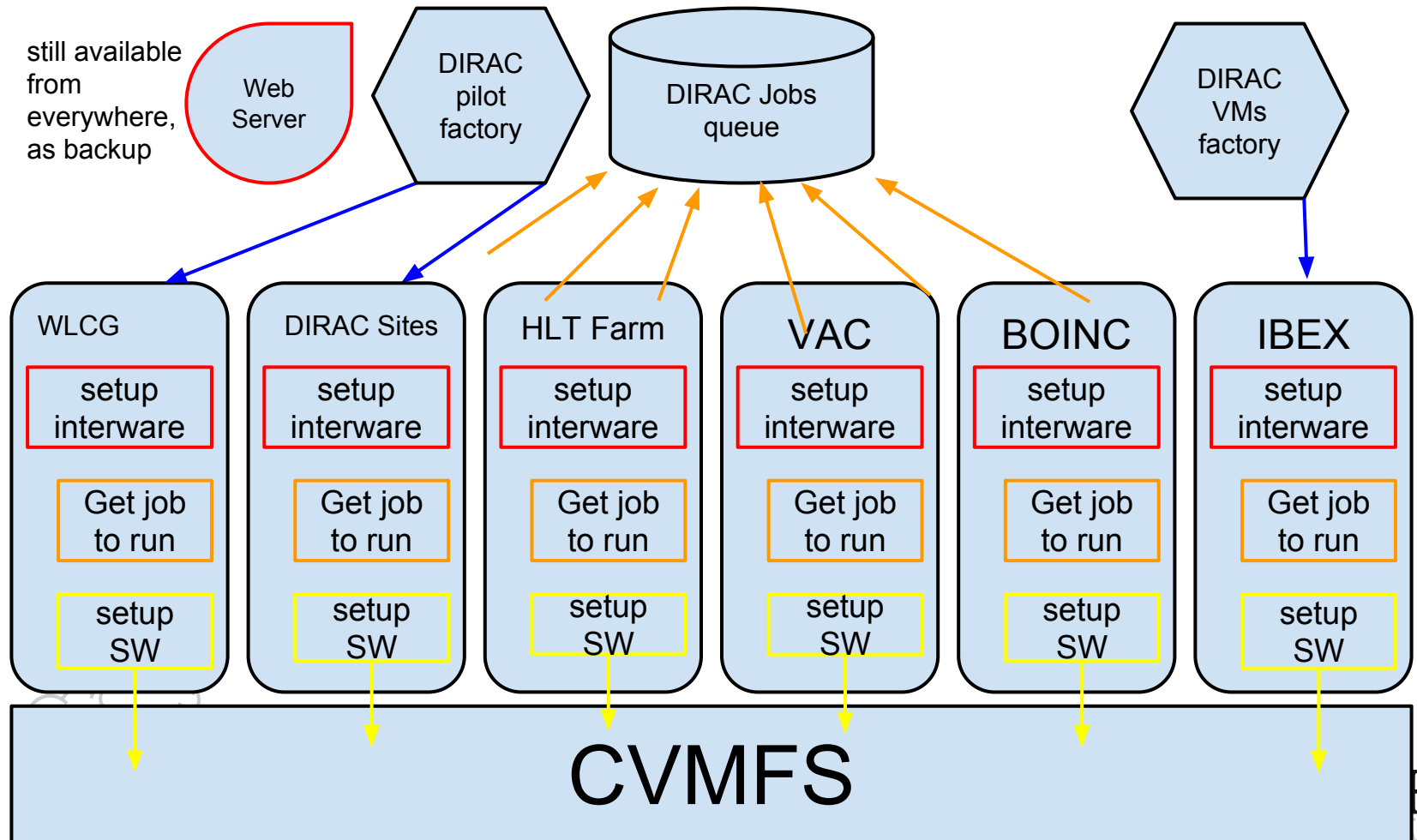
Window from points in time when VMs are active (from timestamp) to JobAgent running (to timestamp).



what we have



what we'll have soon



what we had

WLCG sites - AKA the "Grid"

EDG, EGEE, EGEE-II, EGEE-III, EGI, EMI, EMI2, EMI3...

- WMS
- CEs (LCG, then CREAM)
- Batch queues in front of the WNs

