Virtual PBS

Marc Rodrigues
for
Port d'Informació
Cientifica

Motivations I

- What distinguishes Grid computing from conventional high performance computing systems, such as cluster computing, is that Grids tend to be more loosely coupled, more heterogeneous, and geographically distributed (http://en.wikipedia.org/wiki/Grid_computing)
- Unfortunately the level of heterogeneity is never high enough to satisfy all users

Motivations II

- Changes of architecture, OS upgrades often require big synchronisation efforts between developers and experiments who use the same batch system.
- Hardware improvements often require software updates, e.g. OS updates.
- Experiments may run their software on different Linux platforms and they need to do a extra effort to adapt their software to “default” batch system platform.

Why PBS?

- Why PBS , if it has not support for vitual machines?
  - PIC uses PBS since its initials days.
  - PIC has obtained extended knowledge on PBS.
  - Changing to another batch system is a tremendous challenge in an operative site.
  - PBS looks simple enough to do the project.
  - Why not? No batch system is perfect.

Proposal

- A specific environment , customized for project or single job requirements.
- An isolate scope in a real machine where the project can create a perfect environment to run her jobs without problems.
- A system, transparent to the users and to the middleware .

What we try to do

- Achieve that PBS executes virtual machines.
- PBS must control the virtual machines .
- Any virtual machines can be executed in one node.
- The job owner owns also the virtual machine.
- Try to minimize the disc space requirements.

Achieve that PBS executes virtual machines

- We use PBS prolog and epilog to start, and stop virtual machine

PBS must control the virtual machines

- After some test we decided use KVM instead of XEN because the last one can't be controlled by PBS

Execution of any virtual machines in one node Try to minimize the disc space required

- To avoid the problem of needing an image duplication we use KVM snapshot.
- KVM snapshots permit to re-use the same image for all VM that are running on the same node without interferences

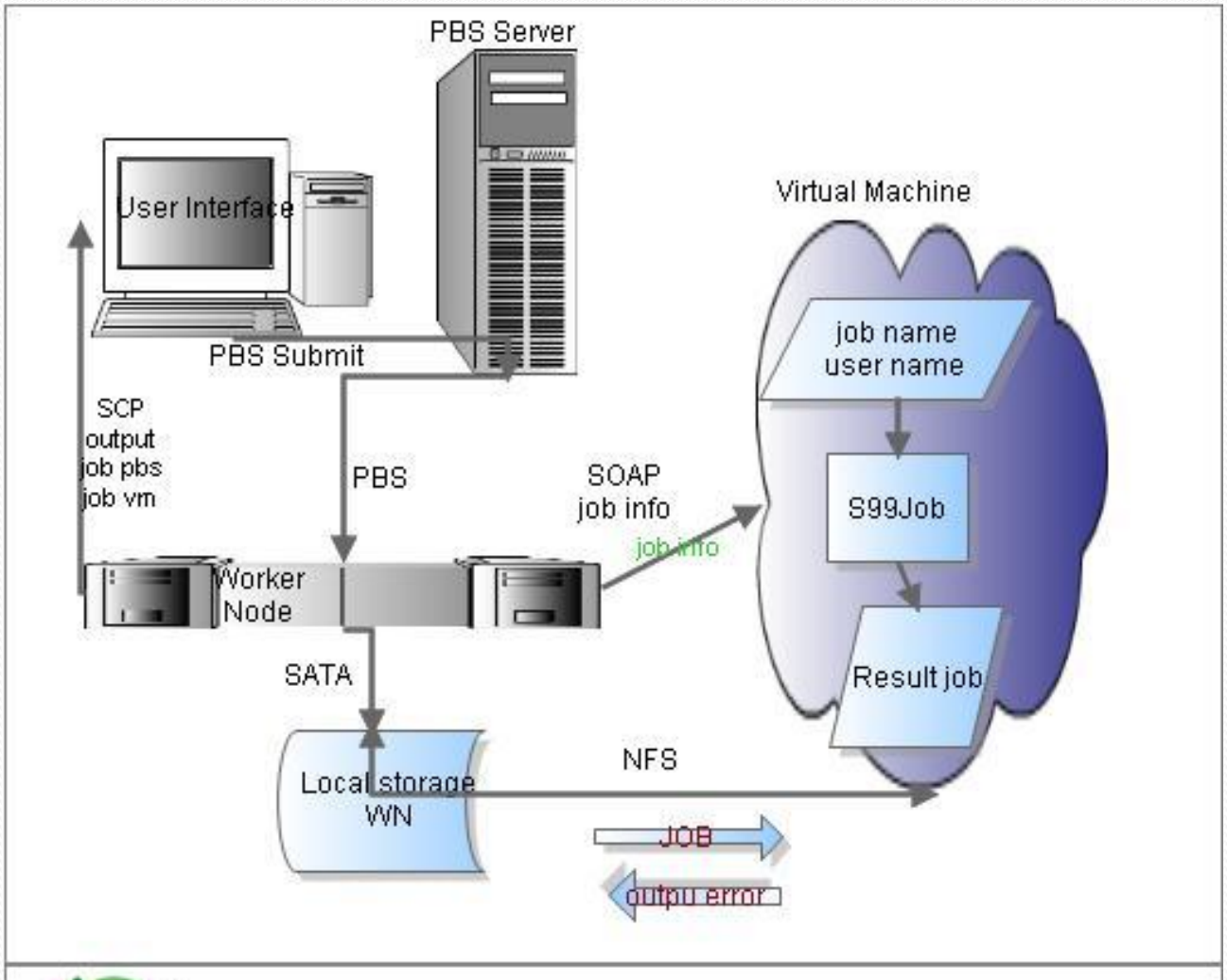
The job owner owns also the virtual machine

- With some modifications KVM can be executed by a normal user , and the process owner is the same one. This is one reason for use KVM vs XEN

The output will be returned by PBS

- After some trying several approaches, the simplest solution, was that the virtual machine writes a stdout, stderror on a tmp file, and at the end re-directs this file to stdout and stderror

SQUEMA V1.1


PBS Server
User Interface
Virtual Machine
job name
user name
PBS Submit
SCP
output
job pbs
job vm
PBS
SOAP
job info
job info
S99Job
Worker
Node
Result job
SATA
NFS
Local storage
WN
JOB
output error

Some tests

First prototype was tested, the first results show :

- CPU bound jobs had very little overhead
- The network bound jobs with virtio driver, had acceptable overhead
- The jobs wih I/O disk bound had high overhead

Real tests

We tested already one type of job . The job was from the MAGIC project and has intensive I/O. It reads files from NFS, applies a filter and stores new files in NFS

- The job run on 31 minutes on a real machine
- The job run on 36 minutes on a virtual machine
- The virtual machine had the same configuration as real machine, on next test we tried to do a fine tuning to the virtual machine to increase the performance

Some issues

- The memory use increases approximately 5%
- The PBS accounting counts from the start to the stop of the execution of the virtual machine .
- For some jobs probably the virtualization is not a solution, but a great number of small projects may find benefits with VMs because they can use their perfect environment and in some cases, the job run faster than hostile environment on real machine.

TO DO V1.1

- Change the method to access job from VM
  - Nowadays are from NFS
- More performance test and more tunings
- Implement an easy method to configure the few things that the VM needs to run on the system
- An initiation script for server scripts.

V2.0

- The V1.1 is functional and the next versionsolves the last issues

  But we try to increase and centralize the VM control and try to decrease de number of CPU cycles lost on start-up and shutdown we are developing a new version, slightly more intelligent than the first version

- To do that we are looking two approach's
  - ADDonPBS
  - DIRAC

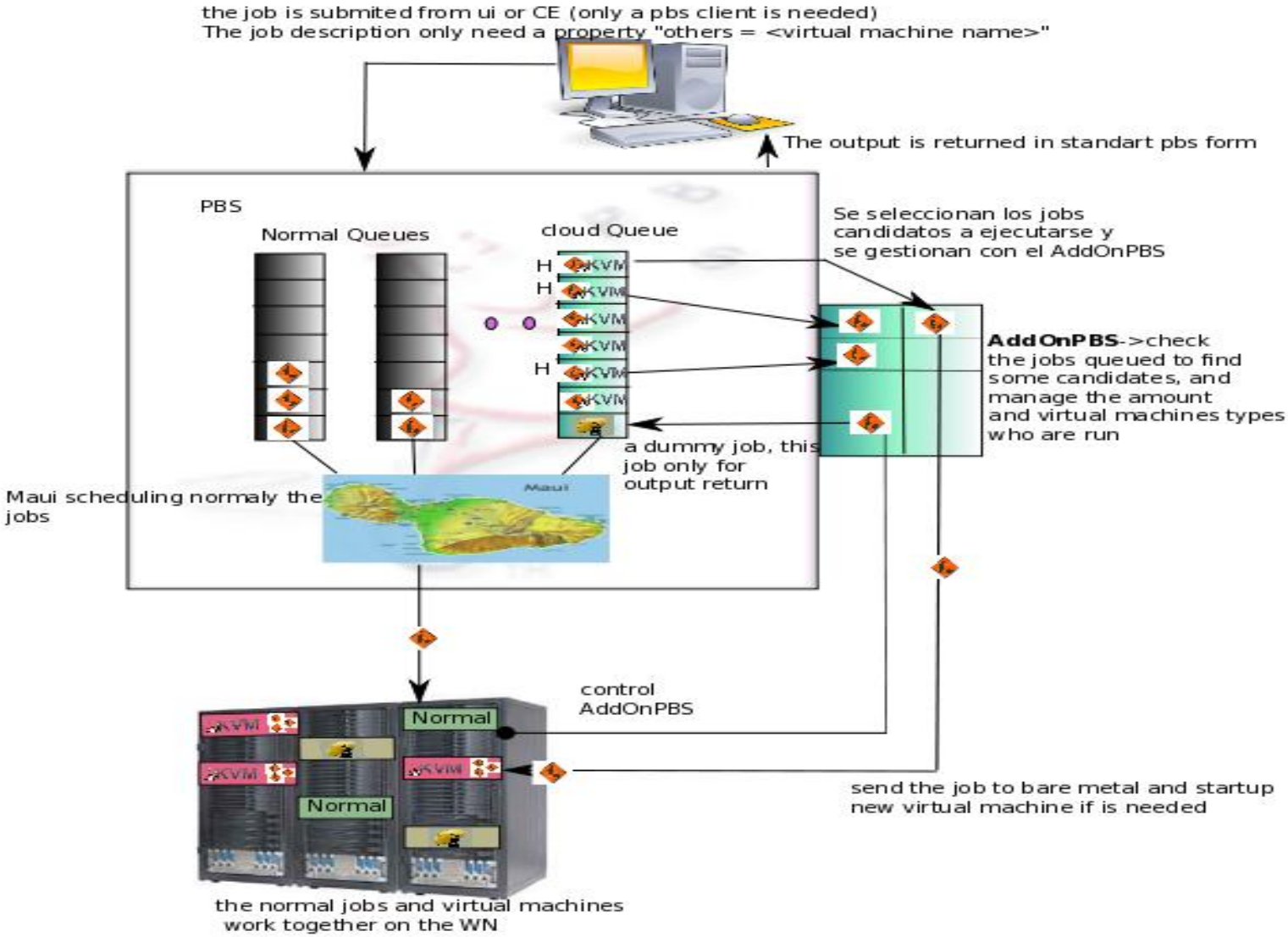
The issues of V2.0

- Balance the number of virtual machines.
- Scheduling the jobs according to the number and types of virtual machines that are running.

ADD on PBS

- ADD on PBS is an additional scheduler for maui, we are doing some modifications like add-on over the maui to control the VM and scheduling it.
- We don’t want to replace the original scheduler ,
- We don’t want to modify the maui scheduler.

AddOnPBS


the job is submited from ui or CE (only a pbs client is needed)
The job description only need a property "others = <virtual machine name>"
The output is returned in standart pbs form
PBS
Normal Queues
cloud Queue
Se seleccionan los jobs candidatos a ejecutarse y se gestionan con el AddOnPBS
AddOnPBS->check the jobs queued to find some candidates, and manage the amount and virtual machines types who are run
a dummy job, this job only for output return
Maui scheduling normaly the jobs
Maui
control AddOnPBS
Normal
Normal
send the job to bare metal and startup new virtual machine if is needed
the normal jobs and virtual machines work together on the WN

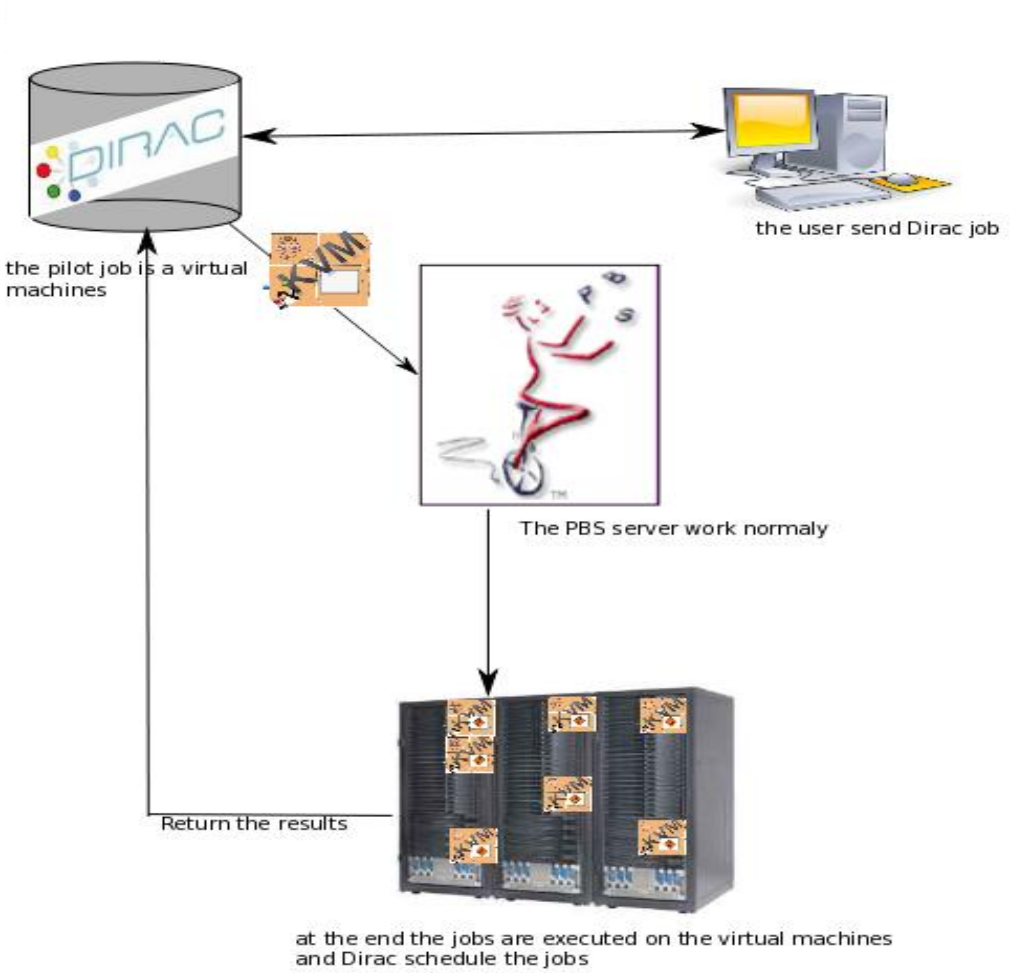
ADD on PBS pros & contras

- benefits
  - Is transparent for PBS clients
  - Permits a better control over the VM jobs through a better scheduler
- problems
  - Is not transparent to PBS admin and need more complex scheduling rules
  - The jobs are execute at a certain time but the results are available at maui earliest atscheduling time,

DIRAC

- The other approach is DIRAC. We want to modify a pilot job to execute a virtual machines.
- On this approach the scheduler are DIRAC system who has the responsibility to organise the jobs correctly

V2.0 DIRAC


DIRAC
the user send Dirac job
the pilot job is a virtual machines
KVM
The PBS server work normaly
Return the results
at the end the jobs are executed on the virtual machines and Dirac schedule the jobs

DIRAC pros & Contras

- benefits
  - Transparent to admin of the batch system
  - Software maintained by others and less development needed
- problems
  - Isn't transparent to the user who need to work with DIRAC platform
  - We need to modify a software developed for others
  - Less control on the job scheduling

Conclusions

- The VM can be a great help for our batch division
  - We can create a heterogeneous cluster under homogeneous platform
- The users can execute their jobs on a big number of platforms without changes in our cluster
- The installation of Worker Nodes made more easy because we don’t worry about compatibility problems between software's
- The lost of performance will be reduced by technology soon.