Virtual Cluster Computing in IHEPCloud

Haibo Li, Yaodong Cheng, Jingyan Shi, Tao Cui
Computer Center, IHEP
lihaibo@ihep.ac.cn

HEPIX Spring 2016

Contents

- Background
- Overview of IHEPCloud
- Virtual Cluster Computing Status
- Future Work

Background

- Low resources utilization rate
  - Utilization of computing resources is less than 60% on average
- Computing resources are non-shared
  - Every experiment such as BESIII,YBJ, has its own computing machine.
- Peak computing requirement
- The penalty of job running on virtual machine is optimistic.

Virtual machine performance test (1)

- BES simulation job
  - Same number of jobs running on physical vs VM, each VM runs one job.
  - The number VM on physical machine(24 cores):1,12,24

```
//This file is generated from " /bes3fs/offline/data/job/665/tmp109/mctest/test1/mctest1" by the program!
RealizationSvc.InitEvtID=103265;
#include "$OFFLINEEVENTLOOPMGRROOT/share/OfflineEventLoopMgr_Option.txt"
#include "$KKMCROOT/share/jobOptions_KKMC.txt"
KKMC.CMSEnergy= 3.686;
KKMC.BeamEnergySpread=0.00092;
KKMC.NumberOfEventPrinted=1;
KKMC.GeneratePsiPrime=true;
#include "$BESEVTGENROOT/share/BesEvtGen.txt"
BesRndmGenSvc.RndmSeed=724432;
#include "$BESSIMROOT/share/G4Svc_BesSim.txt"
#include "$CALIBSVCROOT/share/calibConfig_sim.txt"
RealizationSvc.RunIdList={-8997};
#include "$ROOTIOROOT/share/jobOptions_Digi2Root.txt"
RootCnvSvc.digiRootOutputFile="/scratchfs/bes/offline/shijytest/tmptestfile/mctest_10.rtraw";
MessageSvc.OutputLevel= 6;
ApplicationMgr.EvtMax=4900;
```

Job	alltime	usertime	CPU	slow
1-pm	3318.51	3303.13	99.5%	
1-vm	3427.12	3391.56	98.9%	3.3%
12-pm	3761.75	3740.76	99.5%	
12-vm	3862.58	3828.31	99.1%	2.7%
24-pm	3786.45	3750.01	99.5%	
24-vm	3870.08	3829.19	98.9%	2.2%

- Experiment environment
  - Virtual machine:1CPU cores，2GB memory
  - Physical machine:24CPU cores，16GB memory
- Experiment Result:
  - 1 job :Running time penalty on VM is about 3%
  - 24 job: 2%

Virtual machine performance test (2)

- BES reconstruction job
  - Same number of jobs running on physical vs VM, each VM runs one job.
  - The number VM on physical machine(24 cores):1,12,24

```
DatabaseSvc.ReuseConnection=false;
MessageSvc.OutputLevel= 6;
RawDataInputSvc.InputFiles={"/bes3fs/offline/data/job/700/test/virtual/rec-run_4.raw"};
EventPreSelect.WriteDst=true;
EventPreSelect.WriteRec=false;
EventPreSelect.SelectBhabha=false;
EventPreSelect.SelectDimu=false;
EventPreSelect.SelectHadron=false;
EventPreSelect.SelectDiphoton=false;
WriteDst.digiRootOutputFile="/besfs/offline/data/700-1/test/dst/virtual/rec-run_4.dst";
EventCnvSvc.digiRootOutputFile="/besfs/offline/data/700-1/test/tmp/virtual/rec-run_4.tmp";
ApplicationMgr.EvtMax= 1250000;
```

Job	alltime	usertime	CPU	slow
1-pm	6409.75	6394.53	99.7%	
1-vm	6642.33	6632.84	99.3%	3.6%
12-pm	7333.58	7305.78	99.7%	
12-vm	7639.41	7583.24	99.4%	4.2%
24-pm	7366.25	7333.02	99.7%	
24-vm	8564.37	8286.49	97%	16.3%

Network I/O consumption cause high IOWait

- Experiment environment
  - Virtual machine:1CPU cores，2GB memory
  - Physical machine:24CPU cores，16GB memory
- Experiment Result:
  - 1 job :Running time penalty on VM is about 3%
  - 24 job: 16.3%

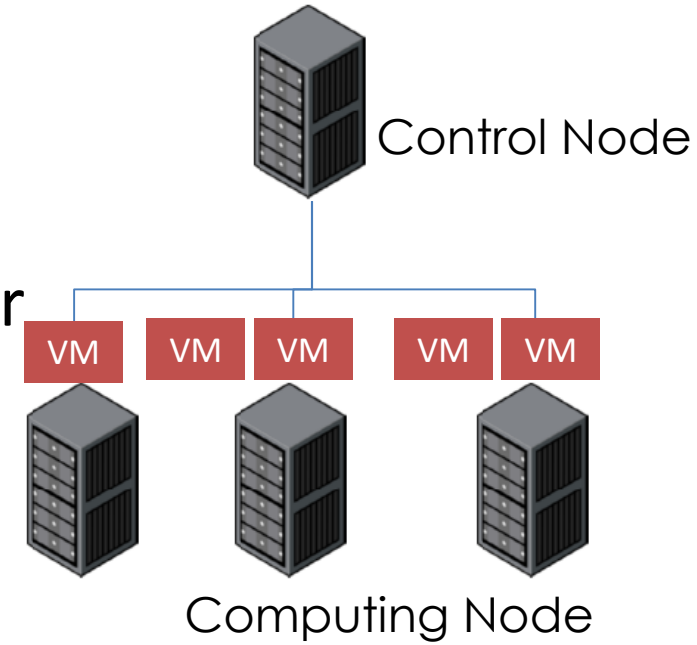
Overview of IHEPCloud

IHEPCloud Introduction

- Provide cloud services for IHEP users and computing demand
- Based on Openstack
  - Launched in May 2014
  - Performed 4 rolling upgrades
  - Currently base on **Kilo**
  - **Virtual Computing Cluster is an use scenario in IHEPCloud**.

IHEPCloud Status

- 1 controller, 29 computing nodes
- ~ 720 CPU cores
- Job queues managed by HTCondor and Torque
- Support LHAASO ,JUNO ,CEPC currently

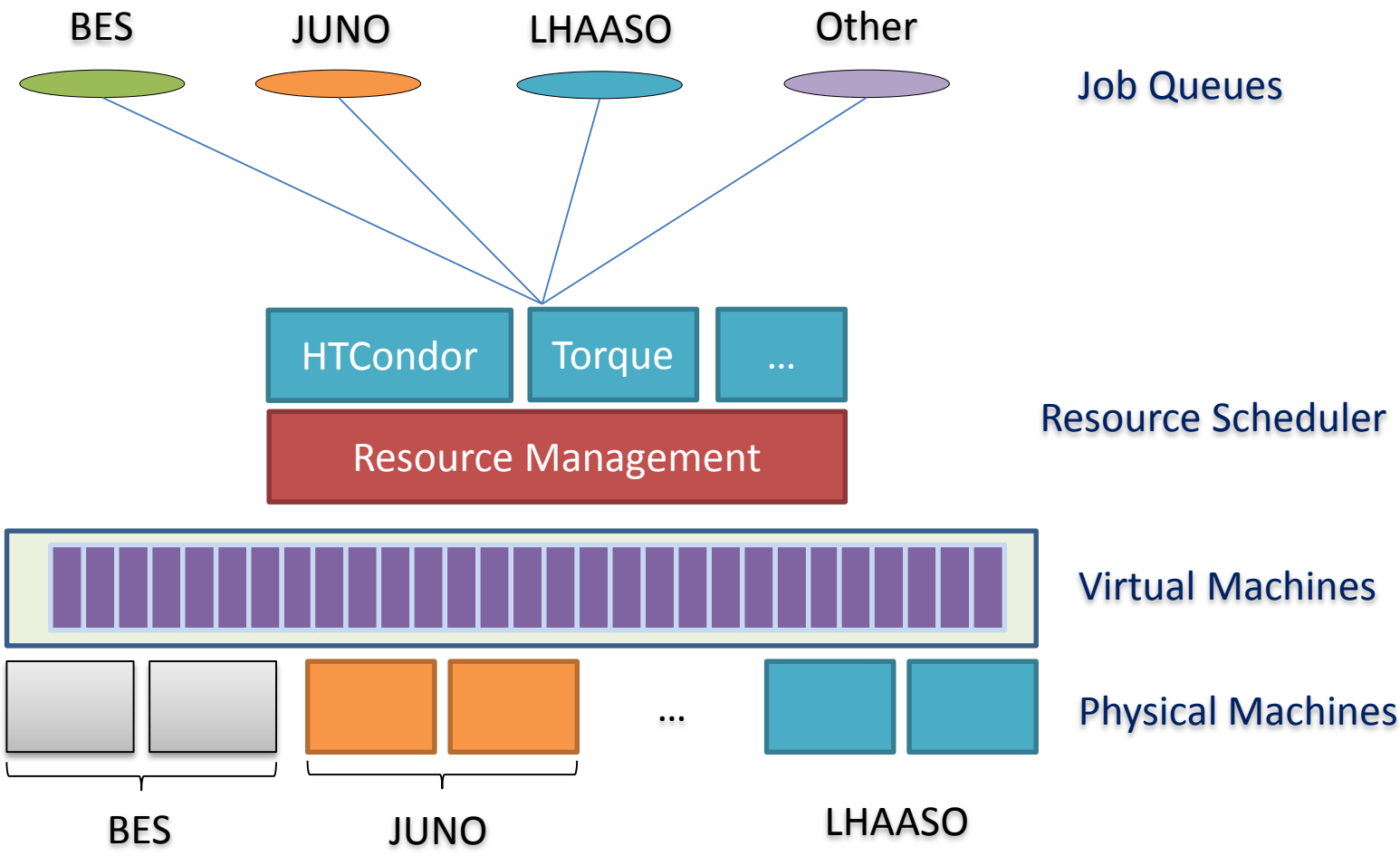


Virtual Cluster Computing Status

Features

- Easy to be used for different experiments
- Provide dynamic virtual resource **on demand**
- User transparent
- Two types of Cluster in IHEPCloud
  - Static virtual cluster
  - Dynamic virtual cluster

Architecture of Virtual Computing Cluster



Key Technology

- Resource Pool Management
- Dynamic Scheduling

Resource pool management

- Different experiments have different resource queue
- Resource Quota management

Resource Name	Min	Max	Available	Reserve_time
LHAASO	100	400	200	600s
JUNO	100	300	200	600s

Dynamic scheduling

- Support different batch systems
  - Torque, HTcondor
- Dynamic VM supplies
  - Virtual machines are created and destoryed as application demand
- Fair-share algorithm
  - Guarantee resources are equally distributed among experiments

Openstack API list

- A list of openstack API written in python to control the VM

```
class CloudControl:
    def get_active_vm(self):#return active vm list
    def get_instance_from_ip(self,ip):#return instance id
    from ip address
    def create_vm(self,sname,imageid,flavorid):#create a
    new vm
    def delete_vm(self,serverid):#destory a vm
    def get_vmstat(self,serverid):#get a vm status
```

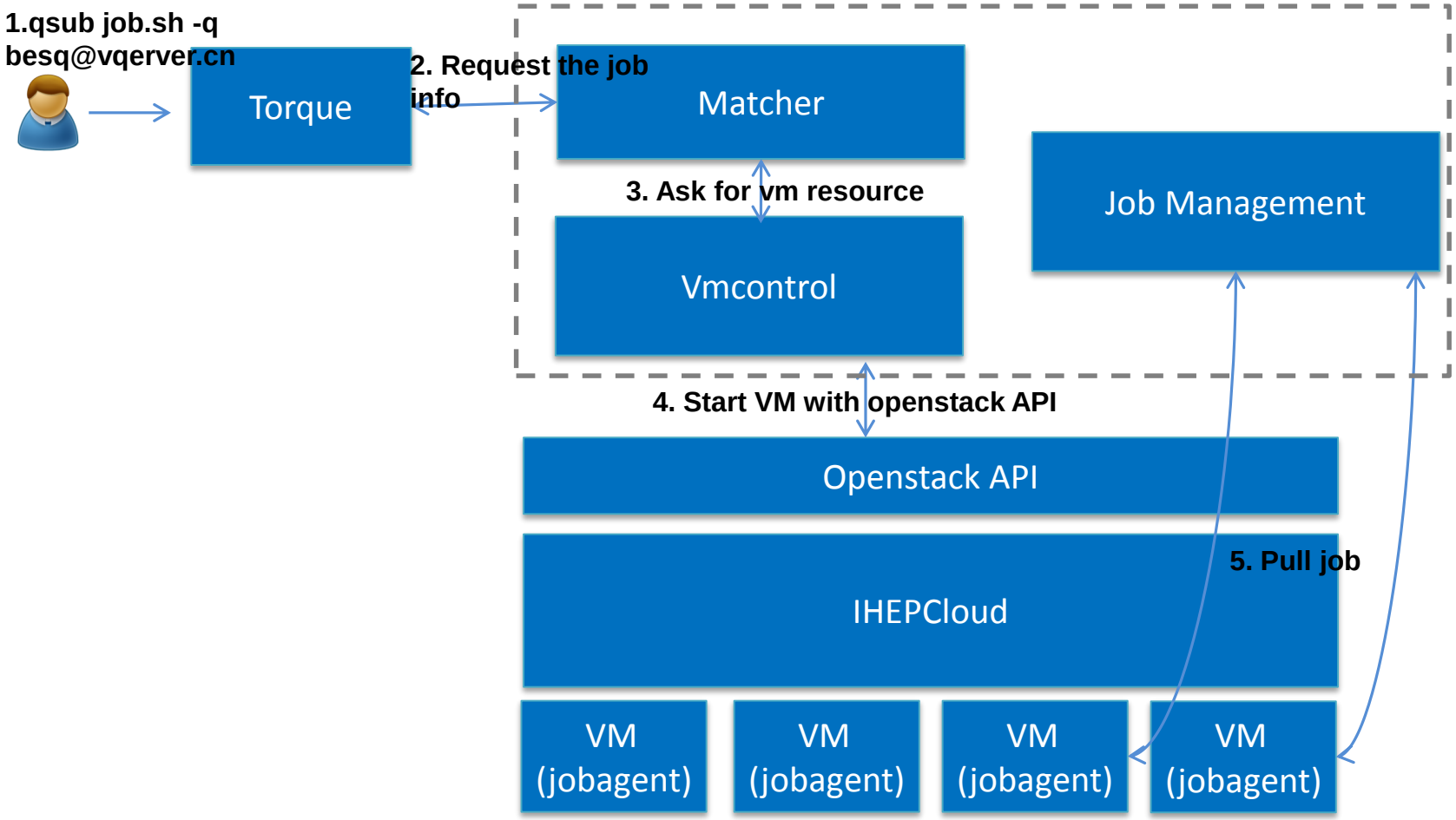
Dynamic Virtual computing

- Virtual PBS (VPBS)
- Virtual Condor (Vcondor)

VPBS

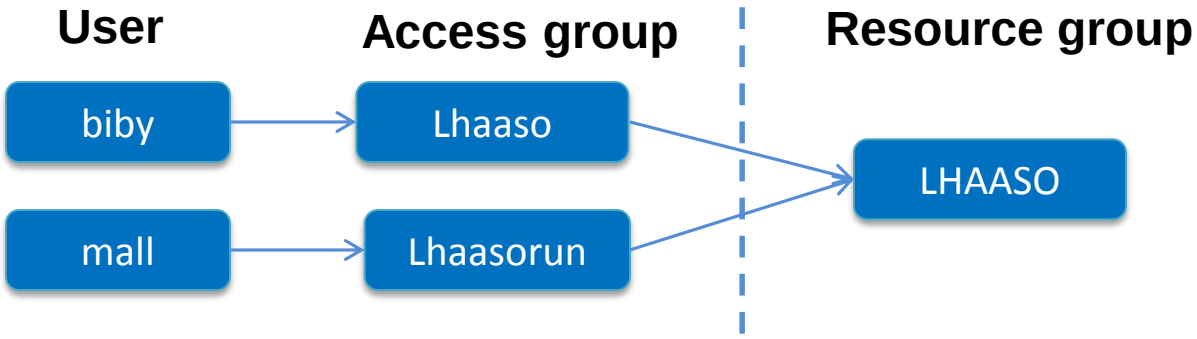
- Integrated with Torque
- Pull-push mode
  - When a job comes, Matcher will ask for the specific virtual machine.
  - When vm starts, it will request jobs.
- Jobagent in VM
  - A pilot running at virtual machine, to pull job and monitor vm status.

VPBS Structure

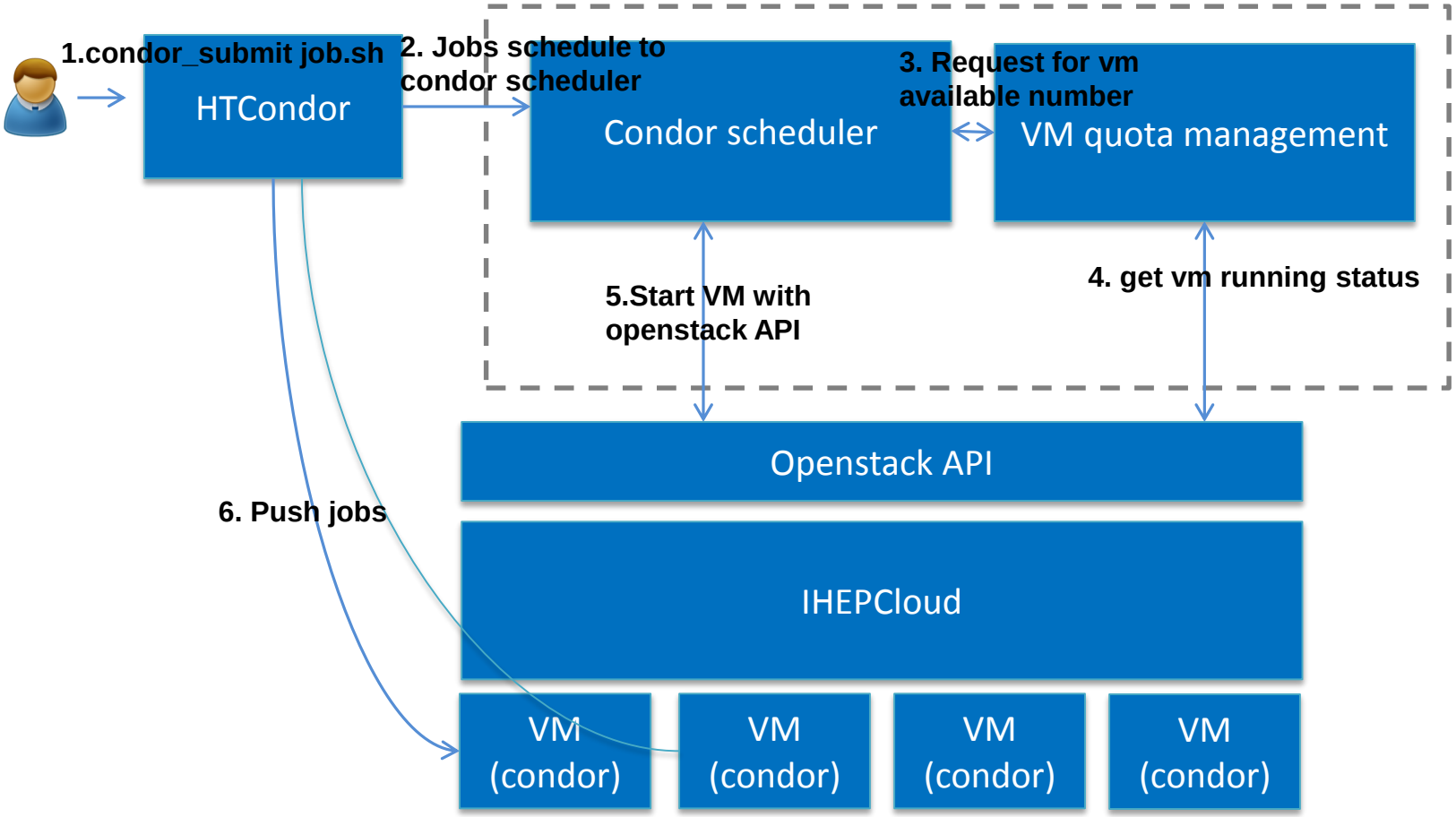


VCondor

- Use HTCondor as the new scheduler
- Push mode
  - Once vm starts, it will join in the server automatically
- How vm join in the right experiment pool?
  - Access group: which group user belongs to.
  - Resource group: which resource a group can use.
  - Mapping table

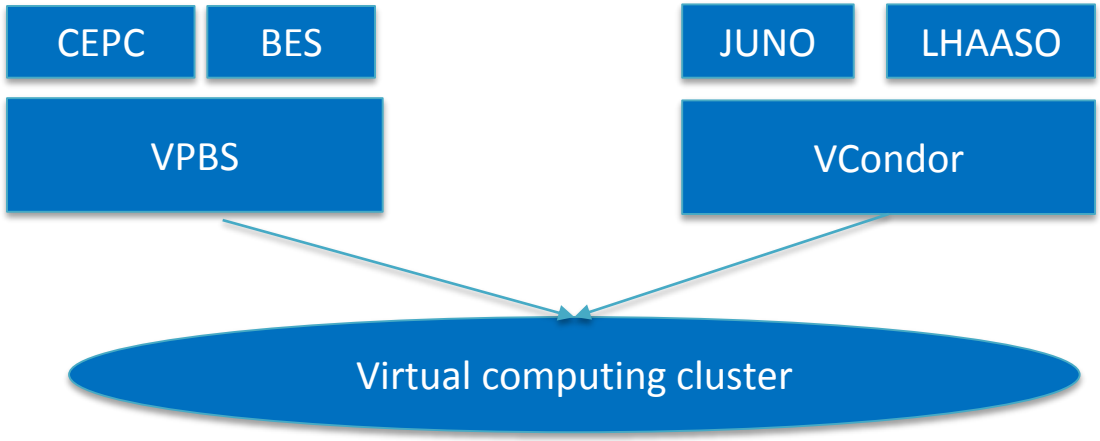


Vcondor structure

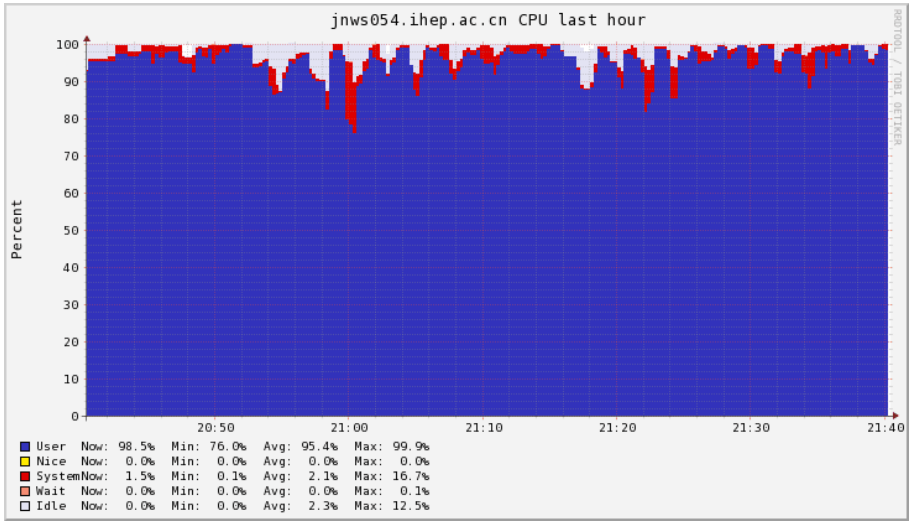


Running Status

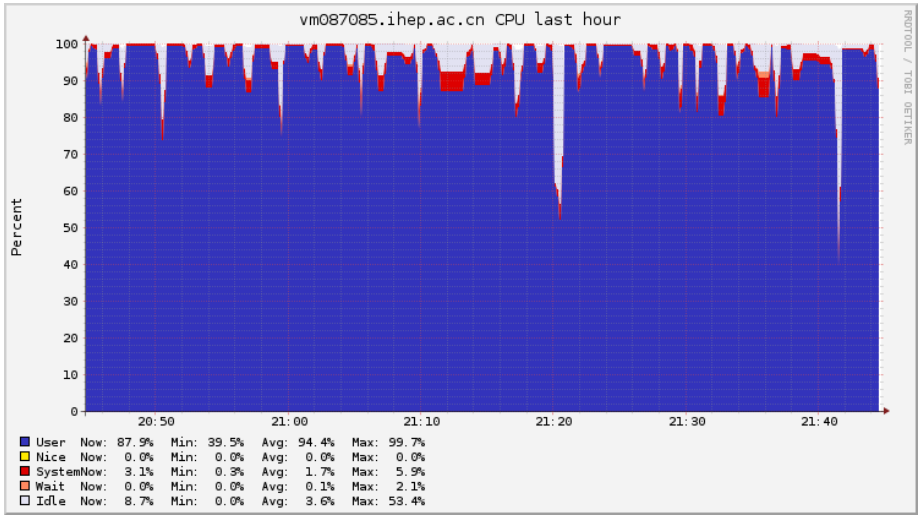
- Support three experiment
  - VPBS CEPC: ~1600 jobs,12300hours a week.
  - Vcondor LHAASO: ~1700 jobs,23500 hours a week.
  - Vcondor Juno: ~ 45500 jobs(mainly short job, average 152s), 1924hours a week.



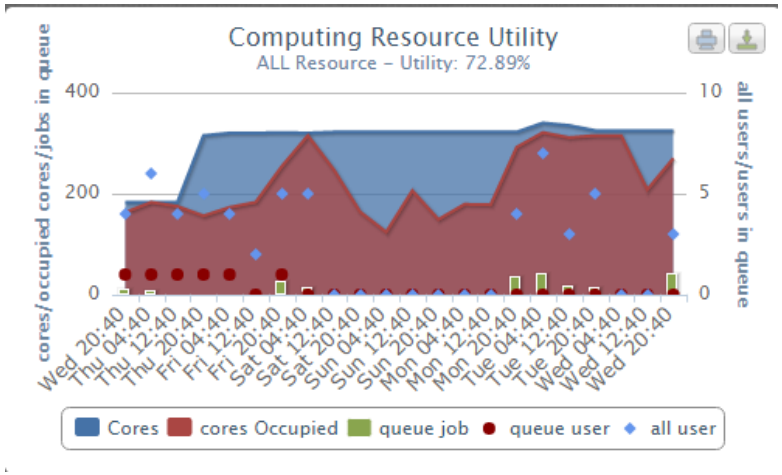
JUNO running performance: physical vs vm machine



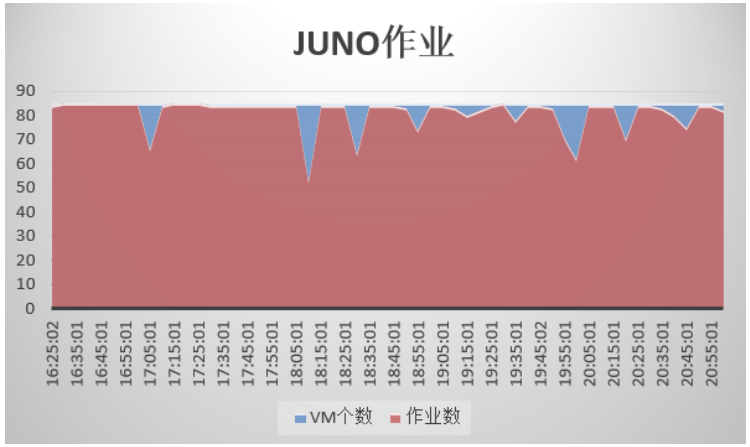
Physical machine
CPU use ratio: avarage 95.4%, highest 99.9%
System consumes: avarage 2.1%,highest 16.7%



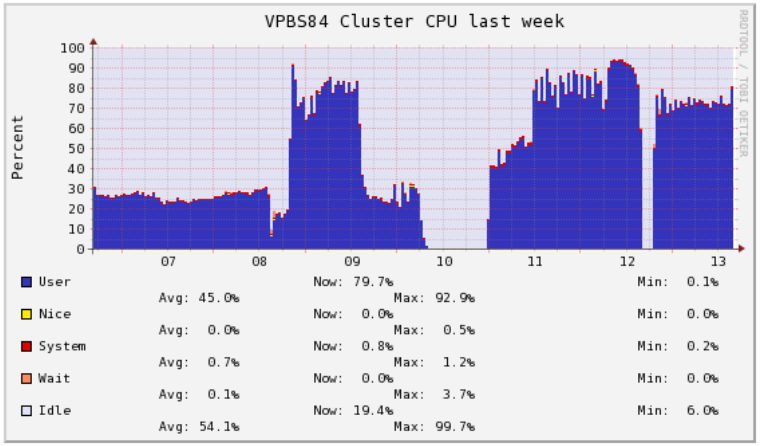
virtual machine
CPU use ratio: avarage 94.4%, highest 99.7%
System consumes: avarage 1.7%,highest 5.9%



Vcondor jobs statistics
(LHAASO - static cluster)



Vcondor jobs statistics
(JUNO: dynamic cluster)



VPBS jobs statistics
(CEPC)

Future work

- Extend Openstack cells and controllers
  - When vm instances reaches a certain range, the cloud management becomes complex.
- Accounting
  - Job accounting
  - Virtual resource accounting
- Job Scheduler Policy
  - Resource preemption, job deleting policy.

Thank you!