Integration of Openstack Cloud Resources in BESIII Computing Cluster

Haibo Li
lihaibo@ihep.ac.cn
Computing Center, Institute of High Energy Physics, Chinese Academy of Sciences
International Symposium on Grids and Clouds (ISGC) 2015
Content

- Motivation and approach
- Openstack in IHEP
- Design and Implementation
- Summary
Motivation

• Offline data processing platform is foundation for particle physics experiment.
  • The particle physics experiment scale expands unceasingly, BEPCII has accumulated more than 5 PB data.
  • Existing computing center faces many challenges from the data-intensive computing.

• The overall computing resources utilization rate in IHEP computing cluster is low.
  ▪ IHEP Computing cluster is composed of many types of experiment, such as BES, Daya Bay, YBJ…
  ▪ Each experiment group has its own computing node, resources cannot be shared.
  ▪ At a certain period of time, some subset are running at full capacity, some subset are idle. Queuing time is too long.
HEP job KVM virtual machine performance testing

- **BES Simulation job**
  - Jos run on physical vs virtual machines, case number:1000
    
    ```
    #include "$BESSIMROOT/share/G4Svc_BesSim.txt"
    #include "$CALIBSVCROOT/share/calibConfig_sim.txt"
    RealizationSvc.RunIdList= {-9989};
    #include "$ROOTIOROOT/share/jobOptions_Digi2Root.txt"
    DatabaseSvc.SqliteDbPath="/panfs/panfs.ihep.ac.cn/home/data/dengzy/pacman_bak/database";
    RootCnvSvc.digiRootOutputFile="/scratchfs/cc/shijy/rhopi-bws0106-1.rtraw";
    MessageSvc.OutputLevel= 5;
    ApplicationMgr.EvtMax=10000;
    ```

- **Experimental environment I**
  - Virtual machine:1CPU cores, 2GB memory
  - Physical machine:8CPU cores, 16GB memory

- **Experimental environment II**
  - Virtual machine:2CPU cores, 2GB memory
  - Physical machine:8CPU cores, 16GB memory

- **Experiment Result:**
  - Experiment 1: running time on vm: 1:58:42, physical: 1:42:04, the performance reduce 16%.
  - Experiment 1: running time on vm: 1:45:05, physical: 1:42:04, the performance reduce 2.9%.
• BES Analysis job
  • Jos run on physical vs virtual machines Application Mgr.EvtMax = 1E9

"/besfs2/offline/data/663-1/jpsi/tmp2/120520/run_0028145_All_file006_SFO-1.dst",
"/besfs2/offline/data/663-1/jpsi/tmp2/120520/run_0028145_All_file006_SFO-2.dst"

// Set output level threshold (2=DEBUG, 3=INFO, 4=WARNING, 5=ERROR, 6=FATAL )
MessageSvc.OutputLevel = 6;
// Number of events to be processed (default is 10)
ApplicationMgr.EvtMax = 1E9;

ApplicationMgr.HistogramPersistency = "ROOT";

• Experimental environment
  ▪ Virtual machine:1CPU cores, 2GB memory
  ▪ Physical machine:8CPU cores, 16GB memory

• Experiment Result:
  ▪ Experiment :running time on vm:8:04:47,physical:7:48:32, the performance reduce ~3%.
  ▪ using virtual machine to run BES job, the performance loss is acceptable.
Approach

• Integrated virtual computing cluster into the original physical cluster to optimize the resource utilization.
• Take fine-grained resource allocation to schedule tasks instead of taking nodes.
• Design flexible allocating policy, considering job types, system load and cluster real-time status.
Why choose Openstack?

✓ Provides an open source platform.
✓ Protects clients current investment with simple path to new technology
✓ Open APIs provides great flexibility and agility
IHEP Cloud

- A private IaaS platform.
- Aims to provide virtual computing environment for IHEP computing system and users.
- Provides virtual computing cluster, user self-service virtual machine platform and distribute computing system.
Virtual computing cluster Design

- Integrated Openstack and PBS to provide virtual computing service based on IHEP cloud.
  - Virtual cluster and physical cluster work together.
  - When a job queue is busy, the jobs can be allocated to a virtual queue.
  - Virtual machines is created as application demand.

![Diagram with steps: Submit job, Check Queue load, Forward job, Start/stop VM, IHEPCloud]

- junoq: 128 CPU cores
System implementation

- **PBS (VM Queue)**
  - Expends the original Torque to support vm queue.

- **VM central controller**
  - A matcher between the various modules.
  - Polling-calculate-release

- **VM job scheduler**
  - Responsible for Job allocation.

- **VM resource controller**
  - Policymaker: make vm allocation strategy.
  - VM controller: used to start or stop vm.
  - CloudAPI: a packaged module based on the openstack api and with some extension.

- **Job agent**
  - Run on vm, monitor job run and vm status.
Workflow

PBS

VM central controller

VM Job Scheduler

VM Resource Controller

Job Agent

Request: each queue's running number and queue number

Reply: The maximum number of job each queue can run

check the VM queue and resource status

Request: queue name

Reply: each queue's job number, queue number, running ip addr

Request: the num of VM to be built, running ip list

Reply: the num of active VM

Request: start/stop VM
System Features

- Resource sharing
- Be transparent to the users
- Fine-grained elastic resource allocation
Elastic Resource Allocation algorithm

• Pull and push mode.
  • Pull: allocate the cpu/core for particle job with suitable resource allocation algorithm.
    ▪ When new job is coming, the PBS will request VM central controller to get vm resource.
    ▪ VM central controller prepares corresponding resource for the job.
  • Push: cluster internal keeps the original way of “push”, sending the job into the appropriate cores or virtual machines.
    ▪ The VM job scheduler will send the job to the above running vms.
Dynamic scheduler

• Issue different queue to different types of job.
  • Monte Carlo simulation, Physical reconstruction: physical cluster.
  • Physical data analysis: virtual cluster.

• Dynamic job scheduling strategy interface.
  • Linear addition and subtraction.
  • And so on…
Current progress

- System function design has been completed.
- Basic environment has been established.
- The primary version is debugging.

Next steps:
- Monitoring.
- Add more job and vm allocation strategy.
- Provide online service in the second half of 2015.
Summary

• Virtual computing cluster is a good supplement for the existing physical cluster.

• Openstack provides a good way to realize the virtual computing with sufficient resource utilization.
Thank you!
Any Question?