CHEP 2018 Conference, Sofia, Bulgaria

Elastic resource allocation for HEP experiments in hybrid cloud

Haibo Li, Qiulan Huang, Yaodong Cheng, Zhenjing Cheng
IHEP Computing Center, CAS
10 July 2018

Outline

- Requirements of scientific computing
- System architecture and implementation
- Use case in LHAASO
- Summary

Large science facilities

- IHEP serves as the backbone of China's large science facilities
  - Beijing Electron Positron Collider BEPCII/BESIII
  - Yangbajing Cosmic Ray Observatory
  - China Spallation Neutron Source (CSNS)
  - Hard X-ray Modulation Telescope(HXMT)
  - Jiangmen Neutrino Underground Observatory (JUNO)
  - Large High Altitude Air Shower Observatory (LHAASO)
  - Accelerator-driven Sub-critical System (ADS)
  - Under planning: BAPS, LHAASO, XTP, HERD, …

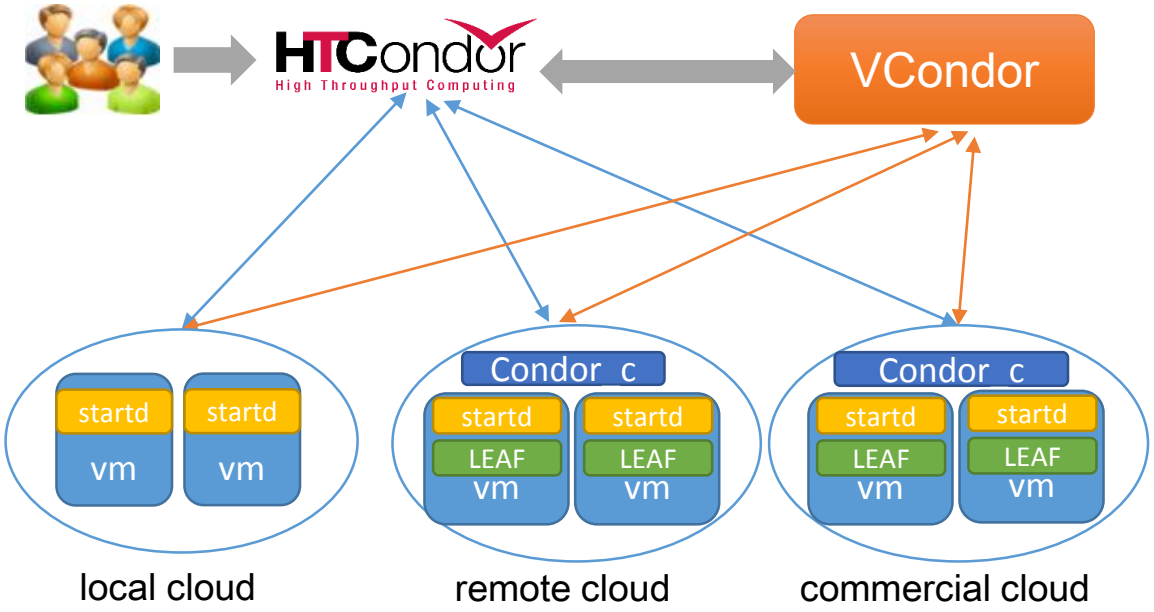
Requirements for cloud computing

- Each experiment has separate computing resources and no sharing
- Different experiments may have different peak times of resource usage
- Massive jobs with a little resource or vice versa
- Resources are distributed in different locations
- Remote site operation and maintenance ability is poor

Cloud computing is a good solution for scientific computing.

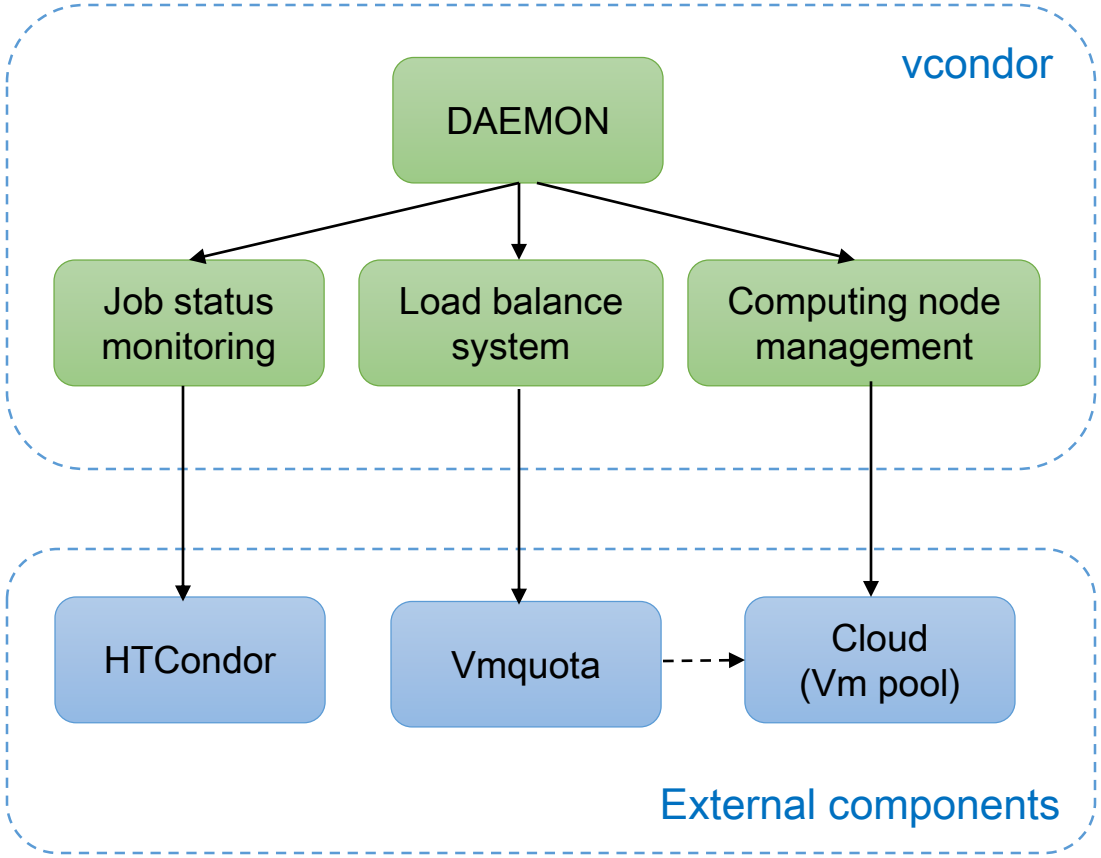
VCondor

- VCondor is a cloud scheduler providing elastic resource allocation service for hybrid cloud based on HTCondor
- A bridge between HTCondor and cloud resources
- Once the jobs submitted to HTCondor, VCondor will allocate dynamic resources for the jobs.



LEAF is a data cache and access system across remote sites. See talk: https://indico.cern.ch/event/567550/contributions/2628857/

Architecture



vcondor

- JobMonitor: query and record job information
- NodeManager: use REST API to create and destroy virtual machines
- DAEMON: Main module, periodically executed

vmquota

- Computing resource share management system

Components

- JobMonitor
  - HTCondor set different groups for each experiment
  - Get job and job queue information periodically from HTCondor
  - Information includes: group name, total job number, running job number, idle job number, etc
- Node Manager
  - According to the scheduling policy to create or stop VMs in hybrid cloud
- Load balancer
  - Interact with Vmquota to realize resource sharing management

Scheduling policy

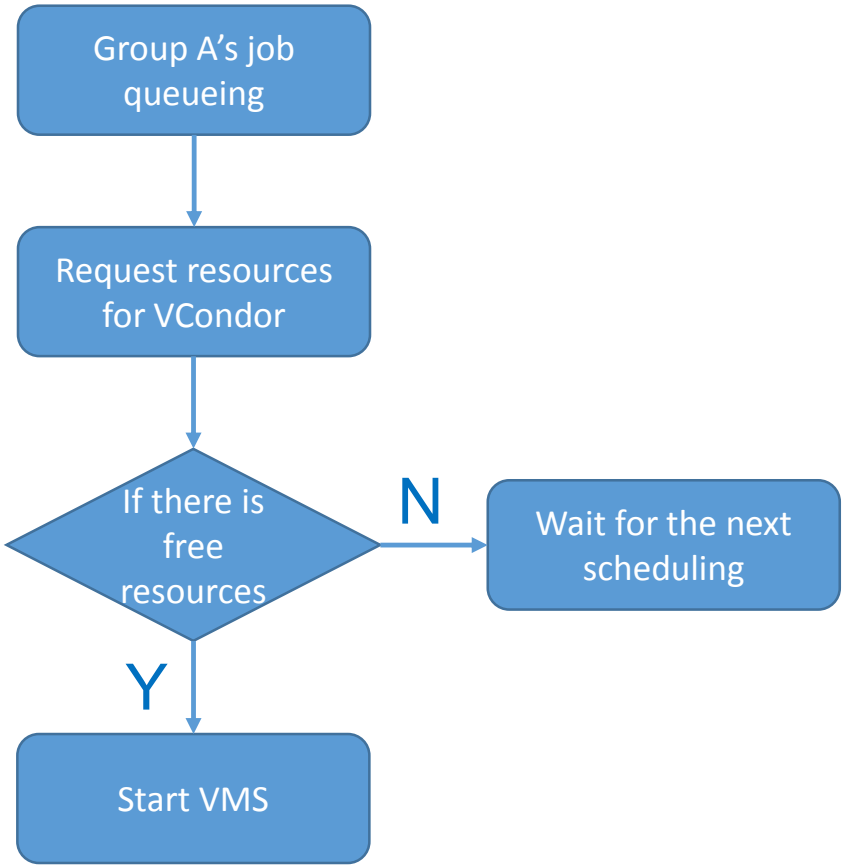
- Local cloud first, free cloud first
- The more resource the experiment shares, the more its jobs can be scheduled
- Jobs from free experiment have high priority

Resource management

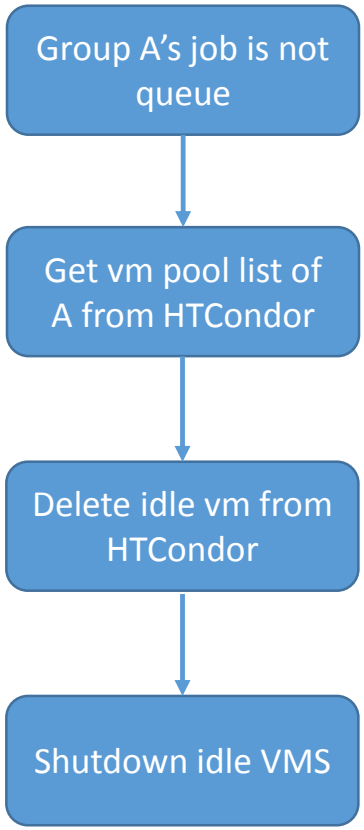
- Each experiment has a queue in vmquota database
- Queue attributes
  - Each queue has a minimum and maximum resource threshold
  - The number of minimum and maximum threshold is based on the contributions of experiments, the value is adjustable

Workflow

Resource expansion



Resource shrink



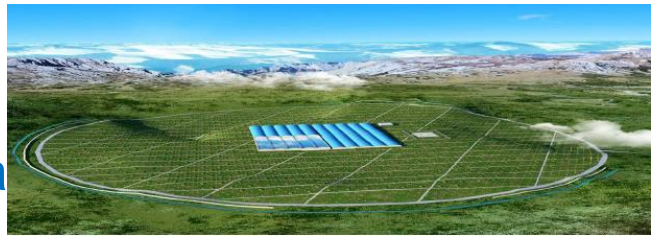
Remark

- VM Image preparation
  - Setup a VM Image with Condor installed
  - Different experiments may have different images
  - Essential softwares
    - AFS and CVMFS for user and software storage
    - LEAF client in outside cloud
- Quota settings
  - Set the minimum and maximum threshold for experiments in vmquota
- User transparent
  - User can use unified job tool and same job command

Current status

- Download VCondor from https://github.com/hep-gnu/VCondor.git
- Basic environment has been established at IHEP
- The primary version is applied to LHAASO
- Test with commercial clouds (such as Ali cloud) is on going

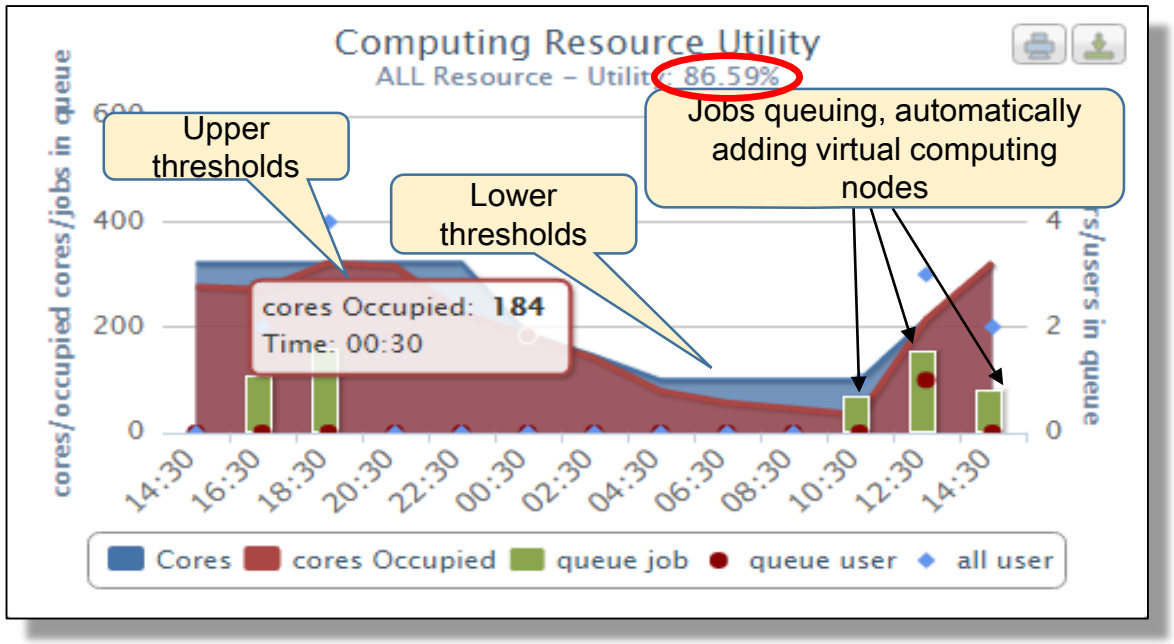
Use case in LHAASO(1/2)



- Large High Altitude Air Shower Observatory
- Located on Mt. Haizi (4410 MASL), Sichuan, China
- ~2 Petabytes (2 million Gigabytes) of data annually generated by the LHAASO detectors
  - 1.7PB of raw data, and >200TB of reconstruction data
  - Totally >20PB for ten years
- >2 Petabytes of data generated by MC simulation
- To build one distributed computing system containing about 6000 CPU cores to process the data
  - ~ 4500 CPU cores for reconstruction, analysis, …
  - ~ 1500 cores for production

Use case in LHAASO(1/2)

- Current Status:
  - Local cloud cluster, 1000 CPU cores
  - ~30,000 jobs, 250,000 CPU hours a week
  - Resource utilization reaches to 86.59%



LHAASO Resource Pool:
Automatically Scale up and down on demand

Summary

- VCondor enables elastic resource management
  - Has been used in IHEP for LHAASO
  - Provide support for HEP application resource sharing plan
- Next steps
  - Federated resources cross domains
  - schedule job across regions transparently

Thanks for your attentions!