PanDA for COMPASS: processing data via Grid

A. Sh. Petrosyan[a], E. V. Zemlyanichkina

Joint institute for nuclear research, 6 Joliot-Curie, 141890, Dubna, Russia

E-mail: [a] artem.petrosyan@jinr.ru

The development of PanDA, production and distributed analysis system as a workload management system for ATLAS, started in 2005. Since that time the system has grown up and in 2013 the BigPanDA project started, aiming to extend the scope of the system to non-LHC experiments. One of the experiments to which production management PanDA is being applied, is COMPASS at CERN. The workflow of the experiment has to be changed to enable Grid for production and user jobs. A large amount of the infrastructure work is being performed on backstage. The PanDA job definition replaces the native batch system job definition, an automatic submission to Condor computing elements comes in place of a console job submission, Grid user certificates identify job submitters instead of AFS user names, Grid storage elements substitute local directories on AFS and EOS. Production software moves from a private directory of production account to CVMFS. Also, a virtual organisation with role management has been established for the experiment. A central monitoring was enabled. The experiment has started to use several computing elements instead of local batch queues. The way how the COMPASS' data are being processed via Grid is presented in this paper.

Keywords: COMPASS, PanDA, workload management system, distributed data management, Grid

Introduction

Historically, data of COMPASS [Abbon et al., 2007] experiment is processed locally: jobs are being sent to CERN cluster (LSF) by production manager, logged in to lxplus.cern.ch with experiment production account. Since 2003 COMPASS collects 1.5 to 3 PB of data every year, and it is already a serious volume of data. In case of production and reproduction it takes long time to be processed on only one computing element still available for the experiment.

Having in mind wide distributed physics community of the experiment, using several computing resources to faster the processing of the experiment's data is a natural choice. Usually, workload management systems (WMS) are used to process data on several sites. WMS' have several key features and, in general, allow to work with many various computing sites as with one local queue. Common features of WMS' may be presented in a short list:

- Provide central queue for users, similar to local batch systems;
- Build a pilot job system;
- Hide middleware while supporting diversity and evolution;
- Hide variation of infrastructure;
- Use the same system for simulation, data processing and user analysis.

Back side of the medal is that installation and configuration of such systems and resources definition in them is quite complicated.

One of the systems, widely used in HEP and beyond is PanDA [Maeno et al., 2014], production and distributed analysis system. Started in 2005 as a WMS for ATLAS, now PanDA has grown to BigPanDA [Klimentov et al., 2015] platform, which is a crucial part of several projects, including several HPC utilisation optimisation projects, Cloud Distributed Operating System and Large Synoptic Survey Telescope.

The difference between ATLAS Grid and COMPASS Grid is that for ATLAS PanDA daily runs hundred thousands jobs on more than 200 sites while COMPASS at the moment has access to a single site. In such situation, many components of the system have to work in limited conditions, several have to be stopped and disabled. To proof that such complicated system may be configured to run COMPASS production jobs, the following list of action items was prepared:

1. PanDA instance installation;
2. Preparation of production chain management software;
3. Grid environment setup;
4. Validation that COMPASS software can work in Grid;
5. Production jobs execution;
6. Physics validation of the job execution results.

Definition of COMPASS' data flow, items 1 and 2 of the list above described in details in [Petrosyan, 2016]. In this article we will concentrate on the rest items.

Grid environment

PanDA designed to distribute jobs to many Grid sites. But, to enable data processing via Grid , several steps had to be performed:

- Virtual organisation (VO) should be established;
- Grid computing element will come in place of local batch queue;
- EOS storage element will replace EOS local directory;
- COMPASS software has to be installed on CVMFS, or on any place, which can be accessed from Condor computing elements.

These steps in common allow to run COMPASS jobs on any site in Grid.

Also, to send jobs to computing elements, auto pilot factory (APF) component of PanDA has to be configured. APF interacts directly with local batch queue management software on computing sites and allows to send as many jobs as site ready to accept and handle. All that user has to do is to declare his jobs in PanDA, and then APF takes care about sending them to the sites.

These steps were applied to COMPASS case. Virtual organization was created on CERN VOMS server (https://lcg-voms2.cern.ch:8443/). Several users were registered, including production account. Quotas on CERN Condor CE's were requested and received. With support of CERN EOS team, access to experiment's directories on the service was granted to users from COMPASS VO with production role. Production software was moved to new location on AFS so that it became visible and accessible from Condor CE's, after that the installation had to be validated.

Validation

In order to confirm that COMPASS software may work in Grid, desired version of software had to be moved to a location where Condor CE pseudo-user could get access to. While testing, permissions were changed in several places to allow read and execution access so that Grid user could successfully execute the software.

Several hundreds jobs were executed to make the chain work from the start till the finish: software version, access rights, production software, EOS storage element and local directories on AFS to store logs have to work as one system. Once this is done then we're ready for the next step, which is execution of real production jobs.

Real production execution

To test a real production, run of 2014 was chosen. It consists of 2804 raw data files, one per job, and, after processing, each job produces 3 result files: histogram, data summary tree, and event dump. Each must be stored on EOS in the directory reserved for files of each type. Besides, each job produces logs of the job, stderr, stdout. This log must be stored separately in the directory of COMPASS production account so that they may be easily analysed in case of errors. When results are ready, they usually being merged and copied to save space on EOS and, later, on long term storage on CASTOR. This part of work also had to be performed during the test production. Such logic of storing each output file in its specific directory is unusual for PanDA payload behaviour in other applications and required changes on PanDA pilot side.

Processing started on one queue at computing element on Condor CE at CERN. The queue allowed experiment to run 150 jobs at the same time. During the processing, second computing element became available. Running on two CE's, up to 300 simultaneously running jobs, all files were processed within one week. Average execution time of each job is approximately 6 hours. Average failure rate was 2%, caused mostly by network instability between elements of computing infrastructure of CERN at JINR. One job was not processed correctly due to network lacks between computing element and EOS storage. No other problems appeared during the processing.

Merging was done after all results of the run finished and produced their results. Due to large volumes of data merging can not be executed as single job, that is why size of result of each job must be tracked and then merging job is being sent with several input files. Merging job results stored to their own directory on EOS.

Physics validation

Results were validated by COMPASS physics coordinators. Several minor errors appeared and jobs were reproduced. The cause of most of the problems is connection lacks between job and PanDA server, the obvious way to solve that problem is to move PanDA server from JINR cloud service infra-

structure [Baranov, Balashov, Kutovskiy, Semenov, 2016], which is used mostly for development and testing, to a production-quality infrastructure with better network connectivity, for example to CERN.

Summary

Results of the performed work show, that PanDA can be used to run COMPASS production jobs, and, even more important, that COMPASS software works correctly in Grid environment. The following goals were achieved during the previous phases of the project:

- Grid environment was prepared for COMPASS, all elements of the chain work as a coherent system, allowing to send jobs to any Grid site which would like to participate in data processing;
- New production management software was built;
- More than 5000 jobs were executed, including 2800 jobs of a real run of 2014.

PanDA server, installed on single virtual machine in JINR cloud service, behaves well and shows impressive productivity and reliability while running thousands COMPASS production jobs in Grid environment and maintains high load of available computing elements.

Grid processing is enabled now for COMPASS community. Starting from this point experiment management may connect as much Grid sites as it's necessary to handle jobs flow and PanDA will distribute the load among them. Large work still awaiting in several areas: central file catalog, distributed data management and careful adaptation of production software for Grid.

At the moment documentation with step by step instructions is being prepared so that next submissions could be done by COMPASS production managers.

Next steps of the project imply:

- Better work with COMPASS production software exit codes so that job would be restarted automatically in case of problems;
- Enabling CVMFS for easy turning on new computing sites;
- Adding several new computing sites and start distributing jobs among them;
- Development of more intelligent production management system, which will take care of whole chain of steps of jobs submission, management and monitoring.

References

Abbon P. et al. The COMPASS experiment at CERN // Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment. — 2007. — Vol. 577, Issue 3. — P. 455-518.

Maeno T. et al. Evolution of the ATLAS PanDA workload management system for exascale computational science // Journal of Physics Conference Series. — 2014. — Vol. 513. — http://inspirehep.net/record/1302031/.

Klimentov A. et al. Next generation workload management system for big data on heterogeneous distributed computing // Journal of Physics Conference Series. — 2015. — Vol. 608. — http://inspirehep.net/record/1372988/.

Petrosyan A. Sh. PanDA for COMPASS at JINR // Physics of Particles and Nuclei Letters. — 2016. — Vol. 13, Issue 5. — P. 708-710.

Baranov A.V., Balashov N.A., Kutovskiy N.A., Semenov R.N. JINR cloud infrastructure evolution // Physics of Particles and Nuclei Letters. — 2016. — Vol. 13, Issue 5. — P. 672-675.