

COMPASS PRODUCTION SYSTEM: PROCESSING ON HPC

A.Sh. Petrosyan

Joint Institute for Nuclear Research, 6 Joliot-Curie, Dubna, Moscow region, 141980, Russia

E-mail: artem.petrosyan@jinr.ru

Since the fall of 2017 COMPASS processes data on heterogeneous computing environment, which includes computing resources at CERN and JINR. Computing sites of the infrastructure work under management of workload management system called PanDA (Production and Distributed Analysis System). At the end of December 2017, integration of Blue Waters HPC to run COMPASS production jobs has begun. Despite an ordinary computing site, each HPC has many specific features, which make it unique, such as: hardware, batch system type, job submission and user policies, et cetera. That is why there is no ready solution out of the box for any HPC, development and adaptation is needed in each particular case. PanDA Pilot has a version for processing on HPCs, called Multi-Job Pilot, which was prepared to run simulation jobs for ATLAS on Titan HPC. To run COMPASS production jobs, an extension of Multi-Job Pilot was performed. COMPASS Production system also was extended to allow to define, manage and monitor jobs, running on Blue Waters.

Keywords: COMPASS, PanDA, workload management system, distributed data management, production system, HPC

© 2018 Artem Sh. Petrosyan

1. Introduction

Processing on High Performance Computers (HPC) machines has several features in comparison with processing on Grid sites, which are traditionally used for processing in High Energy Physics (HEP). These differences are driven by architecture and must be taken into account in order to organize processing smoothly and may be presented in a list of the following items:

- unique hardware setup: each HPC is a unique machine;
- two-factor authentication is a usual type of authentication on HPCs and reduces accessibility of the resource to automatic processing systems;
- shared file system requires careful management of I/O operations in order to at least not to bother other users and system in common;
- batch system: each HPC has its own installation of one of the versions of scheduler, such as PBS or PBS Pro;
- unique set of software packages, managed by system administrators;
- absence of Grid middleware software, such as CVMFS, VOMS clients, data management tools;
- absence of Internet connection at working nodes;
- user and project policy usually require strict usage of the resource during limited period of time. Also, storage is limited and available for a lifetime of the project.

So, despite any Grid site, processing on HPC makes much more demands to each link of chain of processing in order to be to be safe and effective. Having an allocation on HPC does not mean that jobs will be executed faster than jobs of other users and projects. On a traditional Grid site user's job runs in an isolated environment, while on HPC careful management of CPU and I/O signatures of running processes has a crucial role in order to keep system and processes of other users safe.

In 2016 COMPASS [1] received first allocation on Blue Waters HPC [2], located in Urbana Champaign, University of Illinois. In late 2017 project of adaptation of COMPASS Production System [3-5] to work with Blue Waters has started.

2. Blue Waters overview

Blue Waters is a Cray [6] hybrid machine composed of AMD 6276 "Interlagos" processors (nominal clock speed of at least 2.3 GHz) and NVIDIA GK110 (K20X) "Kepler" accelerators all connected by the Cray Gemini torus interconnect. Since COMPASS software, as long as any HEP software, is not yet able to run on Kepler nodes, only Interlagos nodes are used for data processing. There are 22 640 compute nodes with 96 compute nodes having 128 GB and the remaining have 64 GB. There are 4 228 compute nodes with 96 compute nodes having 64 GB and the remaining have 32 GB. There are 26.5PB of storage under management of Lustre available in total. There is also a 250+PB tape storage.

Blue Waters is comprised of a robust and capable Local Area Network coupled with a redundant Wide Area Network to provide leadership class data transfer capabilities and resiliency. Through active monitoring and data collection this network is kept in optimal performance. At the same time, network access from worker nodes is not recommended, and results of processed jobs must be stored on a shared file system in the project or scratch directory in order to be later staged out or written to the tape for long term storage.

It is recommended that Globus Online (GO) [7] is used for file transfers to and from Blue Waters. Blue Waters has dedicated import/export resources to provide superior I/O access to the filesystems.

The batch environment is Torque/MOAB from Adaptive Computing which talk to the Cray's Application Level Placement Scheduler (ALPS) to obtain resource information.

The aprun utility is used to start jobs on compute nodes. Its closest analogs are mpirun or mpiexec as found on many commodity clusters. Unlike clusters, the use of aprun is mandatory on Blue Waters, which is not a Linux cluster but a massively parallel system (MPP), in order to start any jobs including non-MPI ones that run on a single node. If the PBS [8] script does not use aprun to start

the application the latter will start on a service node, which is a shared resource, and that will be a violation of the usage policy. Aprun supports options that are unique to the Cray. There are flags to set process affinity by the node, control thread placement, and set memory policy between the nodes in a job.

3. Production system adaptation for processing on Blue Waters

COMPASS production system automates all steps of data processing, including jobs generation, retries, consistency checks, stage-in and archive data to tape storage system. But, in case of Blue Waters raw data is being delivered manually via Globus Online endpoint by the project manager, thus, steps which cover stage-in and stage-out may be turned off. And, in order to enable processing on Blue Waters, task definition and job execution components of production system had to be changed. Full list of changes in production system is presented below:

- tasks definition: site selection and raw data location on Blue Waters were added to user interface;

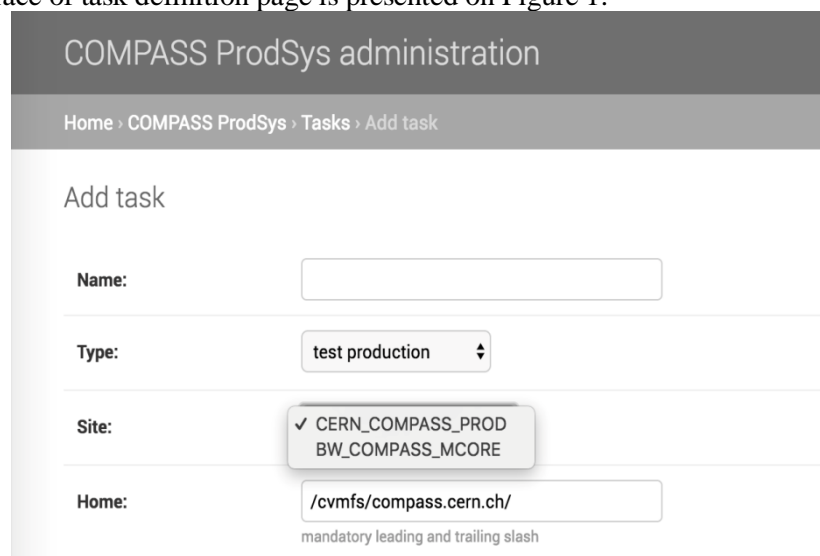
- data management components: stage-in and stage-out are turned off for Blue Waters tasks;

- jobs execution on site: PanDA Multi-Job Pilot [9-11] was used.

More details to each item are presented below.

3.1. Tasks definition

Since automatic data delivery to Blue Waters not yet available, a manual task assignment used in production system. Production manager selects site (Blue Waters) on which set of jobs will be executed. Interface of task definition page is presented on Figure 1.



The screenshot shows the 'COMPASS ProdSys administration' interface. At the top, there is a breadcrumb trail: 'Home > COMPASS ProdSys > Tasks > Add task'. Below this, the page title is 'Add task'. The form contains four fields: 'Name:' with an empty text input; 'Type:' with a dropdown menu showing 'test production'; 'Site:' with a dropdown menu showing 'CERN_COMPASS_PROD' (checked) and 'BW_COMPASS_MCORE'; and 'Home:' with a text input containing '/cvmfs/compass.cern.ch/'. Below the 'Home:' field, there is a note: 'mandatory leading and trailing slash'.

Figure 1. Task definition user interface

3.2. Data management

In case of Blue Waters task data stage-in from Castor tapes to Castor disks is not needed. This step was turned off, together with stage-out step which moves data from EOS to Castor. Both these steps are done at the moment by production manager. All other steps of automation for task definition and jobs generation works in the same manner as for the regular tasks.

3.3. Jobs execution on Blue Waters

For two previous steps there were few changes done in order to enable processing on Blue Waters. Much more changes had to be done in jobs execution layer.

Blue Waters requires two-factor authentication to log in. It means that, in order to submit jobs to local batch system, user must be logged in, credentials can not be generated somewhere outside as it

usually done on a regular Grid sites: authentication is done basing on X509 user proxy of submitter. On HPCs access by X509 user proxy is not available either, even more: proxy can not be generated on the machine due to absence of VOMS clients packages. But, since Internet is enabled on a login nodes, job definitions may be requested from the outside. On HPC systems, if user is logged in, he can run management software which will request jobs from the remote source or simply submit jobs manually.

In case of PanDA jobs execution layer the following scheme was used (Figure 2):

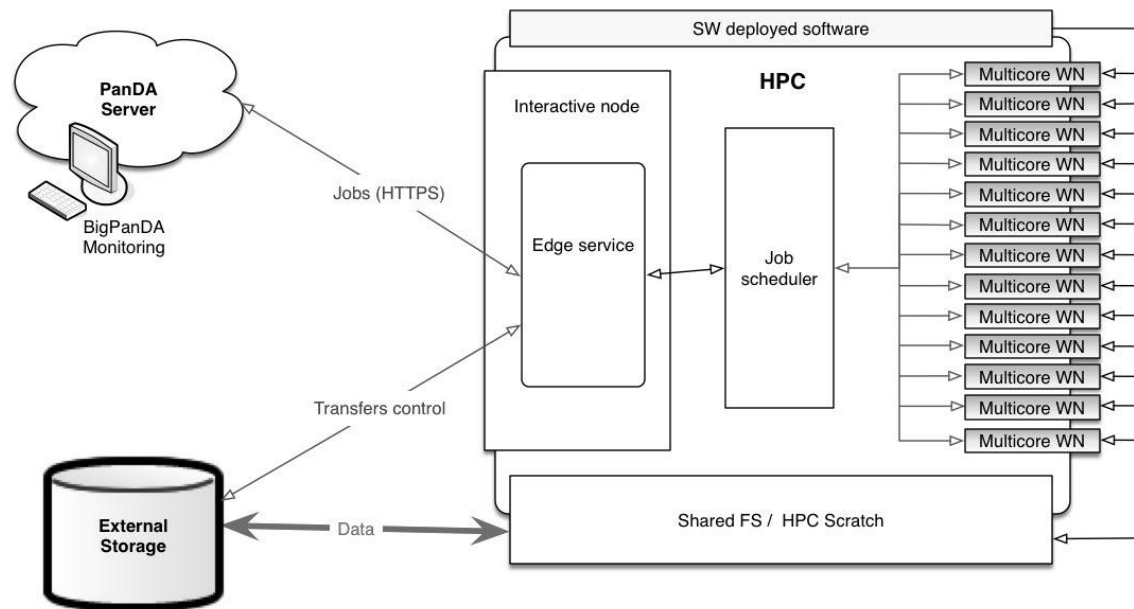


Figure 2. Running jobs via Multi-Job Pilot on HPC

- COMPASS software deployed by Blue Waters production manager in group directory, defined also in task definition interface of production system;
- data is delivered by COMPASS group leader on Blue Waters to group project directory, defined also in task definition interface of production system;
- Python daemon is used in order to organize pilots rotation on the service node (Edge node on the Figure 2). It allows to execute and keep desired amount of pilot processes. It also delivers X509 user proxy;
- PanDA Multi-Job Pilot is used to run COMPASS payloads. Pilot works as management process, which requests and runs desired amount of jobs on HPC. The difference between standard Grid Pilot and Multi-Job Pilot is the following: Grid Pilot is being executed on a Grid node local resource manager while Multi-Job Pilot is also a management software: it prepares, sends submission to a local batch system and provides monitoring. Each Pilot performs one PBS submission. Size of the submission is defined in the configuration, on Blue Waters each Pilot runs 512 jobs on 16 nodes. If no such jobs are available on PanDA server, Pilot submits smaller jobs.

3.4. Calibration database

Each COMPASS reconstruction job reads file from the calibration database. In case of CERN such database runs on a dedicated MySQL node and jobs access it directly via Internet. In case of Blue Waters two new aspects have appeared:

- on the worker nodes Internet connectivity is limited and there is a recommendation from HPC administration not to use Internet connection;
- amount of simultaneously running COMPASS jobs on Blue Waters may reach 200 000, which is more than 10 times more than on resources at CERN and database server simply can not handle such load.

It is clear that in situation with unpredictable load jumps and low throughput, database must be deployed locally on Blue Waters. But even running one or two instances of database can not solve

problem, and a decision was made to run one database on each computing node. This approach has several pro and contra arguments:

- contra:
 - too many database instances — if many submissions occupy 1000 of nodes, 1000 instances of database will be running;
- pro:
 - database is running with one of 32 jobs on the node, first job which comes to the node, starts a database instance. Thus, CPU occupation of the node is 100%, no computing resources being lost;
 - if some job fails because database has not started, only 32 jobs will be lost;
 - no communication between jobs, running on different nodes, is needed. Since database runs on a local node, each job simply connects to a localhost;
 - PanDA Pilot takes care of the running processes and, when all jobs are finished, before leaving the computing node, it removes all zombie and any other running process at the node.

So, each first job on the node, before executing COMPASS payload, starts a database instance. Such approach, in combination of submission size, which is 512 jobs on 16 nodes, allows to solve problem of calibration database overload by too high amount of clients — amount of instances is increasing with amount of running jobs and access to the database is limited by jobs, running on one single computing node.

3.5. Submission tuning

On HPC resources expected time of job execution must be defined before job is submitted. Moreover, execution time must be defined to a bunch of jobs, in COMPASS case to 512 jobs. It is also very important to prepare submission of jobs with as much as possible same expected execution time. It allows to save resources and guarantee uniform load of CPUs on computing nodes. In COMPASS Production system amount of events in raw file is available and used to select and submit jobs, ordered by number of events.

On HPC resource jobs, after being submitted, have to wait in the queue. Each job has a priority which is calculated basing on many parameters: load of the machine, size of submission, priority of project and user, requested execution time and requested queue, etc. In order to make jobs start faster, several techniques, described in the following paragraphs, are being used.

Being an usual Grid jobs, COMPASS jobs are independent and may be run on one multicore machine without interfering each other. Since each single job is not communicating with others while running, no communication is done between nodes either. There is an option (`flags=commtransparent`) which may be set to identify a set of independently running jobs, it allows to start submission faster.

All steps of processing are being performed on Blue Waters: reconstruction, merging of reconstruction job results, merging of histograms and merging of event dumps (filtered events are being selected and stored in event dumps for fast analysis). Each processing type jobs take different time to run. The longest ones are reconstruction jobs, they run up to 18 hours. The fastest are event dumps merging, they usually run 30 minutes. Merging of histograms takes up to 1 hour. OnHPC resource shorter submission usually starts faster then a longer one. In order to optimize submissions, a combination of logical queues are being used in PanDA and Blue Waters: via long queue in PanDA reconstruction jobs are being submitted to normal queue on Blue Waters; via short all merging jobs are being submitted to short queue on Blue Waters with different requested execution time. Using these methods allow jobs to spend less time in the queue.

4. Conclusion

Software services, developed to process data of experiments on Large Hadron Collider, are turning from unique systems to the software products. Projects, such as PanDA, Rucio, AGIS, which were initially developed in the interest of one collaboration, now are used in various areas, not only in HEP experiments, helping to organize heterogeneous computing environment. One of such products, PanDA, is used to run processing of COMPASS experiment data at CERN since 2017, and now is prepared to run jobs on Blue Waters HPC.

COMPASS Production System was designed to run jobs predominantly at CERN, where COMPASS has an allocation, usually limiting amount of simultaneously running jobs to 15 000. When Condor computing element is not overloaded by other experiments, amount of running jobs may reach 20 000. At the same time, there are more than 20 000 computing nodes available on Blue Waters, each of them has 32 CPU. Target amount of jobs to be executed on Blue Waters is 150 000, which is almost 5 000 of occupied nodes. Processing during August 2018 has shown, that setup with one single PanDA server and Multi-Job Pilot on Blue Waters and standard Pilot on CERN Condor may handle 50 000 of simultaneously running jobs under management of one Production System. Such load was achieved on a setup with PanDA database, server and Auto Pilot Factory, running together at the same physical machine, deployed at JINR Cloud Service [12]. But, to reach the target reliably, computing infrastructure has to be changed: PanDA database, Server and Auto Pilot Factory, must run on a separate nodes. In case of very high load, more PanDA servers can be added. Such setup is flexible and reliable enough to handle target amount of running jobs. Migration to PanDA Harvester is also considered.

References

- [1] 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.
- [2] Blue Waters overview. – <https://bluewaters.ncsa.illinois.edu/blue-waters-overview>
- [3] Petrosyan A.Sh. PanDA for COMPASS at JINR // Physics of Particles and Nuclei Letters. – 2016. – Vol. 13, Issue 5. – P. 708-710. – <https://link.springer.com/article/10.1134/S1547477116050393>
- [4] Petrosyan A.Sh., Zemlyanichkina E.V. PanDA for COMPASS: processing data via Grid // CEUR Workshop Proceedings, Vol. 1787. – P. 385-388. – <http://ceur-ws.org/Vol-1787/385-388-paper-67.pdf>
- [5] Petrosyan A.Sh. COMPASS Grid Production System // CEUR Workshop Proceedings, Vol. 2023. – P. 234-238. – <http://ceur-ws.org/Vol-2023/234-238-paper-37.pdf>
- [6] Cray. – <https://www.cray.com/>
- [7] Globus Online. – <https://www.globus.org/>
- [8] Adaptive Computing. – <http://www.adaptivecomputing.com/>
- [9] 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/>
- [10] K. De, A. Klimentov, D. Oleynik, S. Panitkin, A. Petrosyan, J. Schovancova, A. Vaniachine, T. Wenaus on behalf of the ATLAS Collaboration. Integration of PanDA workload management system with Titan supercomputer at OLCF // Journal of Physics Conference Series – 2015. – Vol. 664. – <http://iopscience.iop.org/article/10.1088/1742-6596/664/9/092020>
- [11] 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/>
- [12] 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.