DISTRIBUTED DATA PROCESSING OF THE COMPASS EXPERIMENT

A.Sh. Petrosyan[1,2], D.M. Malevanniy[3]

[1] *Joint Institute for Nuclear Research, 6 Joliot-Curie st., 141980, Dubna, Russia*

[2] *Plekhanov Russian University of Economics, 36 Stremyanny per., 117997, Moscow, Russia*

[3] *Saint Petersburg State University, 7/9 University emb., 199034, Saint Petersburg, Russia*

E-mail: artem.petrosyan@jinr.ru

The implementation of COMPASS data processing in the distributed environment started in 2015. Since the summer of 2017, the data processing system has been working in a production mode, distributing jobs to two traditional Grid sites: CERN and JINR. There are two storage elements, both at CERN: EOS disk storage for short-term storage and Castor tape storage for long-term storage. Processing management services, including the MySQL server, PanDA servers, the APF/Harvester server, a monitoring server, and a production management server, are deployed in the JINR Cloud Service. Thus, the system, which manages distributed data processing of the experiment, is also distributed. The production management system is based on the principles of a service-oriented architecture. Each service of the system is maximally isolated from the others, executed independently and usually performs only one function, for example: sends jobs, checks their statuses, archives results, and so on. During the last year, the system was replenished by a task archiving mechanism, FTS and Harvester services, and a Monte-Carlo processing chain. New HPC machines were also integrated. This article highlights the status, statistics, workflow, data management, infrastructure overview, and future plans.

Keywords: distributed computing, workflow management system, Grid, HPC

Artem Petrosyan, Daniil Malevanniy

1. System overview

A concept of the data processing system in the Grid environment for COMPASS [1] via the PanDA workload management system was presented in 2015 [2]. The first system prototype was prepared in 2016. In 2017, to provide a maximum level of automation for task and job processing, a dedicated workflow management system was developed [3]. Since August 2017, the system has been working in a production mode. During this period, the system was transformed several times. The main reason for the transformation with an upgrade was the need to provide better reliability. Several data processing chains were implemented: real data reconstruction, event filtering [4]. A Monte-Carlo chain was implemented in 2019. It covers MC simulation and reconstruction tasks. In addition, every time the appearance of a new HPC machine, leading to an increase in the load on the system components, has triggered an upgrade of the system. Two Grid sites are involved in data processing on an ongoing basis: CERN and JINR. Volunteer HPC processing sites: BlueWaters (2017-2018), Stampede 2 (2019), Frontera (2019-present time). The current system architecture and workflow are presented in Figure 1.

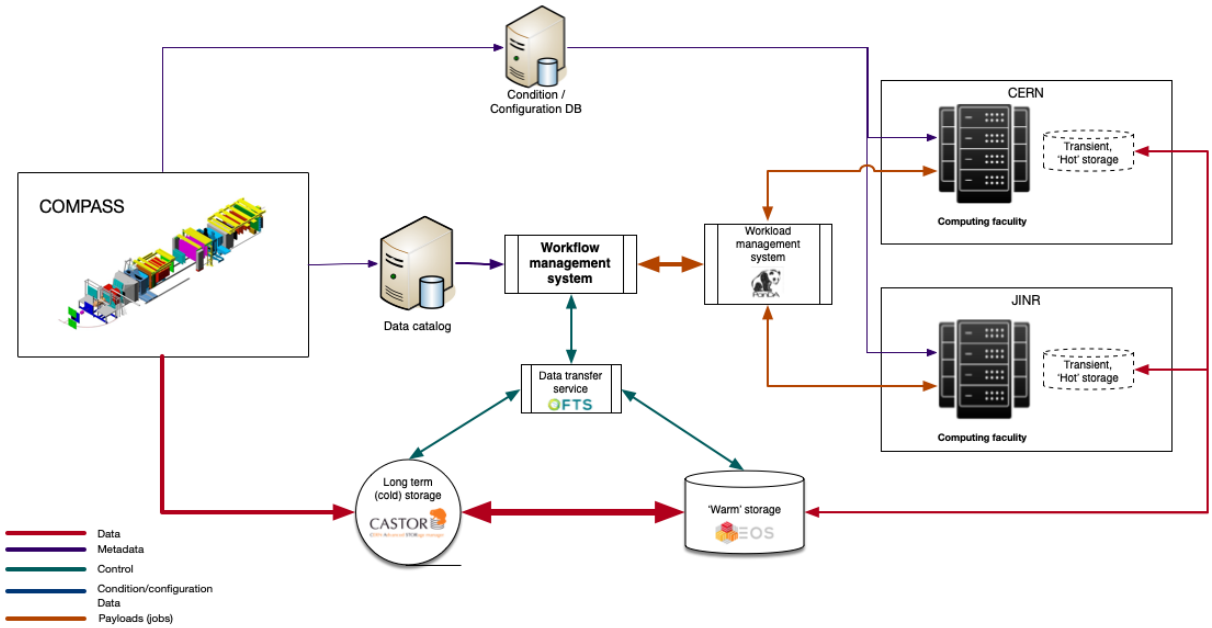

Figure 1. Components of the distributed data processing system

During the last two years, more than 9 million jobs, grouped into more than 400 tasks, were processed by the system. The average wall time of a job in the system is 7.5 hours. Thus, in total, approximately 7.5 CPU years were consumed by jobs managed by the system. The processing rate can reach 100K jobs per day (Fig. 2).

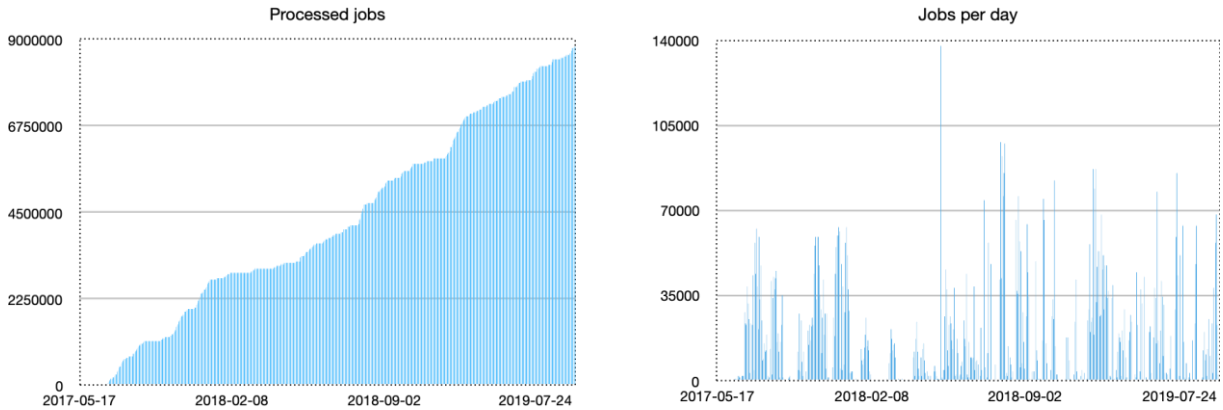

Figure 2. All processed jobs and jobs per day

The system covers all steps of data pre- and post-processing, including archiving to the CERN tape storage: Castor. During two years of operation, more than 700TB of final data were written to Castor. In 2018, support of CERN FTS was implemented to enable asynchronous file transfers between EOS and Castor and reduce the files migration time.

2. Processing on Stampede 2 and Frontera HPCs

Processing on a supercomputer has many significant differences in comparison with working on a regular Grid site and involves a high degree of detail. Running a large number of parallel tasks leads to a big load on the file system. In addition, usually such machines have very strict user policies. COMPASS has a recent experience of using large HPC. During 2018, a prototype as a proof of the concept that the COMPASS production system could run jobs on an HPC machine was developed on BlueWaters HPC of the University of Illinois at Urbana-Champaign [5]. In 2019, an integration of Stampede 2 and Frontera, HPCs of the Texas Advanced Computing Centre [6], was performed. Frontera is one of the most powerful supercomputers in the world; it was at number 5 in the Top 500 in June 2019, Stampede 2 had the number 19 position.

Unlike the scheme used on BlueWaters, with a Multi-Job Pilot as a job management service, a new service named Harvester (Fig. 3) was used as a job management service on the interactive node on HPCs at TACC.

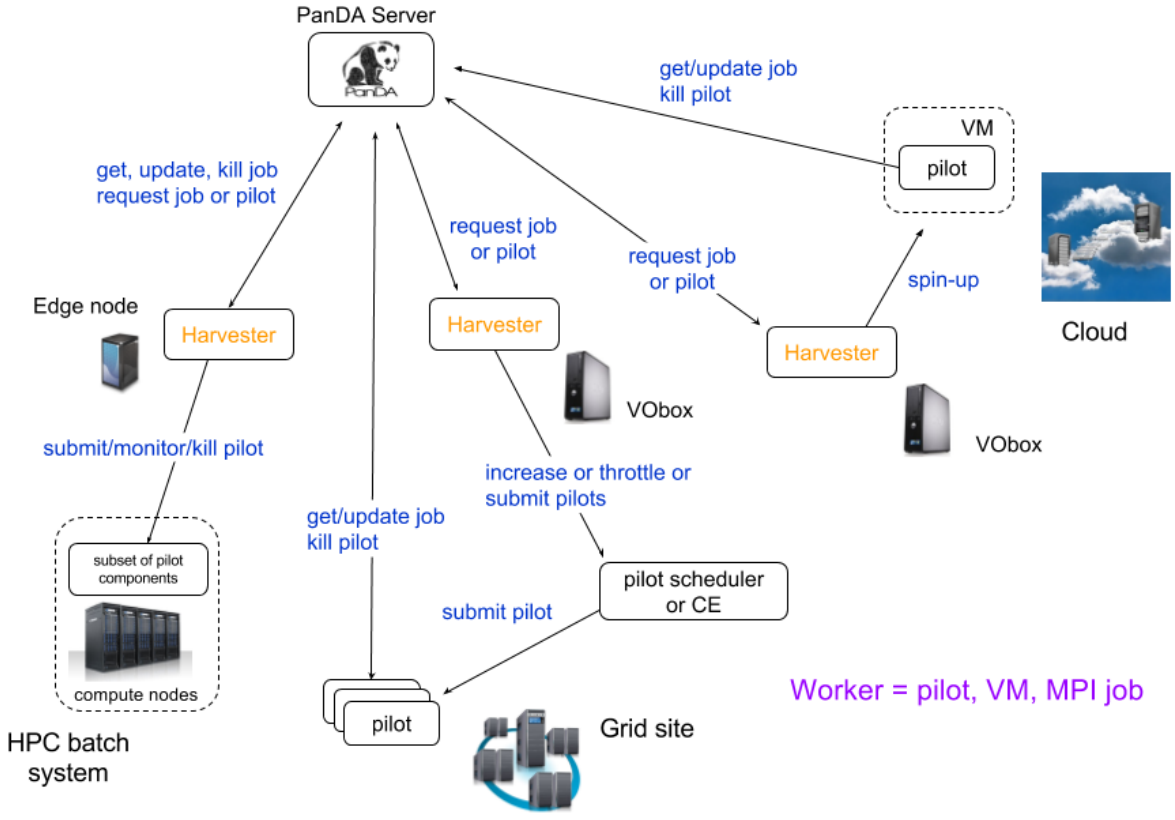

Figure 3. PanDA Harvester workflow

Harvester is a resource-facing service between a workload management system and a collection of pilots [7-8]. It is a lightweight stateless service running on a VO box or an edge node of HPC centres to provide a uniform view for various resources. Harvester was developed, taking into account the ten years' experience of operating the Auto Pilot Factory; it provides much greater configuration flexibility and reliability, as well as monitoring. Harvester was designed so that it could be used to send pilots both to regular Grid sites and HPCs and have an extendable architecture. Everything needed for the COMPASS jobs code, such as local MySQL database execution, payload

management, errors handling and stage-out, was transferred from the Multi-Job Pilot and added as Harvester plug-ins (Fig. 4).

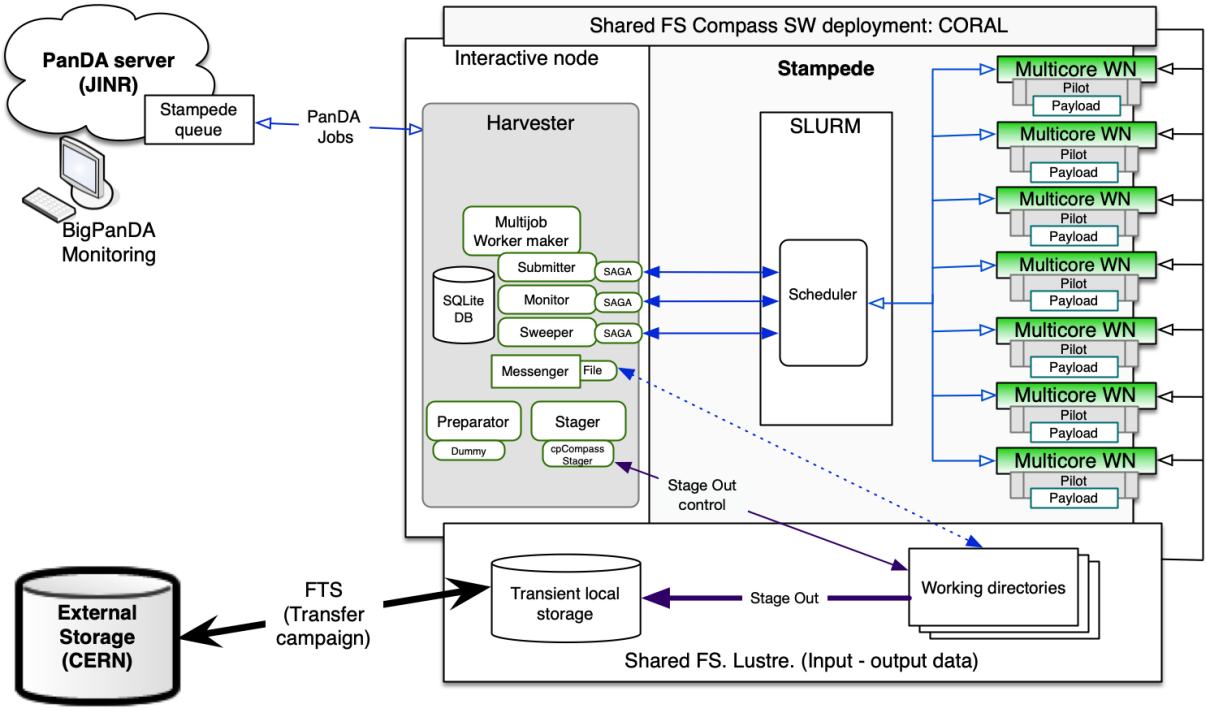

Figure 4. Harvester setup on Stampede 2 and Frontera

A new Harvester-based setup was prepared and tested on Stampede 2, while data processing is carried out on Frontera. Since CPUs on Frontera computing nodes are much more efficient than on Stampede and BlueWaters, much attention has been paid to optimizing data processing in order to reduce the rate of I/O operations. The server side of the setup was also updated to enable processing via Harvester: the latest PanDA server version was installed. Moreover, anticipating a high load on the PanDA server from HPC, an additional PanDA server dedicated exclusively to HPC processing was installed. Thus, there are two PanDA servers in the production setup at present: for Grid and HPC.

3. Summary

The production system continues to provide a reliable platform for handling all types of tasks and is the main tool for distributed processing of data gathered by the physics facility.

In a relatively small experiment as COMPASS, i.e. in the situation with limited resources, it is highly important to rely on central services with proven characteristics, even if they are redundant at first sight. Such services, initially developed for the ATLAS experiment [9], have demonstrated the expected level of scalability and reliability and allowed to use computing resources of different types, including modern HPC facilities.

We see that software systems, initially designed for the needs of one collaboration, after a decade of intensive operation and improvement, have turned into software products that can be used in other experiments, as well as outside the field of high-energy physics for organizing distributed computing.

The development and utilization of the COMPASS distributed processing management system has made it possible to formulate requirements for the components of the Multifunctional Informational and Computing Complex (MICC) [10] in the Laboratory of Information Technologies at the Joint Institute for Nuclear Research on the eve of the start of work on the construction of experimental data processing systems on the Nuclotron-based Ion Collider Facility NICA [11]. There are already ongoing efforts of FTS service deployment, development of unified authentication and authorization services, establishment of local certification authority, etc.

References

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

[2] A.Sh. Petrosyan, PanDA for COMPASS at JINR, Physics of Particles and Nuclei Letters, Vol. 13, No. 5, pp. 708-710, 2016

[3] A. Petrosyan, COMPASS Grid Production System, CEUR Workshop Proceedings, Vol. 2023, pp. 234-238, 2017

[4] A. Petrosyan, COMPASS Production System Overview, EPJ Web Conf., Vol. 214, 2019

[5] A.Sh. Petrosyan, COMPASS production system: processing on HPC, CEUR Workshop Proceedings, Vol. 2268, pp. 139-144, 2018

[6] Texas Advanced Computing Centre, the University of Texas at Austin, available at https://www.tacc.utexas.edu/ (accessed 31.10.2019)

[7] A. Anisenkov, D. Drizhuk, W. Guan, M. Lassnig, P. Nilsson, D. Oleynik on behalf of the ATLAS Collaboration, Global heterogeneous resource harvesting: the next-generation PanDA Pilot for ATLAS, Journal of Physics Conference Series, Vol. 1085, No. 032031, 2018

[8] Harvester web home, available at https://github.com/HSF/harvester (accessed 31.10.2019)

[9] ATLAS collaboration web home, available at https://atlas.cern/ (accessed 31.10.2019)

[10] MICC web home, available at https://micc.jinr.ru/ (accessed 31.10.2019)

[11] NICA web home, available at http://nica.jinr.ru/ (accessed 31.10.2019)