APPLICATION OF THE BIGPANDA WORKLOAD MANAGEMENT SYSTEM TO SUPPORT DISTRIBUTED SUPERCOMPUTER-BASED COMPUTING IN NEUROSCIENCE

K. De[1], A. Klimentov[2,3], R. Mashinistov[1], A. Novikov[3]
[1]University of Texas at Arlington
[2]Brookhaven National Laboratory
[3]National Research Center "Kurchatov Institute"
ruslan.mashinistov@cern.ch

Introduction

PanDA [1] was originally designed specifically for the needs of the ATLAS [2] Experiment at the Large Hadron Collider (LHC) [3], and has proved to be highly successful in meeting all the distributed computing needs of the experiment. The core design of PanDA is however not experiment specific. PanDA is capable of meeting the needs of other data intensive scientific applications, and the BigPanDA project [4] was created to undertake this. The project has generalized PanDA for use by other experiments and extended its reach to High Performance Computing (HPC) platforms.

In February 2017, a pilot project was started between BigPanDA and the Blue Brain Project (BBP)[5] of the Ecole Polytechnique Federal de Lausanne (EPFL) located in Lausanne, Switzerland. BBP has complex scientific workflow which relies on using a mix cluster and supercomputers to reconstruct and simulate accurate models of brain tissue. BBP has own clusters and currently extending available resources to use also the Titan Supercomputer [6] operated by OLCF and cloud based resources (Amazon). But it still lacks central system to manage such complex workflow and all distributed resources at once.

This project was aimed to demonstrate the efficient application of the BigPanDA system to support the complex scientific workflow of the BBP and all target systems. The first "proof of concept" phase of the project was lasted for 6 months and successfully finished in September 2017. Within this period the new BigPanDA software instance was installed in Geneva, adopted and configured in order to support BBP workflow. To meet the specific needs of the BBP workflow additional components were deployed and adopted. The following resources were considered for demonstration:

• Intel x86-NVIDIA GPU based BBP clusters located in Geneva (47 TFlops) and Lugano (81 TFlops),

• BBP IBM BlueGene/Q supercomputer (0.78 PFLops) located in Lugano,

• Titan Supercomputer with peak theoretical performance 27 PFlops operated by the Oak Ridge Leadership Computing Facility (OLCF),

• Cloud based resources such as Amazon Cloud.

General approach based on the BigPanDA WMS

The BigPanDA as a basis technology delivers transparency of data and it's processing in a distributed computing environment to the scientists. It provides execution environments for a wide range of experimental applications, automates centralized data processing, enables data analytics for dozens of research groups, supports custom workflow of individual scientists, provides a unified view of distributed worldwide resources, presents status and history of workflow through an integrated monitoring system, archives and curates all workflow, manages distribution of data as needed for processing or scientists access, and provides other features. The rich menu of features provided, coupled with support for heterogeneous computing environments, makes BigPanDA ideally suited for modern scientific data processing. The Portal integrated the components, which don`t belongs to the BigPanDA system: Web-interface, Data Storage and Data Management System. Some of these components were developed and adopted to meet BBP researchers needs.

The Portal includes the following main components:

• Server. The PanDA server is the heart of the system factorized as a general WMS service. The main components of the server are:

– Database. A system-wide job database, which records comprehensive static and dynamic information on all jobs and resources in the system.

– Brokerage. An intelligent module operates to prioritize and assign jobs to resources on the basis of job type, priority, software and data availability, real time job statistics, and available CPU and storage resources.

– Dispatcher. A component in the PanDA server which receives requests for jobs from pilots and dispatches job payloads.

• Pilots [8]. One of the key features of PanDA approach is to use pilot jobs. Pilot jobs are used for acquisition of processing resources in advance. Jobs are assigned to successfully activated and validated pilots by the PanDA server based on brokerage criteria. This 'late binding' of workload jobs to processing slots prevents latencies and failure modes in slot acquisition from impacting the jobs, and maximizes the flexibility of job allocation to resources based on the dynamic status of processing facilities and job priorities. The pilot is also a principal 'insulation layer' for PanDA, encapsulating the complex heterogeneous environments and interfaces of the grids and facilities on which PanDA operates.

• Web interface and custom API allows users to define the jobs in the system and perform monitoring functions.

• Data Management System allows uniform access to data on distributed storages.

Software components of the BigPanDA based portal for BBP applications

For more efficient use of the portal we have developed an interface to define and run custom computing jobs and monitor their status, control the workflow. This interface consists of several software modules. The portal authenticates users with a given username and password. Direct interaction with the user is provided by the unified web form to define new custom jobs. Throw the form user creates a description of the job and submits it to the server. After that the status of the job can be monitored with the built-in monitoring web interface. The job processing workflow is transparent. The specific settings of the running jobs and some technical spices of server are hidden from the end users. All required actions on jobs also could be done dynamically using HTTP requests to the portal API.

The main components of the portal are:

• GUI/Web services - graphical user interface and web service that provides authentication and simple web interface to make jobs definition. Also support of the API provided at this level.

• PanDA Client is the simple python client for command line interface by which users define job, with all parameters and the input/output files. Then it registers the job at the PanDA server.

• PanDA Server - local Panda server installed and configured on the BBP VM. Server provides an internal API to interact with PanDA Client and Monitor.

• PanDA Monitor provides the overall information about the status of the system, submitted jobs and resources.

• Pilot Schedulers - pilot schedulers manages the pilots submission to available resources defining how many pilots to run on each resource.

• Pilots –several adaptations of pilots were made to support variety computing environments, a new feature was added with initial tests.

• Data Management System (DMS) - lightweight experiment data management tool. It consists of general file catalog to store metadata and distributed file transfer system to move data between heterogeneous data storages. At the moment the development and integration of the DMS is in progress. Globus Online is considered as the Data Transfer System for the portal.

Conclusions

As a proof of concept, the pilot project was aimed to demonstrate efficient application of the PanDA software for the supercomputer-based reconstructions and simulations, offering a radically new approach for understanding the multilevel structure and function of the brain (BBP project).

In the first phase, the goal was to support the execution of BBP software on a variety of distributed computing systems powered by PanDA. The targeted systems for demonstration include: Intel x86-NVIDIA GPU based BBP clusters, BBP IBM BlueGene/Q supercomputer, the Titan Supercomputer operated by the Oak Ridge Leadership Computing Facility (OLCF), and Amazon Cloud.

The project demonstrated that the software tools and methods for processing large volumes of experimental data, which have been developed initially for experiments at the LHC accelerator, can be successfully applied to other scientific fields. Throw the deployed BigPanDA portal an MPI test jobs have been successfully submitted and executed on the BBP distributed resources.

Acknowledgements

We wish to thank our colleagues from the Blue Brain Project especially Prof. Felix Schuermann, Dr. Fabien Delalondre and Alexandre Beche.

This work was funded in part by the U.S. Department of Energy, Office of Science, High Energy Physics and Advanced Scientific Computing Research under Contracts DE-SC0008635, DE-SC0016280; Russian Ministry of Science and Education under Contract no 14.Z50.31.0024; and by Blue Brain Project. We would like to acknowledge that this research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract no. DE-AC05-00OR22725.

References

1. Maeno T., Overview of ATLAS PanDA workload management // J. Phys.: Conf.Ser. – 2011 - vol. 331, 072024.
2. ATLAS Collaboration, The ATLAS Experiment at the CERN Large Hadron Collider // J. Inst. – 2008 - vol. 3, S08003.
3. Evans L., Bryant P., LHC machine // Journal of Instrumentation – 2008 - vol. 3, S08001.
4. Klimentov A. et al. Next Generation Workload Management System for Big Data on Heterogeneous Distributed Computing // J. Phys. Conf. Ser. – 2015 - vol. 608(1):012040.
5. Markram H., The blue brain project // Nat. Rev. Neurosci. - 2006 - vol. 7 - pp.153-160.
6. De K. et al. Integration of PanDA workload management system with Titan supercomputer at OLCF // J. Phys. Conf. Ser. – 2015 - vol. 664:092020.
7. P. Nilsson et al., Experience from a Pilot based system for ATLAS // J.Phys.Conf.Ser. – 2008 - vol. 119:062038.