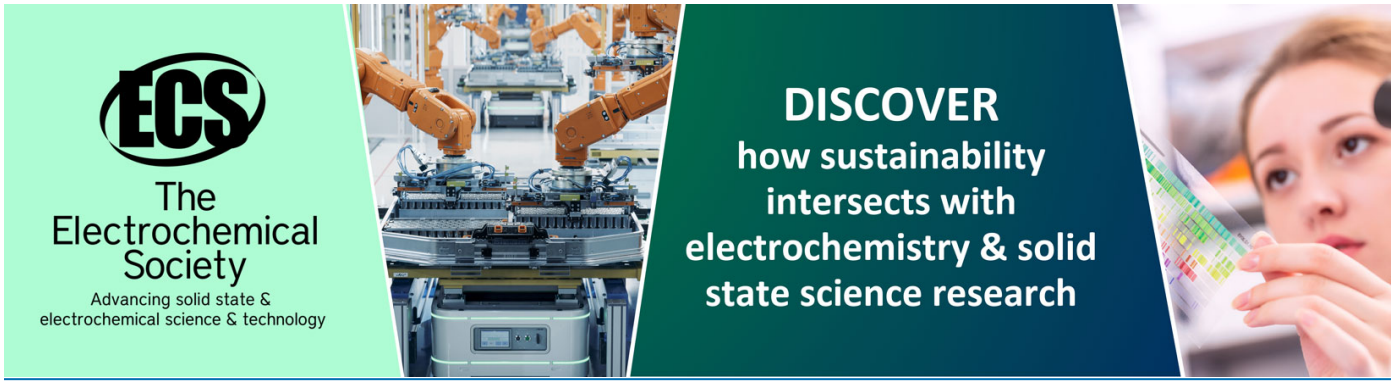
Supercomputers, Clouds and Grids powered by BigPanDA for Brain studies



IOP Conf. Series: Journal of Physics: Conf. Series **1085** (2018) 032003 doi:10.1088/1742-6596/1085/3/032003

Supercomputers, Clouds and Grids powered by BigPanDA for Brain studies

A Beche[1], K De[2], F Delalondre[1], F Schuermann[1], A Klimentov[3] and R Mashinistov[2,a]

[1]Swiss Federal Institute of Technology in Lausanne, Route Cantonale, 1015 Lausanne, Switzerland
[2]University of Texas at Arlington, 701 South Nedderman Drive, Arlington, TX 76019, US
[3]Brookhaven National Laboratory, P.O. Box 5000, Upton, NY 11973-5000, US

[a]rmashinistov@gmail.com

Abstract. The PanDA WMS - Production and Distributed Analysis Workload Management System - has been developed and used by the ATLAS experiment at the LHC (Large Hadron Collider) for all data processing and analysis challenges. BigPanDA is an extension of the PanDA WMS to run ATLAS and non-ATLAS applications on Leadership Class Facilities and supercomputers, as well as traditional grid and cloud resources. The success of the BigPanDA project has drawn attention from other compute intensive sciences such as biology. In 2017, a pilot project was started between BigPanDA and the Blue Brain Project (BBP) of the Ecole Polytechnique Federal de Lausanne (EPFL) located in Lausanne, Switzerland. This proof of concept project is aimed at demonstrating the efficient application of the BigPanDA system to support the complex scientific workflow of the BBP, which relies on using a mix of desktop, cluster and supercomputers to reconstruct and simulate accurate models of brain tissue.

1. Introduction

The Production and Distributed Analysis (PanDA) system [1] was designed to meet ATLAS [2] requirements for a data-driven workload management system for production and distributed analysis processing capable of operating at the LHC data processing scale. It has been used in the ATLAS experiment since 2005 and is now expanding into a BigPanDA [3] project to extend PanDA as a meta application, providing location transparency of processing and data management, for High Energy Physics (HEP) and other data-intensive sciences, and a wider exascale community. The success of the projects to use BigPanDA beyond HEP and traditional HTC resources has drawn attention from other compute intensive sciences such as neuroscience.

In 2017, a pilot project was started between BigPanDA and the Blue Brain Project (BBP) [4] of the Ecole Polytechnique Federal de Lausanne (EPFL) located in Lausanne, Switzerland. This proof of concept project is aimed to demonstrate efficient application of the PanDA Workload Management System, initially developed for HEP applications, for the supercomputer-based reconstructions and simulations offering a radically new approach for understanding the multilevel structure and function of the brain.

In the first phase, the goal of this joint project is to support the execution of BBP software on a variety of distributed computing systems powered by BigPanDA. The targeted systems for demonstration are: x86 BBP clusters located in Geneva and Lugano (Switzerland), the Titan Supercomputer [5] operated by Oak Ridge Leadership Computing Facility (OLCF) and Cloud based resources (Amazon).

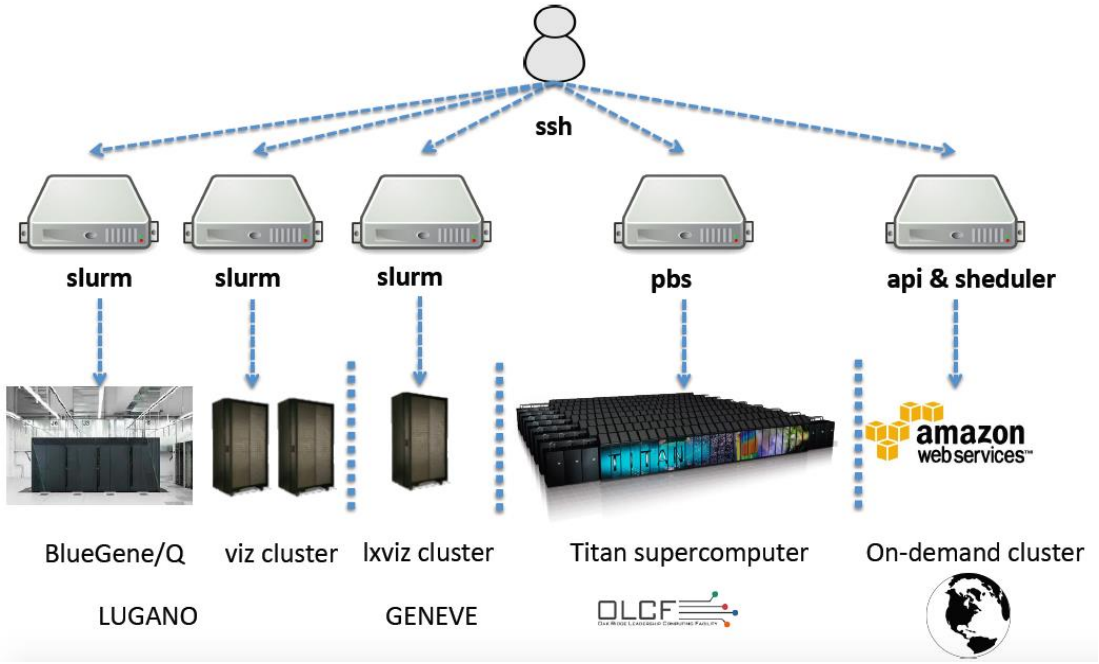


Figure 1. Variety of BBP distributed computing systems considered in the project.

2. Evolution of the computing model of the Blue Brain Project

Brain simulation software applications are highly parallelized using MPI, thus, they should run on HPC clusters with high throughput and low latency interconnect (InfiniBand). The main challenge in the HPC world is that the software stack (scheduler, MPI libraries etc.) is deeply system dependent and portability very difficult to achieve. In order to accelerate science, BBP workflow is no longer exclusively running on dedicated resources but also on world-class supercomputers in Leadership Computing Facilities such as Oak-Ridge National Laboratory (ORNL) - Titan or Argonne National Laboratory (ANL) - MIRA [6]. This heterogeneity in the available systems imposed to streamline the workflow and dataflow in order to lower the entry barrier to various system for scientists. This simplification has began with the software deployment using the NIX[7] package manager allowing an immutable system-agnostic deployment of the whole BBP software stack. The second stone towards simplification is workload scheduling. The current workflow requires every scientist to connect to a remote login node and schedule jobs from there using the local system resource manager (Slurm on BBP systems, Moab/PBS on Titan, Cobalt on MIRA etc.) complexifying the learning curve to run the exact same software. BigPanDA is approaching the problem by confining the scheduling specificities into the pilot job hiding it completely from the scientists and allowing them to run exact same jobs independently of the desired HPC system. Finally the last step will be to standardized on data movement, more precisely to abstract the stage-in stage-out of the data to the job phases to the scientists. While there is strong signal in favor of using Globus GridFTP [8] for achieving it, this will deserve a non-negligible part of the BigPanDA/BBP proof of concept to validate it and perhaps move it towards production one day. To summarize, BBP computing model is moving from system-dependent implementation to system-agnostic workflow with well-defined building blocks to achieve

simplicity: NIX for software deployment, BigPanDA for Job submission and investigating GridFTP for data transfers.

3. BigPanDA based portal for BBP applications

3.1. Accelerating science impact with PanDA WMS

PanDA is a Workload Management System (WMS) designed to support the execution of distributed workloads and workflows via pilots [9]. Pilot-capable WMS enable high throughput of jobs execution via multi-level scheduling while supporting interoperability across multiple sites. PanDA as a basis technology delivers transparency of data and its processing in a distributed computing environment to the scientists [10]. 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 PanDA ideally suited for scientific data processing. HEP and astro-particle experiments COMPASS and AMS has chosen PanDA as WMS for data processing and analysis. Experiments nEDM with LSST will evaluate PanDA and ALICE is interested in PanDA evaluation for OLCF. Joint Institute for Nuclear Researches (Russia) is considering PanDA as main WMS for NICA collider. Also another several PanDA instances deployed beyond ATLAS at OLCF, Taiwan, Amazon EC2, Russia.

3.2. Software components of the BigPanDA Portal

Two most important components of the PanDA WMS are server and pilots. Jobs are submitted to the PanDA server via simple python-based client application. The PanDA server is the main component, which provides a task queue managing all job information centrally. The PanDA server receives jobs into the task queue, upon which a brokerage module operates to prioritize and assign work on the basis of job type, priority, software availability, input data and its locality, and available computing resources. Pilots launched on the supercomputer or cluster retrieve jobs from the PanDA server in order to run them via the native resource manager like pbs, slurm or others. Pilots use resources efficiently; they exit immediately if no job is available and the submission rate is regulated according to workload. Each pilot executes a job, detects zombie processes, reports job status to the PanDA server, and recovers failed jobs.

To hide execution complexity and simplify manual tasks by end-users, we an interface providing the reduced intuitive feature set like computing tasks definition and submission for execution on the chosen resource, monitoring the states of the system, jobs and resources.

Authentication/authorization module of the Portal integrated with BBP SSO and LDAP groups aware. Unified web form to define new custom tasks allows users to select the applications from the predefined list, provide the input/output files names together with additional parameters for execution. Then defined task can be submitted for execution on specific resource linked to the predefined queue at the PanDA server. After that the status of the tasks can be monitored with the built-in monitoring web interface. Also the portal provides the REST API streamlining the bulk tasks submission.

We have created the prototype of the data management system (DMS) that allows to connect to the storages of different types. The system supports the file replication mechanism and ensures consistency of the replicas. All files involved in the computing, their replicas and access rules are described in a special file catalog. Together distributed data management system and file catalog allows quick integration of third-party storage systems. The DMS system being a part of the portal not yet integrated with the BBP environment and workflow. Wide variety of the used computing resources requires the core data movement and storage technologies to be chosen. This strategic choice will be

strictly followed later while development of the DMS plugins and integration with storage technologies. Currently the Globus - a secure, unified interface to research the data - is considered as underlying technology for DMS. Globus was built by the computer scientists at the University of Chicago and Argonne National Laboratory to meet the needs and requirements of the research community. It presents a secure, unified interface to identities and storage across Globus-connected sites, within the visibility and access control limits set by each site. The portal scheme is shown on figure 2.

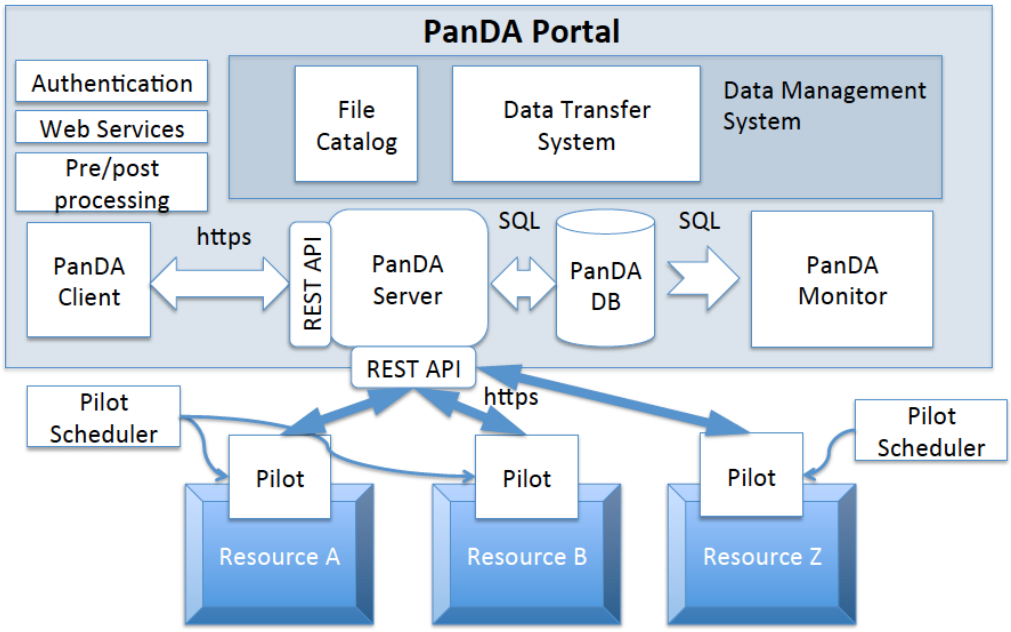


Figure 2. Software components of the PanDA Portal

The main components are:

- GUI/Web services - graphical user interface and web service that provides authentication and simple web interface to make a jobs definition. Also support of the API provided at this level.
- Pre/post processing tool - software component that support the pipelines - several sequenced computational steps. The tool translates the user task definition to the sets of standard PanDA jobs corresponding the pipeline steps and submit them to the PanDA server via the client. Jobs belonging to the same step can run in parallel while jobs of different steps are running by defined order.
- PanDA Server - local Panda server installed and configured on the BBP VM. The server is the heart of the system factorized as a general WMS service. Server provides mapping of all jobs in the system to all available resources. 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.
- The pilot launched on resource retrieves jobs from the server and handles payload execution. Resource specific schedulers are responsible to launch the pilots and maintain appropriate number of running pilots.
- Pilot Schedulers - pilot schedulers manages the pilots submission to available resources defining how many pilots to run on each resource.

- Data Management System - 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 an integration of the DMS is in progress.

The following resources were integrated for the Portal at BBP: Intel x86-NVIDIA GPU based BBP clusters located in Geneva (47 TFlops) and Lugano (81 TFlops); BBP IBM BlueGene/Q supercomputer (0.78 PFLops and 65 TB of DRAM memory) [10] located in Lugano; the Titan Supercomputer with peak theoretical performance 27 PFlops operated by the Oak Ridge Leadership Computing Facility (OLCF); Cloud based resources such as Amazon Cloud.

4. Adaptation of the PanDA pilots

Significant role within supported computing resources for this project is played by the Titan supercomputer. This is one of the largest supercomputers in the world. Titan is a hybrid-architecture Cray XK7 system with a theoretical peak performance exceeding 27 petaflops. Titan features 18,688 compute nodes, (each with one 16-core AMD Opteron CPU and 1 NVIDIA Kepler K20X GPU), 299,008 x86 cores, a total system memory of 710 terabytes, and a high-performance proprietary network. The combination of these technologies allows Titan to achieve up to 10 times the speed of its predecessor, the Jaguar supercomputer, while consuming the same average power load and occupying the same physical footprint. The BigPanDA project provided the first important demonstration of the capabilities that a workload management system (WMS) can have on improving the uptake and utilization of LCF from both application and systems points of view. This demonstration was done for ATLAS and now same approach is being applied with the variety of diverse scientific applications even outside of HEP area.

On Titan the pilots run on the front-end nodes, which allows them to communicate with the PanDA server, since front end nodes have connectivity to the Internet. Worker nodes use Cray's Gemini interconnect for inter-node MPI messaging but have no network connection to the outside world. The interactive front-end machines and the worker nodes use a shared file system which makes it possible for the pilots to stage-in input files that are required by the payload and stage-out the produced output files at the end of the job. The pilots submit payloads to the worker nodes using the local batch system (PBS) via the SAGA (Simple API for Grid Applications) interface [11]. It also uses the SAGA facilities for monitoring and management of PanDA jobs running on Titan's worker nodes.

All of the target systems for demonstration except Titan supercomputer use SLURM as local resource management system. In this case we use a wrapper script that will launch pilot jobs in adopted environment at clusters and supercomputers.

A payload also can be executed in cloud environment, which will need special pilot adaptation. In this case pilots launching system must be aware of cloud environment and have permissions and other means to request new virtual machines (VMs) and destroy unnecessarily. Some extra configuration of local resource management system will also be required to deal with VMs. At the moment the adaptation of the pilots for this case is in progress.

For specific cases users also need to strictly separate working environment of their workflows and usage of their own quota at clusters. That also imply slight modification of pilot jobs. Simple yet powerful approach can use wrapper that will switch user context using SUID bit in Linux operation systems. That wrapper can also do other cluster specific actions with user environment or administrative tasks. We have successfully conducted initial tests for this case.

5. Conclusion

The first phase of the pilot project lasted 6 months with the goal to support the execution of BBP software on a variety of distributed computing systems powered by BigPanDA. The targeted systems chosen for demonstration for demonstration shown on figure 3 and included: Intel x86-NVIDIA GPU based BBP clusters located in Geneva and Lugano, BBP IBM BlueGene/Q supercomputer [12] located

in Lugano, 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.

Next "preproduction" phase of the project is under investigation currently. The next goal is to implement the Data Management System for considered scope of the systems and seamlessly integrate it into the BigPanDA portal. Globus is considered as the basic technology. The main target is to establish PanDA portal as a service with 8/5 support.

Acknowledgement
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 et al 2011 *J. Phys.: Conf. Ser.* vol 331 072024 (Bristol: IOP Publishing) p 6
[2] ATLAS Collaboration, Aad J et al 2008 *J. Inst.* vol 3 S08003 (Bristol: IOP Publishing) p 407
[3] Buncic P et al 2015 *J. Phys.: Conf. Ser.* vol 608 012040 (Bristol: IOP Publishing) p 8
[4] Markram H 2006 *Nat. Rev. Neurosci.* vol 7 (London: Nature Publishing Group) pp 153-160
[5] Titan supercomputer: https://www.olcf.ornl.gov/titan
[6] MIRA supercomputer: https://www.alcf.anl.gov/mira
[7] NIX – The Purely Functional Package Manager: https://nixos.org/nix/
[8] Globus GridFTP: http://toolkit.globus.org/toolkit/docs/latest-stable/gridftp/
[9] Nilsson P et al 2011 *J. Phys.: Conf. Ser.* vol 331 062040 (Bristol: IOP Publishing)
[10] De K et al 2016 *EPJ Web of Conferences* vol 108 01003 (Les Ulis: EDP Sciences) p 10
[11] SAGA progect, "SAGA" [software], version 0.47.2, 2018. Available from https://github.com/radical-cybertools/saga-python [accessed 2018-02-28]
[12] BlueGene/Q SC: https://www-03.ibm.com/systems/technicalcomputing/solutions/bluegene/