EPJ Web of Conferences **214**, 03050 (2019) https://doi.org/10.1051/epjconf/201921403050
*CHEP 2018*

# BigPanDA: PanDA Workload Management System and its Applications beyond ATLAS

*Pavlo* Svirin[1*], *Kaushik* De[2], *Alessandra* Forti[3], *Alexei* Klimentov[1], *Rasmus* Larsen[1], *Peter* Love[4], *Tadashi* Maeno[1], *Ruslan* Mashinistov[1], *Swagato* Mukherjee[1], *Andrei* Nomerotski[1], *Danila* Oleynik[2], *Sergey* Panitkin[1], *Hye Yun* Park[1,5], *Erin* Sheldon[1], *Anze* Slosar[1], *Jack* Wells[6], and *Torre* Wenaus[1]

[1]Brookhaven National Laboratory (BNL), Upton, NY USA
[2]University of Texas at Arlington (UTA), Arlington, TX USA
[3]School of Physics and Astronomy, University of Manchester, Manchester, United Kingdom
[4]Physics Department, Lancaster University, Lancaster, United Kingdom
[5]Stony Brook University (SBU), Stony Brook, NY USA
[6]Oak Ridge National Laboratory (ORNL), Oak Ridge, TN USA

**Abstract.** Modern experiments collect peta-scale volumes of data and utilize vast, geographically distributed computing infrastructure that serves thousands of scientists around the world. Requirements for rapid, near real-time data processing, fast analysis cycles and need to run massive detector simulations to support data analysis pose special premium on efficient use of available computational resources. A sophisticated Workload Management System (WMS) is needed to coordinate the distribution and processing of data and jobs in such environment. The ATLAS experiment at CERN uses PanDA (Production and Data Analysis) Workload Management System for managing the workflow for all data processing on over 150 data centers. While PanDA currently uses more than 250,000 cores with a peak performance of 0.3 petaFLOPS, it runs around 2 million jobs per day on hundreds of Grid sites and serving thousands of ATLAS users. In 2017 about 1.5 exabytes of data were processed with PanDA. In 2012 BigPanDA project project was started with aim to introduce new types of computing resources into ATLAS computing infrastructure, but also to offering PanDA features to different data-intensive applications for projects and experiments outside of ATLAS and High-Energy and Nuclear Physics. In this article we will present accomplishments and discuss possible directions for future work.

## 1 Introduction

Production and Distributed Analysis Workload Management System (PanDA WMS) [1] is the system initially developed for ATLAS experiment [2] and was designed as high-level intellectual layer on WLCG [3] grid-infrastructure.

PanDA WMS allows the efficient use of WLCG infrastructure and provided multiple benefits for running ATLAS payloads (Figure 1). The most significant features provided by PanDA are the following:

*Corresponding author: pavlo.svirin@cern.ch

- A common layer over heterogeneous computing resources which hides the complexity of different computing resources and middlewares from users.
- An implementation of the late-binding principle for the job payload with CPU slots using the Pilot component. This approach allows to prevent latencies and failure modes in slot acquisition from impacting the jobs.
- PanDA's brokerage system takes into account multiple static and dynamic information about jobs and computing resources and distributes the workload efficiently among the available sites.
- A comprehensive monitoring system offering detailed drill-down into job, site and data management information for problem diagnostics.
- Support for running arbitrary user payloads as in conventional batch submission. There is no ATLAS specificity or workload restrictions in the design.

In 2012 the BigPanDA project funded by Advanced Scientific Computing Research (ASCR) program [4] was started with aim to introduce HPC computing resources like Titan [5] operating at the Oak Ridge Leadership Computing Facility (OLCF) into the ATLAS computing model. Another goal of the project was to offer PanDA features to projects and experiments beyond ATLAS and High-Energy Physics (HEP).

Since 2015 the BigPanDA project evolved into BigPanDA++ as a collaboration between Brookhaven National Laboratory (BNL), Oak Ridge National Laboratory (ORNL), University of Texas Arlington (UTA) and Rutgers University.

## 2 PanDA outside ATLAS and CERN: instances at OLCF and EC2, edge services and user environment

### 2.1 PanDA Server instances in EC2 Amazon Cloud and OLCF

ATLAS PanDA Server depends on Oracle database backend, which is a proprietary product. In order to meet licensing requirements, a version of PanDA Server has been developed to

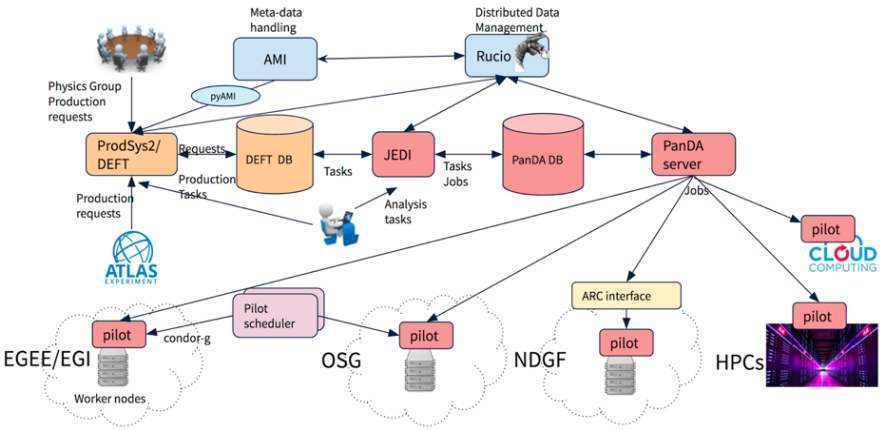


**Figure 1:** The components of ATLAS distributed computing software: production management (amber elements), metadata and distributed data management (blue elements), workload management (red elements)

run with MySQL database, an open-source product published under GPLv2 license. We have ported database description to MySQL excluding ATLAS-specific stored procedures. PanDA Monitor has also been tested for operability with MySQL.

In order to run payloads for non-ATLAS experiments a dedicated instance of MySQL-based PanDA Server has been deployed in Amazon EC2. PanDA Monitor was also configured to display job information for this instance, as well as AutoPilot [6] which submits pilots into Grid environment. As of today, there are 36 computing element endpoints configured with this instance of AutoPilot. This instance also serves HPC resources at BNL, NERSC and Jefferson Lab (JLab) via Harvester (section 2.2) instances deployed on the front nodes of these resources. Currently this instance serves payloads for experiments and projects like LSST/DESC and LQCD mentioned in sections 3.5 and 3.6.

Another PanDA Server instance to serve experiments and projects at OLCF (Figure 2) was deployed as a container-based application in OpenShift cluster management system at OLCF. This approach can be reused for fast deployments in the future. Currently this instance serves experiments and projects like nEDM (section 3.7), Computational Biology (section 3.1), IceCube (section 3.3) and Molecular Dynamics (section 3.2). To isolate the workflows of different groups and experiments, dedicated queues were defined for both PanDA Server instances.

### 2.2 From Pilot to Harvester

Pilot-based architecture [7] allows PanDA to integrate the wide diversity of the computing architectures like Grid, HPCs and Clouds. However, for correct operation pilot requires outbound network connection from worker nodes, which might be absent on worker nodes of HPCs, like Titan. In this case a persistent edge-service has to be used to submit jobs to local batch system.

Harvester [8] is a resource-facing service between PanDA server and collection of workers on a compute site. It is intended to operate as a universal edge service in PanDA, capable of working with different types of computing resources – Grid sites, clouds and supercomputers. Worker is a generalization of a pilot concept and can be, depending on a resource and

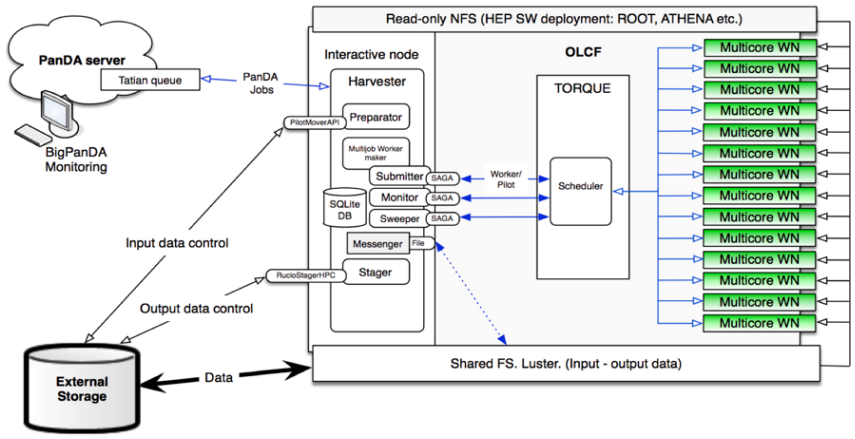


**Figure 2:** Harvester deployment on Titan and its interaction with external PanDA services and Titan's batch system

workflow, a pilot, an MPI job, or a virtual machine. Harvester has modular multi-threaded design to support heterogeneous resources and to accommodate for special workflow requirements.

Harvester provides for flexible scheduling of job execution and asynchronous data transfer to and from the controlled resource.

We successfully used Harvester for integration of resources for various projects inside and outside of OLCF (Figure 2). Harvester was used to run jobs on Titan for nEDM and LSST/DESC projects. It was used to integrate computational resources at BNL, JLab and NERSC for the Lattice QCD computations.

### 2.3 Client tools: job description and description of workflows

PanDa jobs are usually described using PanDA Framework entities. In order to provide an easier way to describe jobs a new job description format based on YAML has been introduced. This format includes necessary information to execute jobs on HPC resources: required walltime, payload script, etc. This new format also allows to describe simple workflows which can be executed by PanDA Server without any dependency on third-party services. This workflow description workflow supports description of independent jobs, jobs with multiple inputs and multiple outputs. In order to generate multiple similar parametrized workflows a template engine for workflow generation has also been introduced.

For easier job management a set of client tools has been developed. It allows the submission of jobs in YAML format, as well as job status polling and cancelation. A separate tool has been developed which allows to generate YAML job descriptions from templates. Workflow management tools are currently being developed.

## 3 Experiments And Projects

### 3.1 PanDA for Computational Biology and Genomics

In collaboration with Center for Bioenergy Innovation at ORNL, the PanDA based workflow for epistasis research [9] was established.

The GBOOST [10] application, a GPU-based tool for detecting gene-gene interactions in genome-wide case control studies, was used for initial tests. Input data were located in a set of eight input directories of 152 MB each. Every PanDA job was configured to process single input directory in backfill mode on one Titan's worker node and walltime of 30 min, the output data volume was around 11 MB per job.

In 2019 we are going establish the workflow for larger scale computation runs on Titan and Summit [11].

### 3.2 PanDA for Molecular Dynamics

In collaboration with department of Chemistry and Biochemistry at the University of Texas Arlington we implemented a test to try out PanDA to support the Molecular Dynamics [12] study "Simulating Enzyme Catalysis, Conformational Change, and Ligand Binding/Release". The CHARMM (Chemistry at HARvard Macromolecular Mechanics) application was chosen as a basic payload tool. CHARMM is a molecular simulation program which primarily targets biological systems including peptides, proteins, prosthetic groups, and others; it is designed for hybrid MPI/OpenMP/GPU computing. For initial tests with PanDA we configured two types of jobs with different run time and output data volumes. These tests were successfully run on Titan using job submission via PanDA server at OLCF .

### 3.3 PanDA for IceCube

Together with experts from the IceCube [13] experiment we implemented the demonstrator for IceCube job submission to Titan using PanDA. IceCube is a particle detector at the South Pole that records the interactions of a nearly massless subatomic particle called the neutrino. Demonstrator includes the use of GPU simulation tools for atmospheric neutrinos, packed in Singularity container and remote stage-in/-out the data from a remote storage. Test jobs ran in one node mode with walltime of 120 minutes. After successful tests of this realistic IceCube jobs, we intended to perform full scale processing of 4500 files.

### 3.4 PanDA for BlueBrain

In 2017, a pilot project was started between BigPanDA and the Blue Brain Project (BBP) [14] of the Ecole Polytechnique Federale de Lausanne (EPFL) located in Lausanne, Switzerland. This proof of concept project is aimed at demonstrating the efficient application of the PanDA 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.

In the first phase of this joint project we supported the execution of BBP software on a variety of distributed computing systems powered by PanDA. The targeted systems for demonstration included: Intel x86-NVIDIA GPU based BBP clusters located in Geneva and Lugano, BBP IBM BlueGene/Q supercomputer in Lugano, the Titan supercomputer, and Cloud based resources such as Amazon Cloud.

In addition to standard PanDA components we developed the set of components to support BlueBrain user management, data flow, monitoring etc. Complete set of the tools provided within this project refers as PanDA portal and includes: Web-interface, Data Storage and Data Management System.

### 3.5 PanDa for LSST/DESC

A goal of Large Synoptic Survey Telescope (LSST) project is to conduct a 10-year survey of the sky that is expected to deliver 200 petabytes of data after it begins full science operations in 2022. The project will address some of the most pressing questions about the structure and evolution of the universe and the objects in it. It will require a large amount of simulations, which model the atmosphere, optics and camera to understand the collected data. The LSST Dark Energy Science Collaboration (LSST/DESC) [15] will prepare and perform a variety of cosmological analyses with the data received from LSST. This will allow to provide the community with state of the art analysis tools which, in turn, will help to expand the knowledge about dark energy.

For running LSST/DESC simulations with the PanDA WMS we have established a distributed testbed infrastructure that employs the resources of several sites on GridPP [16] and Open Science Grid [17] as well as the Titan supercomputer. In order to submit jobs to these sites we have used a PanDA server instance deployed on the Amazon AWS Cloud. Current LSST/PanDA computing environment includes 2 sites in Open Science Grid (US), 8 sites in GridPP (UK) and LAPP Grid Site (France), in total, there are 36 Grid endpoints that support LSST/DESC payloads. During our last tests we've reached the level of 3000 simultaneously running LSST/DESC jobs via PanDA infrastructure without any dedicated resources. Job duration was 4-24 hours with average duration around 7.5 hours. LSST/DESC workflows defined in YAML were tested with PanDA and Harvester as an edge service on NERSC

Cori [18]. "2-point Statistics Analysis" workflow (Figure 3) contains several steps where output from completed step is passed to the input of a next step. There were also intermediate steps that performed data preparation which had to be executed locally on a front node, main steps for the analysis were executed on worker nodes of Cori in Shifter containers prepared by BNL Astro group.

### 3.6 PanDA for Lattice QCD computations

Quantum chromodynamics (QCD) is the fundamental theory to describe the dynamics of quarks and gluons in hadrons. Lattice QCD (LQCD) simulations are the most efficient tools for nonperturbative analysis of QCD. Current LQCD payloads can be characterized as massively parallel, occupying thousands of nodes on leadership-class supercomputers.

The computations typically proceed in two phases: in the first phase, one generates thousands of configurations of the strong force fields (gluons), colloquially referred to as gauge fields. This computation is a long-chain Monte Carlo process, requiring the focused power of leadership class computing facilities for extended periods. In the second phase, these configurations are analyzed on various HPC resources [19]. Until a few years ago, the analysis phase would often account for a relatively small part of the cost of the overall calculation. In recent years, however, focus has turned to more challenging physical observables and new analysis. As a result, the relative costs have shifted to the point where analysis often requires an equal or greater amount of computation than gauge field generation.

In 2017, as a part of SciDAC-4 funded project, a collaboration was formed between several US LQCD groups and BigPanDA team with the goal to adopt PanDA WMS for the needs of the SciDAC-4 LQCD computational program.

In order to run LQCD payloads we have used our PanDA Server in Amazon cloud. Production campaigns were executed on BNL Institutional Cluster through a dedicated instance of Harvester. During the period between April and June 2018 13 TB of input data were processed, producing output of 176 GB. LQCD jobs used around 15000 GPU hours with average job duration around 12 hours.

Another branch of LQCD activities is developed by Jefferson Lab. They have their own instance of Harvester deployed on the front node of JLab internal computing environment and an instance of Harvester on NERSC. The goal is to create a private computing segment which will consist of several JLab clusters and NERSC resources (Figure 4).

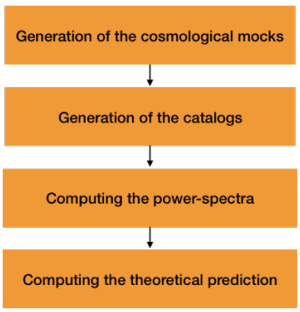


**Figure 3:** 2-point Statistics Analysis workflow

Different LQCD groups are planning to share input data generated on Titan among other computing resources. We have successfully tested these data transfers performed by Harvester between OLCF and NERSC via Globus file sharing [20].

In terms of preparation to future LQCD campaigns an instance of Harvester was successfully installed and tested on the front node of SummitDev supercomputer. It will be also available to other experiments and projects after Summit will be in full production.

### 3.7 PanDA for nEDM

Precision measurements of the properties of neutron help to study the violations of fundamental symmetries in the Standard Model of electroweak interactions. This research will allow to explain the dominance of matter over antimatter in the Universe. The goal of the nEDM [21] experiment at ORNL is to further improve the precision of this measurement by another factor of 100.

Current nEDM payloads run detector simulations using GEANT4 [22] with walltime around 20 minutes. We have successfully tested nEDM payloads executed through Harvester instance deployed on a login node of Titan supercomputer. No specific data movement was required from nEDM; input and output data remain on Titan's Lustre filesystem. nEDM team is planning to run their data challenges in 2019 through OLCF PanDA Server instance.

## 4 Conclusions

In terms of BigPanDA project we are making research on how PanDA features can be useful not only for ATLAS, but to non-ATLAS and non-HEP experiments and projects. The most remarkable achievements for now for BigPanDA project beyond ATLAS are the following:

- we created a version of PanDA Server, which does not depend on proprietary software
- several PanDA Server instances were deployed to serve different non-ATLAS payloads
- application from several projects (LSST/DESC, nEDM, BlueBrain, Molecular Dynamics, Computational Biology, LQCD) were executed via PanDA on various Grid and HPC resources
- production campaign has been started for LQCD computations on BNL Institutional Cluster

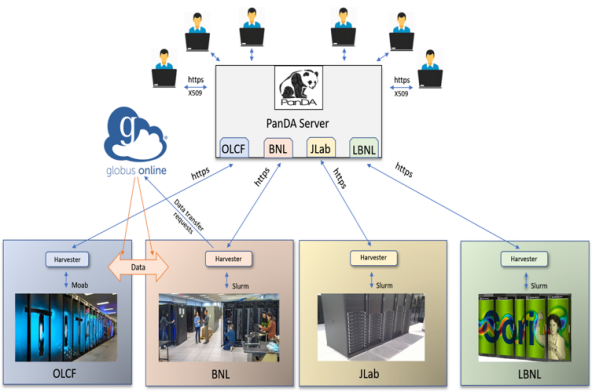


**Figure 4:** Future computing environment for Lattice QCD

- automated data transfer instruments have been tested with PanDA for non-ATLAS experiment on heterogeneous resources

The future goals for BigPanDA regarding non-ATLAS projects and experiments are:

- to integrate new Grid and HPC resources that will be served with PanDA Servers
- to prepare for payloads and workflows that will be run on next-generation HPCs like Summit

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. 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] T. Maeno, Journal of Physics: Conference Series **119**, 062036 (2008)
[2] G. Aad, E. Abat, J. Abdallah et al., Journal of Instrumentation **3**, S08003 (2008)
[3] I. Bird, Annual Review of Nuclear and Particle Science **61**, 99 (2011)
[4] *DOE ASCR web page*, `https://science.energy.gov/ascr/`
[5] *Titan web page*, `https://www.olcf.ornl.gov/titan/`
[6] X. Zhao, J. Hover, T. Wlodek, T. Wenaus, J. Frey, T. Tannenbaum, M. Livny, Journal of Physics: Conference Series **331**, 072069 (2011)
[7] P. Nilsson, Journal of Physics: Conference Series **119**, 062038 (2008)
[8] F. Megino, K. De, A. Klimentov, T. Maeno, P. Nilsson, D. Oleynik, S. Padolski, S. Panitkin, T. Wenaus, Journal of Physics: Conference Series **898**, 052002 (2017)
[9] P.A. Gros, H. Le Nagard, O. Tenaillon, Genetics **182**, 277 (2009)
[10] L.S. Yung, C. Yang, X. Wan, W. Yu, Bioinformatics **27**, 1309 (2011)
[11] *Summit web page*, `https://www.olcf.ornl.gov/summit/`
[12] B. Brooks, C. Brooks, A. Mackerell et al., Journal of Computational Chemistry **30**, 1545 (2009)
[13] F. Halzen, S.R. Klein, Review of Scientific Instruments **81**, 081101 (2010)
[14] H. Markram, Nature Reviews Neuroscience **7**, 153 EP (2006)
[15] *LSST DESC web page*, `http://www.lsst-desc.org`
[16] P.J.W. Faulkner, L.S. Lowe, C.L.A. Tan et al., Journal of Physics G: Nuclear and Particle Physics **32**, N1 (2005)
[17] R. Pordes, M. Livny, T. Wenaus et al., Journal of Physics: Conference Series **78**, 012057 (2007)
[18] *Cori web page*, `http://www.nersc.gov/users/computational-systems/cori/`
[19] R. Babich, M.A. Clark, B. Joo et al., *Scaling Lattice QCD beyond 100 GPUs*, in *SC11 International Conference for High Performance Computing, Networking, Storage and Analysis Seattle, Washington, November 12-18, 2011* (2011)
[20] *Globus data transfer web page*, `https://www.globus.org/data-transfer`
[21] S.K. Lamoreaux, R. Golub, Journal of Physics G: Nuclear and Particle Physics **36**, 104002 (2009)
[22] S. Agostinelli, J. Allison, K. Amako et al., Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **506**, 250 (2003)