Journal of Physics: Conference Series **513** (2014) 032062 doi:10.1088/1742-6596/513/3/032062

Evolution of the ATLAS PanDA workload management system for exascale computational science

T Maeno [1], K De [2], A Klimentov [1], P Nilsson [2], D Oleynik [2], S Panitkin [1], A Petrosyan [2], J Schovancova [1], A Vaniachine [3], T Wenaus [1], D Yu [1] and the ATLAS Collaboration

[1] Brookhaven National Laboratory, NY, USA
[2] University of Texas at Arlington, TX, USA
[3] Argonne National Laboratory, IL, USA

E-mail: tmaeno@bnl.gov

Abstract. An important foundation underlying the impressive success of data processing and analysis in the ATLAS experiment [1] at the LHC [2] is the Production and Distributed Analysis (PanDA) workload management system [3]. PanDA was designed specifically for ATLAS and proved to be highly successful in meeting all the distributed computing needs of the experiment. However, the core design of PanDA is not experiment specific. The PanDA workload management system is capable of meeting the needs of other data intensive scientific applications. Alpha-Magnetic Spectrometer [4], an astro-particle experiment on the International Space Station, and the Compact Muon Solenoid [5], an LHC experiment, have successfully evaluated PanDA and are pursuing its adoption. In this paper, a description of the new program of work to develop a generic version of PanDA will be given, as well as the progress in extending PanDA's capabilities to support supercomputers and clouds and to leverage intelligent networking. PanDA has demonstrated at a very large scale the value of automated dynamic brokering of diverse workloads across distributed computing resources. The next generation of PanDA will allow other data-intensive sciences and a wider exascale community employing a variety of computing platforms to benefit from ATLAS' experience and proven tools.

1. Introduction

The Production and Distributed Analysis (PanDA) system has been developed to meet ATLAS production and analysis requirements for a data-driven workload management system capable of operating at LHC data processing scale. PanDA has a highly scalable and flexible architecture. PanDA scalability has been demonstrated in ATLAS thorough the rapid increase in usage over the past four years, and is expected to easily meet the growing needs over the next decade. PanDA was designed to have the flexibility to adapt to emerging computing technologies in processing, storage, networking and distributed computing middleware. This flexibility has also been successfully demonstrated through the past six years of evolving technologies adapted by computing centers in ATLAS which span many continents. This proven scalability and flexibility makes PanDA well suited for adoption by a wide variety of exabyte scale sciences.

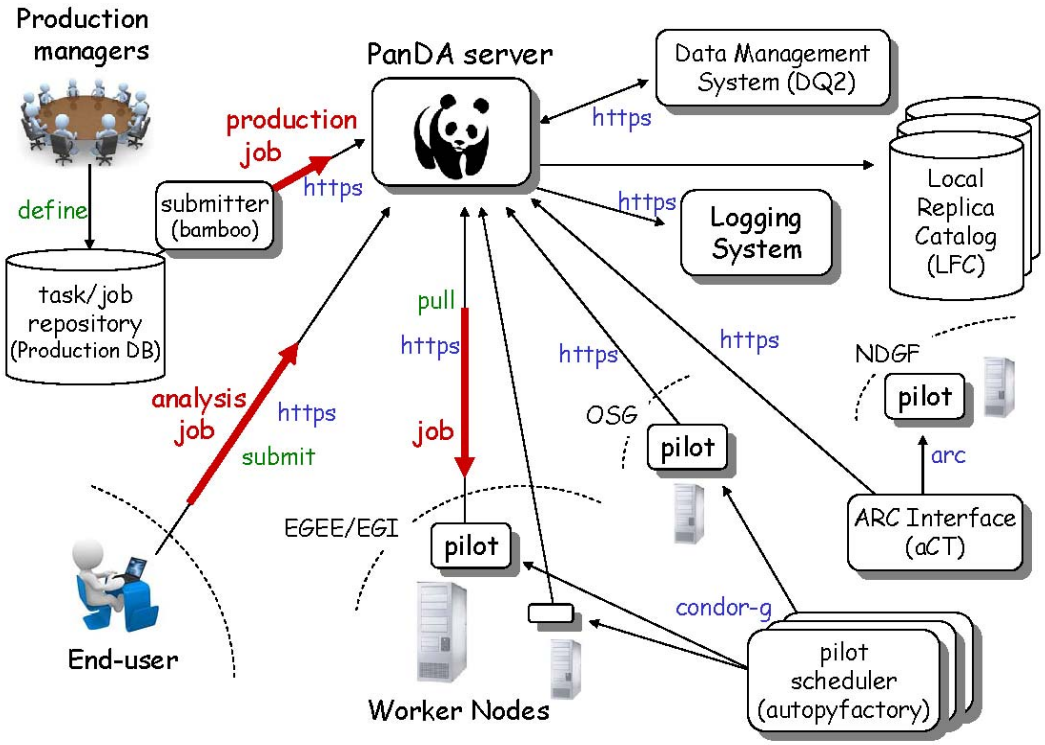

Figure 1. Schematic view of the PanDA System which is composed of the PanDA server, the PanDA pilot, pilot scheduler, and logging system.

The PanDA project began in 2005 as part of the US ATLAS program and it took over US ATLAS production responsibilities in December 2005. In 2008 PanDA was adopted as the workload management system (WMS) for the entire ATLAS collaboration. PanDA has performed well for data processing, simulation and analysis, while actively evolving to meet rapidly changing physics needs. Today, it is successfully managing more than 100 sites, about 5 million jobs per week, and about 1500 users. An overview of the PanDA system is shown in figure 1.

Through a work package supported by the Open Science Grid, PanDA was successfully used for molecular dynamics simulations of protein folding with the CHARMM molecular modeling software [6]. The implementation has been published for use with other molecular dynamics programs and other grids. In 2011, the Alpha Magnetic Spectrometer (AMS) experiment began testing PanDA for possible adoption. AMS dedicated a three FTE effort in 2012 to install PanDA in the AMS Payload Operations and Computing co-Center. AMS physicists successfully adopted PanDA and conducted their first Monte-Carlo simulation campaign in April 2012. AMS has set up the infrastructure and is reconfiguring computing facilities to have PanDA as the primary WMS of the experiment. The Compact Muon Solenoid (CMS) experiment has successfully evaluated PanDA after intensive work in collaboration with the core PanDA team and CERN IT, and is putting PanDA into production for analysis [7].

The interest in PanDA by other big data sciences provided the primary motivation to make a proposal titled “Next Generation Workload Management and Analysis System for Big Data” to the Advanced Scientific Computing Research program [8]. The idea was to generalize PanDA, providing location transparency of processing and data management, for High Energy Physics (HEP) community and other data-intensive sciences, and a wider exascale community. The proposal was approved and

the BigPanDA project has been active since September 2012. We will present in this paper a brief overview of BigPanDA, current status, and plans for the future.

2. Overview of BigPanDA

There are three dimensions for PanDA's system evolution; making PanDA available beyond ATLAS and HEP, extending PanDA beyond the Grid, and integration of network as a resource in workload management. The following four work packages have been identified:

- WP1 (Factorizing the core). Factorizing the core components of PanDA to enable adoption by a wide range of exascale scientific communities.
- WP2 (Extending the scope). Evolving PanDA to support extreme scale computing clouds and Leadership Computing Facilities [9].
- WP3 (Leveraging intelligent networks). Integrating network services and real-time data access to the PanDA workflow.
- WP4 (Usability and monitoring). Real time monitoring and visualization package for PanDA.

The work packages are described in the following sections. ATLAS remains the largest user community and provides the primary focus to PanDA development although PanDA is not only for ATLAS any longer. BigPanDA will encompass one set of software packages for all experiments including ATLAS, while experiment-specific code will be encapsulated into plug-ins. The effort has to be incremental and coherent with many challenging developments in ATLAS [10-12].

2.1. WP1 (Factorizing the core)

It is relatively straightforward to take the core components of PanDA and package them in an experiment neutral package. PanDA consists of several subsystems. Each of these is being factorized into general components and customizable layers. The experiment specific layers will be configurable. Advanced features will have sensible defaults and can be turned on for demanding applications. The main components to be factorized are the PanDA server, the PanDA database, the PanDA pilot system, Auto pilot factory, and PanDA monitoring.

2.2. WP2 (Extending the scope)

The PanDA system had used the Grid infrastructure for large scale production. An objective of the work package is to add extra resources such as computing clouds and Leadership Computing Facilities to those supported by PanDA. Extending PanDA beyond the Grid will further expand the potential user community and the resources available to them.

2.3. WP3 (Leveraging intelligent networks)

Many of the research efforts into dynamic network provisioning, quality of service and traffic management have made their way into production services on major research and education networks globally. These capabilities include:

- Extensive standardized monitoring data from network performance monitoring (perfSONAR [13]) instances located within backbone networks, national networks, regional networks and end-sites.
- Traffic engineering capabilities that allow transparent re-routing of high impact flows onto separate infrastructures.

- Dynamic circuit capabilities allowing the temporary construction of point-to-point network circuits with dedicated bandwidth for specific durations.
- Intelligent networking systems, such as Virtual Network on Demand [14], that have capabilities to create on demand virtual network domains with necessary bandwidth guarantees along hops.

The question for globally distributed, data-intensive, scientific collaborations is how best to take advantage of these services to improve their ability to do their science. We will provide a means of doing this by:

- Integrating these services within existing and evolving scientific infrastructures. For example, PanDA currently does not interface with any of the advanced network provisioning technologies available.
- Providing intelligent decision support for when such services are beneficial to the task.
- Automating the discovery and use of such services transparently to the scientists. As an example, instead of exposing intelligent network services directly to scientists, PanDA will directly interface with them without any involvement of scientists.

2.4. WP4 (Usability and monitoring)

The PanDA monitoring has been identified to require a special effort for factorization and generalization since each experiment has own workflow and visualization needs. BigPanDA will provide a generic PanDA monitor browser view and skeleton from which experiment (ATLAS and other) browser views are derived customizations. Also generic components and APIs will be provided for user communities to easily implement and customize their monitoring views and workflow visualizations.

3. Current Status and Plans

The three year plan is as follows: The goal of the first year was to set the collaboration and define algorithms and metrics. The hiring process was completed and a three FTE development team formed in June 2013. Implementation and prototyping will take place in the second year, and production and operations in the third year. The current status and near future plans are described in the following sections.

3.1. WP1& WP4

The PanDA server has been improved to support MySQL in addition to Oracle. Each experiment can choose the database backend in the configuration file. A PanDA server instance has been installed on the Amazon Elastic Computing Cloud [15]. It is running on a MySQL backend and will server non-LHC experiments like the Large Synoptic Survey Telescope [16]. Refactoring of the PanDA server is well underway to decompose experiment-specific code to plug-ins. The PanDA pilot is being refactored in the context of the Common Analysis Framework project. New experiment specific classes have been introduced, which enables better organization of the code. Changes are being introduced gradually, to avoid affecting current production. The details of the PanDA pilot evolution are described in reference [10]. The next step is to refactor the PanDA monitoring.

3.2. WP2

Google allocated Google Computing Engine [17] resources for ATLAS for free (~5M cpu hours and 4000 cores for ~2 months). Resources were organized as a PanDA queue based on HTCondor. CentOS6-based custom build images and SL5 compatibility libraries were used to run ATLAS software. We have successfully demonstrated transparent inclusion of cloud resources into the ATLAS

grid system, as reported in reference [18]. Those resources were delivered to ATLAS as a production resource and not as an R&D platform, the idea being to test long term stability while running a cloud cluster similar in size to Tier 2 site in ATLAS.

In collaboration with Oak Ridge Leadership Computing Facility, we have been gaining experience with all relevant aspects of the Titan platform [19] and workload, such as job submission mechanism, job output handling, local storage system details, outside transfers details, security environment and monitoring model. A pilot/agent model is being developed for Titan. Figure 2 shows a possible scenario to run ATLAS jobs on Titan. MC generators will be initial use case.

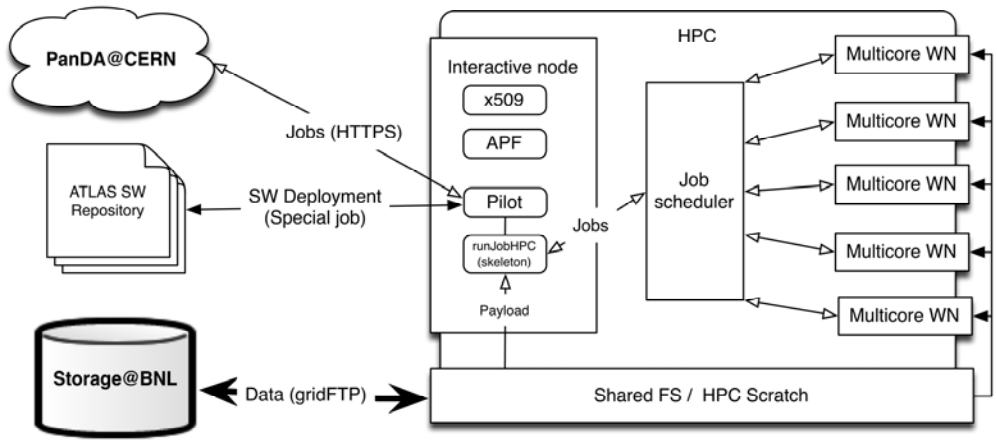

Figure 2. A possible scenario to run ATLAS jobs on Titan. PanDA pilot runs on the interactive node and communicates with the PanDA server. Each pilot splits a job to multiple sub-jobs and they are submitted to the local job scheduler. ATLAS software will be installed to the shared file system by using special jobs. Output data are transferred from the shared file system to a remote storage system by ATLAS data management system.

3.3. WP3

The initial goals are to optimize the site selection algorithm in the brokerage based on network capability and to use network provisioning for dynamic data placement. The effort has been synchronized with two other projects; one is the US ATLAS project for integration of the Federated ATLAS Xrootd (FAX) [20] with PanDA, and the other is the Advanced Network Services for Experiments project [21]. A three layered software architecture has been designed as follows to give network information to the PanDA server.

- "Collector" collects network performance information from various sources such as perfsonar, Grid sites status board, FAX, etc,
- "AGIS" [22] stores historical network information which are provide by "Collector", and
- "Calculator" retrieves information from "AGIS" and calculates weights which the brokerage takes into account for site selection.

4. Conclusions

The PanDA system played a key role during LHC Run 1 data processing, simulation and analysis with great success, while actively evolving to meet rapidly changing physics needs. The interest in PanDA

by other big data sciences provided the primary motivation to generalize the PanDA system. The BigPanDA project gives us a great opportunity to evolve PanDA beyond ATLAS. There is progress in various areas: networking, VO independent PanDA instance, cloud computing, and HPC.

Acknowledgments
This manuscript has been authored by employees of Brookhaven Science Associates, LLC under Contract No. DE-AC02-98CH10886 with the U.S. Department of Energy. The publisher by accepting the manuscript for publication acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References
[1] The ATLAS experiment http://www.atlas.ch
[2] The LHC experiment http://lhc.web.cern.ch
[3] Maeno T 2011 Overview of ATLAS PanDA Workload Management *J. Phys. Conf. Ser.* **331**
[4] Aguilar M, et al 2002 The Alpha Magnetic Spectrometer (AMS) on the International Space Station *Physics Reports* **366** 331-405
[5] The CMS experiment http://cms.web.cern.ch
[6] The CHARMM molecular modeling software http://www.charmm.org
[7] Spiga D 2013 The Common Analysis Framework Project *Int. Conf. on Computing in High Energy and Nuclear Physics 2013* Amsterdam
[8] The Advanced Scientific Computing Research program http://science.energy.gov/ascr/
[9] Ledership Computing Facilities http://science.energy.gov/ascr/facilities/
[10] Potekhin M 2013 Task Management in the New ATLAS Production System *Int. Conf. on Computing in High Energy and Nuclear Physics 2013* Amsterdam
[11] Nilsson P 2013 Next Generation PanDA Pilot for ATLAS and Other Experiments *Int. Conf. on Computing in High Energy and Nuclear Physics 2013* Amsterdam
[12] Dimitrov G 2013 Next generation database relational solutions for ATLAS distributed computing *Int. Conf. on Computing in High Energy and Nuclear Physics 2013* Amsterdam
[13] perfSONAR http://www.perfsonar.net/
[14] Katramatos D 2012 Virtual Network On Demand: Dedicating Network Resources to Distributed Scientific Workflows *Int. Workshop on Data Intensive Distributed Computing 2012* Delft
[15] The Amazon Elastic Computing Cloud http://aws.amazon.com/ec2/
[16] The Large Synoptic Survey Telescope http://www.lsst.org/lsst/
[17] The Google Computing Engine https://cloud.google.com/products/compute-engine
[18] Panitkin S 2013 ATLAS Cloud Computing R&D *Int. Conf. on Computing in High Energy and Nuclear Physics 2013* Amsterdam
[19] The Titan platform http://www.olcf.ornl.gov/titan/
[20] Vukotic I 2013 The True Cost of Data Access in ATLAS *Int. Conf. on Computing in High Energy and Nuclear Physics 2013* Amsterdam
[21] Melo A 2013 Integrating the Network into LHC Experiments: Update on the ANSE (Advanced Network Services for Experiments) Project *Int. Conf. on Computing in High Energy and Nuclear Physics 2013* Amsterdam
[22] Anisenkov A 2013 AGIS: The ATLAS Grid Information System *Int. Conf. on Computing in High Energy and Nuclear Physics 2013* Amsterdam