Journal of Physics: Conference Series **664** (2015) 092018 doi:10.1088/1742-6596/664/9/092018

Integration of Russian Tier-1 Grid Center with High Performance Computers at NRC-KI for LHC experiments and beyond HENP

A Belyaev[1], A Berezhnaya[1], L Betev[2], P Buncic[2], K De[3], D Drizhuk[1], A Klimentov[1,4], Y Lazin[1], I Lyalin[1], R Mashinistov[1,5], A Novikov[1], D Oleynik[3,6], A Polyakov[1], A Poyda[1], E Ryabinkin[1,7], A Teslyuk[1], I Tkachenko[1] and L Yasnopolskiy[1]

[1] National Research Center "Kurchatov Institute", Moscow, Russia
[2] CERN, Geneva, Switzerland
[3] University of Texas, Arlington, TX, USA
[4] Brookhaven National Laboratory, Upton, NY, USA
[5] P.N. Lebedev Institute of Physics (Russian Academy of Sciences), Moscow, Russia
[6] Joint Institute of Nuclear Research, Dubna, Russia
[7] Moscow Institute for Physics and Technology, Moscow, Russia

E-mail: ruslan.mashinistov@cern.ch

Abstract. The LHC experiments are preparing for the precision measurements and further discoveries that will be made possible by higher LHC energies from April 2015 (LHC Run2). The need for simulation, data processing and analysis would overwhelm the expected capacity of grid infrastructure computing facilities deployed by the Worldwide LHC Computing Grid (WLCG). To meet this challenge the integration of the opportunistic resources into LHC computing model is highly important. The Tier-1 facility at Kurchatov Institute (NRC-KI) in Moscow is a part of WLCG and it will process, simulate and store up to 10% of total data obtained from ALICE, ATLAS and LHCb experiments. In addition Kurchatov Institute has supercomputers with peak performance 0.12 PFLOPS. The delegation of even a fraction of supercomputing resources to the LHC Computing will notably increase total capacity. In 2014 the development a portal combining a Tier-1 and a supercomputer in Kurchatov Institute was started to provide common interfaces and storage. The portal will be used not only for HENP experiments, but also by other data- and compute-intensive sciences like biology with genome sequencing analysis; astrophysics with cosmic rays analysis, antimatter and dark matter search, etc.

1. Introduction

The data intensive sciences, such as high energy and nuclear physics, astrophysics, and biology will generate exabytes of data in the near future. The challenges posed by such data-intensive sciences are numerous and not limited to the unprecedented size of the data. The scientific data is often highly distributed and accessed by large international collaborations.

To address an unprecedented multi-petabyte data processing challenge, the LHC collaborations are relying on the computational grid infrastructure consisting of hundreds of distributed computing centers deployed by the Worldwide LHC Computing Grid (WLCG) [1]. Using the massive data processing power of the Grid, more than 10 000 scientists analyse LHC data in search for physics discoveries. The Tier-1 facility at Kurchatov Institute (NRC-KI) in Moscow is a part of WLCG and it will process, simulate and store up to 10% of total data obtained from ALICE [2], ATLAS [3] and LHCb [4] experiments. Sophisticated Workload Management Systems (WMS) are used to manage the data processing, simulations and analysis.

One of the scalable WMS developed for high-energy physics is PanDA [5]. The ATLAS experiment uses PanDA for managing the workflow for all data processing on the WLCG. Through PanDA, ATLAS physicists see a single computing facility, even though the data centers are physically scattered all over the world. PanDA is now being generalized and packaged, as a software system already proven at exabyte scales, for the wider use in other sciences.

2. ATLAS Use Case

2.1. Hierarchical Distributed Computing

In ATLAS, WLCG computing facilities are organized into tiered hierarchy. CERN is the source of all primary data, referred to as Tier-0. There are eleven Tier-1 centers including the NRC-KI Tier-1. Each Tier-1 center hierarchically supports 5-20 Tier-2 centers in a cloud. PanDA is deployed at all ATLAS Tier-1 and Tier-2 centers. Each PanDA site provides a grid accessible Compute Element (CE) and a Storage Element (SE). The pilot jobs are continuously and automatically scheduled at the CE of each site. The pilot jobs are used for acquisition of processing resources. The PanDA server based on brokerage criteria assigns workload jobs to successfully activated and validated pilots. 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. When the pilot jobs start execution, they contact PanDA Apache servers, which then dispatch the execution workload. PanDA maintains a central database of all activities, and consequently a central queue of all workflows. This architecture provides an integrated view of all resources managed by PanDA. The pilot based system also enables integration of non-grid based resources. Local resources at universities are integrated using local pilot submission factories. New cloud-based resources are also added to PanDA using the CE model.

2.2. Hierarchical Workflow

In LHC data processing, the task became a main unit of computation (instead of a job). A computational task is a collection of similar jobs that could be executed in parallel. In petascale data management scientists deal not with individual files but with large datasets - collections of similar files. Similarly, a task – not a job – is a major unit in data processing. Splitting of a large data processing task into jobs is similar to the splitting of a large file into smaller TCP/IP packets during the FTP data transfer [6]. Generally, during a file transfer, application scientists are not concerned how many TCP/IP packets were dropped. The network performance is an area of concern of the network researchers and engineers. Similarly, scientists do not care about transient job failures, when data processing delivers “six sigma quality” performance for the petascale data processing campaigns with thousands of tasks [7]. In order to manage the variety of LHC physics (exceeding 35K physics samples per year), the individual data processing tasks are organized into workflows [8].

3. Common ATLAS Tier-1 site for Grid resources and supercomputer at NRC-KI

3.1. Extending PanDA to HPC facilities

Initially PanDA has been used only on the facilities managed by WLCG. The support for High Performance Computers (HPC) expands the potential user community for PanDA and also, in the near term, benefit LHC experiments by running jobs on HPC facilities with direct access to the data hosted by Grid centers, such that LHC experiments can acquire supplementary CPU resources. Extending PanDA beyond the Grid further expands the potential user community and the resources available to them, for example, making it possible for non-LHC and non-HENP experiments to use PanDA.

3.2. PanDA extension for supercomputer at NRC-KI for ATLAS

Kurchatov Institute has the second-generation supercomputer HPC2 based on Intel(R) Xeon(R) E5450@3.00GHz, with peak performance of 122,9 TFLOPS. Currently two worker nodes (16 cores) are dedicated for ATLAS. Integration schema of PanDA WMS with Kurchatov's supercomputer is shown in figure 2. In order to launch ATLAS tasks the local Auto-Pilots Factory (APF) was installed. APF - an independent subsystem manages the delivery of "pilots" to worker nodes via a number of schedulers ('pilot factories') serving the sites at which PanDA operates. Pilots are python scripts used to gather the information about computing resources, request working task, handle input/output data. Working tasks are passing by PanDA server to successfully activated and approved pilots on the basis of choice of resources. The server part component named "dispatcher" operates the task requests from pilots. It dispatches the tasks following the allocated resources.

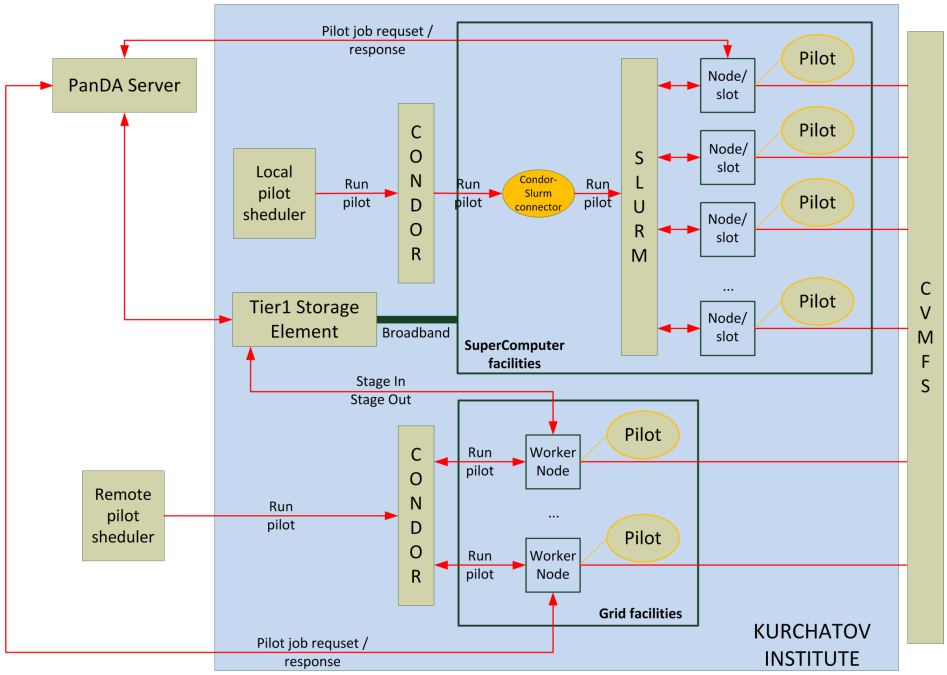


Figure 1. Integration of PanDA WMS with Kurchatov's supercomputer

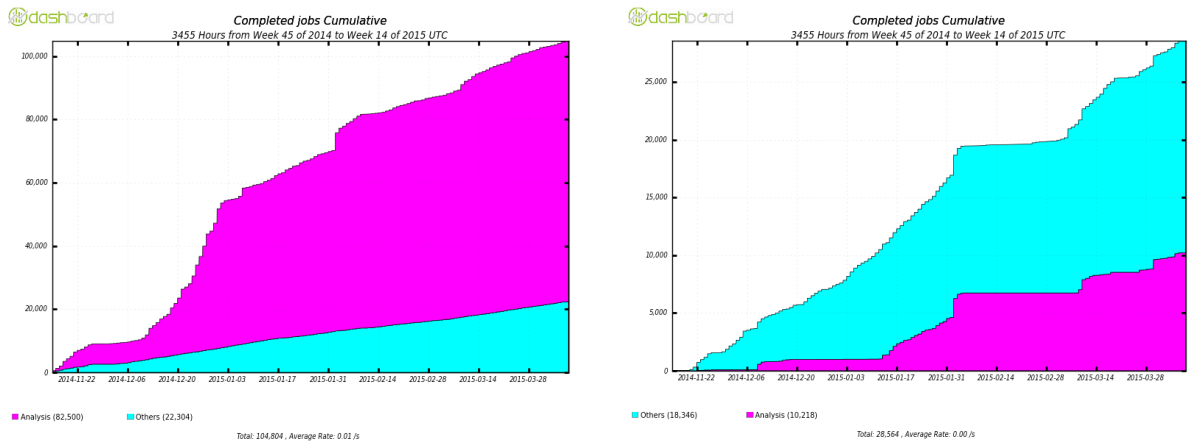
The integration has been done with minimal modifications in code of PanDA subsystems. The interface Condor – SLURM connector has been developed for integration of the APF with the supercomputer. Current version of APF could maintain only narrow set of the LRMS (Local Resource Management System) systems, namely CondorG and CondorLocal. However the HPC2 management is possible only via SLURM LRMS system.

The integration was done following the basic approach where one pilot is running on one core as currently ATLAS tasks are not parallelizing. The worker nodes of the supercomputer have access to Internet, this allows not to take additional action for data and SW moving and for connectivity of pilot tasks with PanDA server. For support of the ATLAS analysis task submission the installation of CVMFS [9] (CERN Virtual Machine File System) on the working nodes was done. CVMFS provide access to full set of ATLAS SW releases. In addition the required libraries and compilers were installed.

3.3. ATLAS site validation

For ATLAS analysis tasks new PanDA site "ANALY_RRC-KI-HPC" within ATLAS Tier-1 site of National Research Center "Kurchatov Institute" was defined. The system-wide site/queue information database recording static and dynamic information used throughout PanDA to configure and control system behavior from the region level down to the individual queue level. The database is an information cache, gathering data from grid information systems, the data management system and other sources. It is used by pilots to configure themselves appropriately for the queue they land on.

New site contains own PanDA resource/queue description to address ATLAS user analysis tasks to the worker nodes of supercomputer HPC2. About 200 user analysis jobs are running and successfully and finishing daily. One of the most important studies dedicated to be solved on the new supercomputer is the reconstruction of *pp* events for ATLAS Transition Radiation Tracker (TRT) performance at high occupancies (done by the ATLAS TRT software group). Figure 2 shows statistics of the completed jobs at the portal for Grid Tier-1 and Supercomputer resources.



(a) (b)

Figure 2. Completed jobs (cumulative) at the NRC-KI Tier-1 site (November - April). Grid Tier-1 (a); Supercomputer (b)

4. PanDA beyond ATLAS at NRC-KI

4.1. PanDA instance at NRC-KI

In 2014 the pioneering work on implementation of the portal, which will combine the Kurchatov Institute's Tier-1 site and the supercomputer HPC2 into Kurchatov Institute's Tier-1 center has begun. This center should allow starting tasks optionally not only on Grid but on the supercomputer as well and to collect the results in the common storage. As the base technology has been used WMS PanDA. In further work there is a plan to extend the developed portal into single system for task/data access and management in federative heterogeneous resources. The work includes several directions: increasing the number of aggregated heterogeneous resources types, expanding their geography, increasing the number of maintained experiments etc. The developed portal should be used not only

for the HENP tasks but for tasks in other scientific areas, for example, in the biology for genome sequencing data processing, in the astrophysics for researching the cosmic rays structure, for searching the antimatter and the dark matter. The development and integration of new big data management technologies with re-engineered PanDA system is necessary to overcome these challenges. The new system (called megaPanDA) will work at megascience scale having a state-of-the-art data management component.

The installation of PanDA WMS has been done at NRC-KI. Local instance consists of next main components: server, auto-pilots factory, monitor and database server. MySQL was chosen as database backend technology. Auto pilot factory is configured in a way that it works with standard pilots to run ATLAS-jobs derived from production server at CERN and also factory operates with HPC-pilot to run non-ATLAS jobs derived from local PanDA server at NRC-KI. The PanDA monitor realizes overall monitoring of the tasks. It provides detailed information about the tasks and site for their status diagnostics [10].

4.2. PanDA pilot scheme extending

The ATLAS jobs are single core. Main demand for non-ATLAS jobs processing is to support multi core jobs on par with single core. So PanDA pilot modification was required. Figure 3 shows basic implementation of the PanDA pilots.

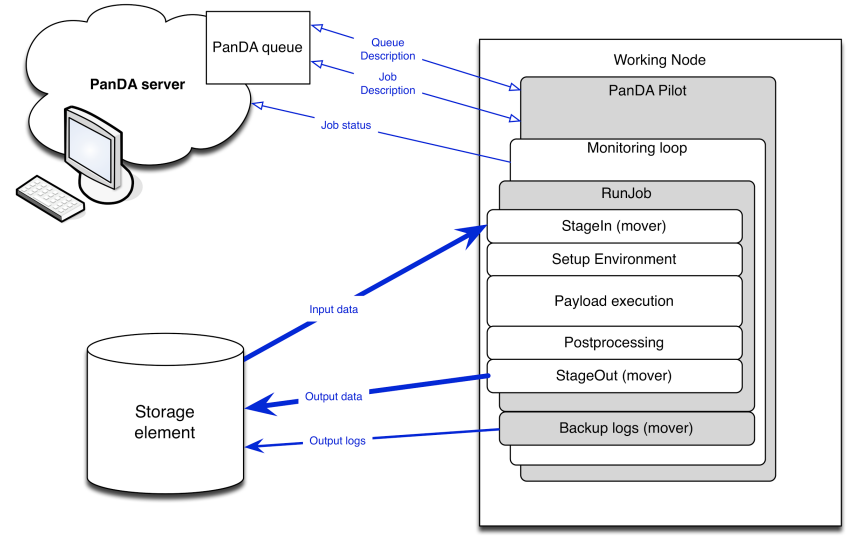


Figure 3. Basic pilot scheme for the ATLAS experiment

The server stores the information about all resources and jobs in the system. The server sorting out the jobs among various queues, associated with computing sites. The pilot runs in the same computing node as the requested job. Main three modules of the pilot are:

- Monitor, which control the payload and pilot execution and update PanDA server
- RunJob, which launch the payload at the working node and process its outputs
- Mover, which get the input data and transfer the output data and logs and register them into the DMS

All pilot modules run on the same node and use local access to data as shown at figure 3.

While this scheme works well for ATLAS jobs, it cannot be used for multicore jobs with the number of cores unknown beforehand. Therefore, the pilot allocates unappropriate amount of resources, and then either pilot fail to satisfy the needs of the job or part of the resources will idle.

To solve this issue we extended the pilot scheme to work in “remote job launch” mode. The extended scheme is shown at figure 4.

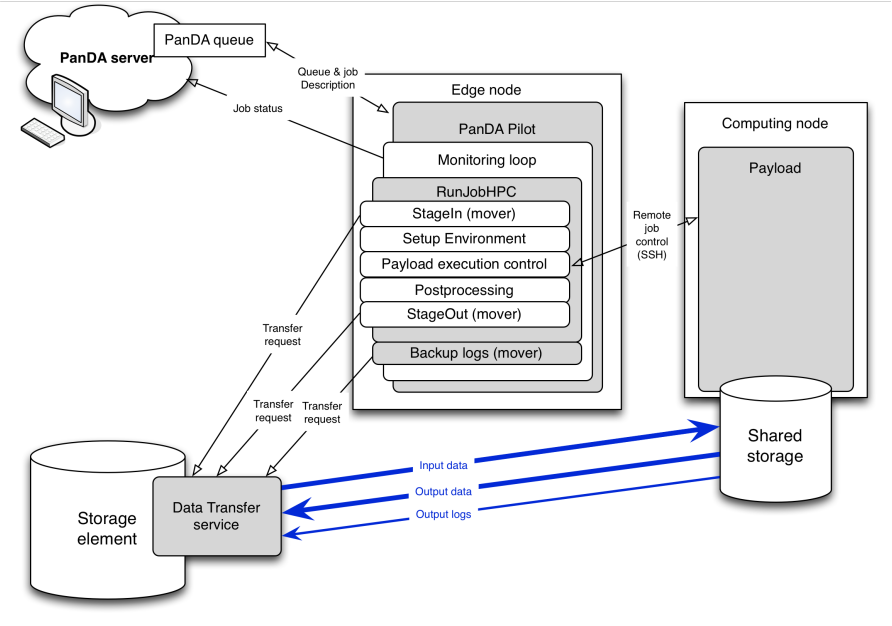


Figure 4. Pilot scheme for remote job launch mode

In this scheme the pilot runs different node from jobs. To implement this scheme we extended pilot modules to support remote control mode through SSH tunnels.

- The RunJob module launches payload by remote call of batch system. In this scheme exact amount of resources could be requested from batch system accordingly to job definition.
- Mover plugin, which allow transfer data through requests to special data-transfer system, was realized.

4.3. PanDA beyond HENP experiments: the biology use case

The developed portal suits well for integration of heterogeneous resources with various architectures as well as for different compute-intensive sciences. We have adapted workflow management system PanDA for grid/Cloud/supercomputer facilities with shared data storage and extended job lifecycle management schemes to support multicore parallel jobs, developed web interface to simplify job definition and submission by users. We extended PanDA users community beyond LHC by implementation of local authentication mechanism.

The rich menu of features provided, coupled with support for heterogeneous computing environments, makes PanDA ideally suited for scientific data processing. One of the aims of this project was to run user applications for scientific areas other than HENP where we can use advantages of the architecture of our workload management system. We provided this portal at NRC "Kurchatov institute" to run compute-intensive bioinformatics jobs including multicore ones of genome sequencing analysis. Originally biology group used common approach of batch system scheduling. Users ran their personal jobs independently without central management. For highly repetitive operations on large datasets customized scripts were written. PanDA WMS ideally cover this needs and provide a wide range of features like central management of the workload with automated leveraging of the jobs priorities, online monitoring of the jobs, resources and data statuses, and archive of all actions in the system. In future we are going to expand the application area of the portal into other scientific disciplines.

As a pioneer application we chose genome sequence de-novo assembly and short sequence reads to long reference alignment. We prepared software packages Bowtie2 [11] for sequence alignment, ABySS 1.9.0 [12], SPAdes 3.5 [13] - for de-novo assembly and QUAST [14] package for providing

detailed and user-friendly reports of de-novo assembly results. SPAdes and Bowtie2 support OpenMP while ABySS can take advantage of MPI for parallelization.

User-friendly Web-interface was developed to simplify jobs submission and results gathering procedures. Interface provides local user authentication, graphics menu to select desirable distributive, upload input files, specify running command and initiates job submission to PanDA workflow. Main features of the interface:

1. Web-interface and API were implemented to interact with scientists and external integrated workflow composers. A new authorization endpoint allows us to remove restrictions of CERN certificates and Virtual Organizations.
2. The modular structure supports connection with various external storage systems including unsupported by original PanDA. All files passing through job's lifecyrcle are described in lightweight file catalog in generalized style.
3. The interface provides monitoring of submitted jobs during their lifecycle, fetches output files and produces report based on log files analysis.

Our service was first probed on synthetic data. Further we used our de novo assembly service for assembly of microbial reads that were sequenced in Kurchatov institute's genomics lab. Our web service is actively used by bioinformaticians since it combines computing power of supercomputer facilities at NRC KI with user-friendly application specific web interface. Highly scalable and flexible architecture of PanDA allows to involve additional cloud computing resources is supercomputers nodes are busy.

5. Summary and conclusions

The LHC experiments preparing for the precision measurements and further discoveries that will be made possible by much higher LHC collision rates from early 2015 (Run2). The need for simulation, data processing and analysis and would overwhelm the expected capacity of WLCG computing facilities unless the range and precision of physics studies were to be curtailed. To meet this challenge the integration of the opportunistic resources into LHC computing model is highly important.

The Tier-1 facility at Kurchatov Institute (NRC-KI) in Moscow is a part of WLCG and it will process and store up to 10% of total data obtained from ALICE, ATLAS and LHCb experiments. In addition Kurchatov Institute has supercomputers with peak performance 0.12 PFLOPS. The delegation of even a fraction of super-computing resources to the LHC Computing will notably increase total capacity.

In 2014 a pioneer work has been started to develop a portal combining a Tier-1 and a supercomputer in Kurchatov Institute. This portal is aimed to provide interfaces to run Monte-Carlo simulation at the Tier-1 Grid and supercomputer, using common portal and storage. PanDA (Production and Distributed Analysis) workload management system having great results at the ATLAS was chosen as underlying technology.

An implementation of HPC-pilot at NRC-KI was used to run first test genome sequencing tasks. We have demonstrated that our system is effective for two biological applications: short read alignment using bowtie2 tool and genome de-novo assembly using ABySS software. These two tools are used often for next generation sequencing data analysis. We have used two approaches for data processing parallelization: input data partitioning and OpenMP thread parallelization for bowtie2 and MPI+OpenMP combination for ABySS. Both approaches showed to be effective when processing genomic datasets varying from 1 to 100 Gbytes.

The portal is provided to run ATLAS tasks on Grid Tier-1 and/or supercomputer resources. It will be used not only for HENP, but also for other data-intensive sciences like biology with genome sequencing analysis; astrophysics with cosmic rays analysis, antimatter and dark matter search, etc. New PanDA site contains resources description to address ATLAS user analysis tasks to the HPC2 supercomputer. About 200 user analysis jobs are running and finishing successfully daily. One of the most important dedicated studies to be solved on new supercomputer is reconstruction of high mu pp events for Transition Radiation Tracker performance at high occupancy study (ATLAS TRT SW

group). First result got on the new resource has been verified by crosschecking with same results gathered on Kurchatov Institute's Tier-1 grid site.

Contributing in the development the PanDA dedicated to be applicable to different supercomputers and HPC architectures and to provide the ability of multicore parallel job running via MPI technology the NRC-KI team had performed the successful prototype of genome sequencing data processing task.

Acknowledgments

We wish to thank all our colleagues who contributed to ATLAS data processing and Monte-Carlo simulation activities and to PanDA software development and installation. We wish to thank OLCF computing team for helping to port PanDA on Titan. 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 No. DE-SC0012704, No. DE-AC02-98CH10886 and Contract No. DE-AC02-06CH11357. NRC-KI team work was funded by the Russian Ministry of Science and Education under Contract No 14.Z50.31.0024. Supercomputing resources at NRC-KI are supported as a part of the center for collective usage (project RFMEFI62114X0006, funded by Ministry of Science and Education of Russia).

References

[1] Bird I, Computing for the Large Hadron Collider. *Annu. Rev. Nucl. Part. S.* 2011; 61: 99.
[2] Aamodt K *et al*. 2008 The ALICE experiment at the CERN LHC, *JINST* **3** S08002
[3] Aad G *et al*. 2008 ATLAS Experiment at the CERN Large Hadron Collider. *JINST* **3** S08003.
[4] Alves, A *et al*. 2008 The LHCb Detector at the LHC, *JINST* **3** S08005
[5] Maeno T *et al*. 2008 PanDA: Distributed Production and Distributed Analysis System for ATLAS *J. Phys.: Conf. Ser.* **119** 062036
[6] Vaniachine A V on behalf of the ATLAS and CMS Collaborations, Advancements in Big Data Processing in the ATLAS and CMS Experiments. 2012 In: *Proc. of the Fifth International Conference "Distributed computing and Grid-technologies in Science and Education"* (Dubna, July 16-21, 2012), Dubna, JINR, p. 224.
[7] Vaniachine A on behalf of the ATLAS Collaboration. 2011 ATLAS detector data processing on the Grid. *IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)* p. 104
[8] Multilevel Workflow System in the ATLAS Experiment. Borodin M, De K, Garcia Navarro J, Golubkov D, Klimentov A, Maeno T and Vaniachine A on behalf of the ATLAS Collaboration, 2015, *J. Phys.: Conf. Ser.* 608:012015
[9] R Meusel et al.; 2015 J. Phys.: Conf. Ser. 608 012031"Recent Developments in the CernVM-File System Server Backend", *16th International workshop on Advanced Computing and Analysis Techniques in physics research (ACAT2014),* 6 pp.
[10] Klimentov A, Nevski P, Potekhin M and Wenaus T 2011 The ATLAS PanDA Monitoring System and its Evolution *J. Phys.: Conf. Ser.* **331** 072058
[11] Johns Hopkins University, 2015, Bowtie2
[12] Simpson *et al*. 2009, ABySS 1.9.0
[13] Bankevih *et al*. 2012, SPAdes 3.5
[14] Gurevich *et al*. 2013, QUAST