HEP Computing Tools, Grid and Supercomputers for Genome Sequencing Studies

K De¹, A Klimentov²³, T Maeno², R Mashinistov¹, A Novikov³, A Poyda³, I Tertychnyy³ and T Wenaus²

¹ University of Texas at Arlington, Arlington, TX 76019, USA.
² Brookhaven National Laboratory, Upton, NY 11973, USA.
³ National Rerearch Center “Kurchatov Institute”, Moscow, Russia.

E-mail: Ruslan.Mashinistov@cern.ch

Abstract. PanDA - Production and Distributed Analysis Workload Management System has been developed to address ATLAS experiment at LHC data processing and analysis challenges. Recently PanDA has been extended to run HEP scientific applications on Leadership Class Facilities and supercomputers. The success of the projects to use PanDA beyond HEP and Grid has drawn attention from other compute intensive sciences such as bioinformatics. Recent advances of Next Generation Genome Sequencing (NGS) technology led to increasing streams of sequencing data that need to be processed, analysed and made available for bioinformaticians worldwide. Analysis of genomes sequencing data using popular software pipeline PALEOMIX can take a month even running it on the powerful computer resource. In this paper we will describe the adaptation the PALEOMIX pipeline to run it on a distributed computing environment powered by PanDA. To run pipeline we split input files into chunks which are run separately on different nodes as separate inputs for PALEOMIX and finally merge output file, it is very similar to what it done by ATLAS to process and to simulate data. We dramatically decreased the total walltime because of jobs (re)submission automation and brokering within PanDA. Using software tools developed initially for HEP and Grid can reduce payload execution time for Mammoths DNA samples from weeks to days.

1. Introduction

Modern experiments face unprecedented computing challenges. Heterogeneous computer resources are distributed worldwide, thousands of scientists analyse the data remotely, the volume of processed data is beyond the exabyte scale, while data processing requires more than a few billion hours of computing usage per year. The Large Hadron Collider (LHC) [1], operating at the international CERN Laboratory in Geneva, Switzerland, is leading Big Data driven scientific explorations. Experiments at the LHC explore the fundamental nature of matter and the basic forces that shape our universe, and were recently credited for the discovery of a Higgs boson [2]. ATLAS [3], one of the largest collaborations ever assembled in the sciences, is at the forefront of research at the LHC. To address an unprecedented multi-petabyte data processing challenge, the ATLAS experiment is relying on a heterogeneous distributed computational infrastructure [4].

Another example is the modern biology studies in particular the genome sequencing data analysis. The input data can be up to hundreds of gigabytes, and the computing tasks itself are very resource intensive. Running the typical task on a single multi-core server can take a few months. Therefore, the use of the distributed and parallel data processing methods can significantly reduce the execution time.
The authors have designed, developed and deployed the portal to submit scientific payloads to heterogeneous computing infrastructure. As a basis technology PanDA [5] (Production and Distributed Analysis) Workload Management System has been chosen. The ATLAS experiment at LHC uses PanDA for managing the workflow for all data processing on over 140 data centers [6]. The authors have extended PanDA to be used beyond HEP and have tested the portal within bioinformatics domain. The portal combines Tier-1 Grid center, Supercomputer, and academic cloud at the National Research Center “Kurchatov Institute” (NRC KI) [7].

2. The portal architecture
PanDA as a basis technology of the portal delivers transparency of data and its 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 PanDA ideally suited for scientific data processing [8]. The generalized architecture of the portal at NRC KI can be schematically illustrated as shown in figure 1.

Figure 1. The portal generalized architecture.
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 that records comprehensive static and dynamic information on all jobs in the system. Relational databases implement the job queue and all metadata and monitoring repositories.
  - **Brokerage.** An intelligent module operates to prioritize and assign work on the basis of job type, priority, software availability, input data and its locality, 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 by taking priorities, resource allocation policy, and retry strategies into account.

- **Pilots.** Pilot jobs are used for acquisition of processing resources. Workload 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 pipeline in the system and perform monitoring functions.

- **Data management system.** Allows uniform access to data on distributed storages.

### 3. Software components of the Portal

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

In some cases a large number of computing tasks need to be defined and submitted. Thus the use of the web interface can be inconvenient. In such cases customized software applications or scripts can be used to create descriptions of the tasks dynamically. To integrate such applications with the portal the API providing calls for all required actions on tasks using HTTP requests was implemented. We have created data management system that allows to connect to the storages of different types without limitations of solutions built for HEP experiments. The system supports the file replication mechanism and ensure 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 (eg, Dropbox, Google Drive, etc.).

The portal scheme is shown in figure 2. The main components are:

- **GUI/Web service.** Graphical user interface and web service that provides a simple web interface to create a pipeline - global task definition. Also support of the API provided at this level.

- **Pipeline management tool.** Software component that supports the pipelines - several sequenced computational steps. The tool translates the pipeline definition to the sets of standard PanDA jobs corresponding the pipeline steps. 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 in the NRC KI
• Local Pilot Scheduler - pilot scheduler manages the pilots submission to available resources defining how many pilots to run on each resource.
• Resource gateway - interactive nodes that run pilot and remotely run payload to the target resource. For bioinformatics our portal instance in the NRC KI integrates two resources: Supercomputer and cloud computing platform.
• Resource Balancer - component for dynamic balancing of virtual machines within the cloud infrastructure 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. Our portal was connected with Lustre, NFS and local ftp storages.
• Data Transfer Node - the node used by Data Management System in order to transfer data from one storage to another and between storages and computing elements.

![Diagram](image)

**Figure 2.** The portal architecture at the NRC KI.

The following resources were integrated for the Portal at NRC KI:

• The supercomputing facilities at Kurchatov Centre for Collective Usage which include high performance cluster with peak performance of 122.9 TFLOPS. It operates since September 2011. It consists of 1280 nodes each with 2 CPUs of 4 cores (10240 cores total) interconnected with high throughput InfiniBand DDR network. Its total RAM is 20.5 TB and
local data storage has 144 TB in Lustre 2.0 parallel file system. Nodes are operated by SLURM batch system in CentOS Linux.

- OpenStack based academic cloud having performance 1.5 TFLOPS and providing 16 nodes, 256 cores, 512 GB RAM, 60 TB at storage system and InfiniBand connectivity.
- One of the largest Russian WLCG Tier-1 centers at NRC KI. The Tier-1 facility will process, simulate and store up to 10% of total data obtained from ALICE [9], ATLAS and LHCb[10] experiments.

4. Using the portal for Genome Sequencing Studies

The success of the projects to use PanDA beyond HEP (and Grid) has drawn attention from other compute intensive sciences such as bioinformatics [11]. Modern biology uses complex algorithms and sophisticated software, which is impossible to run without access to significant computing resources.

DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule. The advent of rapid DNA sequencing methods has greatly accelerated biological and medical research and discovery. Several new methods for DNA sequencing were developed in the mid to late 1990s and were implemented in commercial DNA sequencers by the year 2000. Together these were called the "next-generation" sequencing (NGS) methods. NGS platforms perform massively parallel sequencing, during which millions of fragments of DNA from a single sample sequenced in unison.

Special software pipelines are used to analyse a DNA. Pipelines include sets of software components which allow fast NGS data processing. PALEOMIX [12] is the one of the most popular pipeline developed by researchers from Ludovic Orlando’s group at the Centre for GeoGenetics, University of Copenhagen, Denmark.

In purpose of testing PALEOMIX pipeline the woolly mammoth sequenced DNA data analysis was conducted at the 80-cores server with 512 GB RAM. More than 350 gigabytes of data, containing more than 900 million paired reads, was processed. That took about two months. Besides, it took a large amount of manual routines to prepare data processing, some of them with the involvement of PALEOMIX experts. Specialized computing infrastructure was required and still it was time-consuming.

To hide execution complexity and manual routines from end-users we introduced original pipelines control system and seamlessly integrated it into the Portal. Also the authors took advantage of the fact that some of the PALEOMIX components can handle the input data fragments independently of the other fragments. The parallelization of the PALEOMIX pipeline execution can be implemented by dividing the input data into a large number of smaller files, then parallel processing of these files with final assembly the output data from the intermediate results. This approach of large data files processing, adopted and widely used in high-energy physics.

Every software pipeline that can be handled via the portal processes in several sequential computational steps as shown in figure 3. The system automatically (and without user prompt) handles input data of each step.

Pipeline workflow steps are the following:

- Pipeline initialization either via the web interface either via API
- Pre-processing. On the first step single PanDA job processes pipeline inputs in order to prepare data to parallel run.
- Step 2 consists of three parts. Special PanDA job splits input data into multiple files. Then each file is used as input for PALEOMIX in parallel mode on a distributed infrastructure. After that all the results are merged into a single output file within a special merge job.
- Post-processing. Step 3 is used to finish PALEOMIX pipeline.
- Outputs of the step 3 are transferred to user’s storage by data management system.
The developed portal has been successfully applied to the mammoth DNA analysis computing tasks. This approach allows:

- Run genome sequencing data analysis task using PALEOMIX pipeline on a heterogeneous distributed computing infrastructure. The size and number of files - the fragments of the original data - may be adjusted to use computing resources with maximum efficiency. File size is directly related to the required volume of RAM and processing time.
- Use for a single task various computing resources with different infrastructure. In particular, to run the task of mammoth DNA analysis supercomputer and academic cloud at NRC KI were used. In its original form the PALEOMIX task can’t run on Kurchatov supercomputer or on similar clusters because of the high requirements for RAM per core.
- Monitor and control the data and tasks automatically, reducing user interaction to a minimum. The approach provides the flexibility and reliability of computing. Authorized users interact with the system via the user-friendly interface.

The results showed that the software tools and methods for processing large volumes of experimental data, which have been developed initially for HEP in particular for experiments at the LHC accelerator, can be successfully applied in other scientific areas.

5. Conclusion
The aim of this study was to research the possibility of using the experience of distributed computing based on PanDA system in HEP to other scientific areas. To run the genome sequencing applications using developed portal we used the approach well known in high energy physics to handling large files: source file is cut into a large number of fragments of smaller size; each of the resulting file is treated as an independent computing job on different nodes of a supercomputer; the results of all jobs are combined into single output file. This approach allowed to significantly reduce the overall genome sequencing processing time from several weeks to a few days.

This method is most applicable when the task can be represented as follows. A set of common procedures on data is represented as software scripts, which transform the input data into outputs. Input data are the data files and software configurations files. The output of the script are processed data files. The task running process is consists of parallel and sequential execution of a subtasks - jobs, where each job corresponds to the transformation software script. In this paradigm the task can be adjusted to be executed on modern high-performance computing resources as a scope of jobs and successfully be finished within a reasonable period of time. This approach allows to ensure optimum utilization of the available computing resources (a supercomputer with a common shared access) in automated mode and quickly release them at the expiration time for the task.
The project showed that the software tools and methods for processing large volumes of experimental data, which have been developed initially for HEP in particular for experiments at the LHC accelerator, can be successfully applied in other scientific fields.

6. Acknowledgment
NRC KI laboratory team work is conducted with support from Ministry of Education and Science Russian Federation, contract № 14.Z50.31.0024 with partial support from RFBR (Russian Foundation for Basic Research) project № 16-37-00249 mol.a. Supercomputing resources at NRC-KI are supported as a part of the center for collective usage. Many thanks to PanDA SW team.

References
[1] Evans L, Bryant P 2008 LHC machine. *J. Inst.* 3 S08001.
[2] ATLAS Collaboration 2012 Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC, *Physics Letters B*, Volume 716, Issue 1, pp. 1–29.
[3] ATLAS Collaboration 2008 The ATLAS Experiment at the CERN Large Hadron Collider *J. Inst.* 3 S08003.
[4] The Worldwide LHC Computing Grid (WLCG). http://wlcg.web.cern.ch/LCG
[5] Maeno T 2011 Overview of ATLAS PanDA workload management *J. Phys.: Conf. Ser.* 331, 072024.
[6] Klimentov A, De K, Jha S, Maeno T, Nilsson P, Oleynik D, Panitkin S, Wells J and Wenaus T 2016 Integration Of PanDA Workload Management System With Supercomputers for ATLAS and Data Intensive Science. *J. Phys.: Conf. Ser.* 762, 012021
[7] Belyaev A et al. 2015 Integration of Russian Tier-1 Grid Center with High Performance Computers at NRC-KI for LHC experiments and beyond HENP. *J. Phys.: Conf. Ser.* 664, 092018
[8] Aamodt K et al. 2008 The ALICE experiment at the CERN LHC *J. Inst.* 3 S08002
[9] Alves A et al. 2008 The LHCb Detector at the LHC *J. Inst.* 3 S08005
[10] Maeno T, De K, Klimentov A, Nilsson P, Oleynik D, Panitkin S, Petrosyan A, Schovancova J, Vaniachine A and Wenaus T 2014 Evolution of the ATLAS PanDA workload management system for exascale computational science. *J. Phys.: Conf. Ser.* 513, Track 3
[11] De K, Klimentov A, Maeno T, Mashinistov R, Nilsson P, Oleynik D, Panitkin S, Ryabinkin E and Wenaus T 2016 Accelerating Science Impact through Big Data Workflow Management and Supercomputing. *EPJ Web of Conferences* Volume 108
[12] Schubert M, Ermini L, Sarkissian C D, Jonsson H, Ginolhac A, Schaefer R, Martin M D, Fernandez R, Kircher M, McCue M, Willerslev E, Orlando L 2014 Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat Protoc.*; 9:1056-1082.