Scientific Data Lake for High Luminosity LHC project and other data-intensive particle and astro-particle physics experiments

A Alekseev\textsuperscript{1,2,3}, A Kiryanov\textsuperscript{2,4}, A Klimentov\textsuperscript{5}, T Korchuganova\textsuperscript{1,2,3}, V Mitsyn\textsuperscript{6}, D Oleynik\textsuperscript{2,6}, A Smirnov\textsuperscript{2,7}, S Smirnov\textsuperscript{7} and A Zarochentsev\textsuperscript{2,8}

\textsuperscript{1} Compiler Technology Department, Ivannikov Institute for System Programming of the RAS, Moscow, 109004, Russia
\textsuperscript{2} Cloud technologies and Big Data analytics laboratory, Plekhanov Russian University of Economics, Moscow, 117997, Russia
\textsuperscript{3} ATLAS group, Universidad Andrés Bello, Santiago, 7550196, Chile
\textsuperscript{4} Department of computing resources and technologies, NRC “Kurchatov Institute” - PNPI, 188300, Gatchina, Russia
\textsuperscript{5} Physics Department, Brookhaven National Laboratory, Upton, NY 11973, USA
\textsuperscript{6} Laboratory of Information Technologies, Joint Institute for Nuclear Research, Dubna, 141980, Russia
\textsuperscript{7} Institute of Nuclear Physics and Engineering, National Research Nuclear University MEPhI, Moscow, 115409, Russia
\textsuperscript{8} Laboratory of Ultra-High Energy Physics, Saint Petersburg State University, Saint Petersburg, 198504, Russia

Abstract. The next phase of LHC Operations – High Luminosity LHC (HL-LHC), which is aimed at ten-fold increase in the luminosity of proton-proton collisions at the energy of 14 TeV, is expected to start operation in 2027-2028 and will deliver an unprecedented scientific data volume of multi-exabyte scale. This amount of data has to be stored and the corresponding storage system should ensure fast and reliable data delivery for processing by scientific groups distributed all over the world. The present LHC computing and data processing model will not be able to provide the required infrastructure growth even taking into account the expected hardware technology evolution. To address this challenge the new state-of-the-art computing infrastructure technologies are now being developed and are presented here. The possibilities of application of the HL-LHC distributed data handling technique for other particle and astro-particle physics experiments dealing with large-scale data volumes like DUNE, LSST, Belle-II, JUNO, SKAO etc. are also discussed.

1. Introduction

Modern scientific experiments generate hundreds of petabytes of data and the processing of those volumes requires totally new data management, distribution and storage approaches compared with the existing ones. A good example of such a project is one of the largest scientific instruments in the world – the Large Hadron Collider (LHC) \cite{1}, which operates at the CERN Laboratory in Geneva, Switzerland for already more than decade.
The other examples of modern experiments which are supposed to produce exabyte volumes of scientific data are the projects at the future NICA collider (Dubna, Russia) and the FAIR complex (GSI, Germany), Deep Underground Neutrino Experiment (DUNE) at Fermi Lab. (USA), the Jiangmen Underground Neutrino Observatory (JUNO) in China, Belle-II experiment (KEK, Japan), Large Synoptic Survey Telescope (LSST, Vera Rubin Observatory) – a new kind of optical telescope under construction in Chile. One should also mention the scientific projects beyond physics and astronomy – in molecular biology and bioinformatics (genom sequencing), in computational neurobiology (e.g. the project to create a digital brain model BlueBrain at EPFL in Switzerland).

2. Distributed computing software tools for High Energy Physics

The LHC experiments at CERN explore the fundamental nature of matter and the basic forces that shape our Universe, generating an unprecedented amount of physics data of an exabyte scale. To address this data processing challenge, the experiments are relying on the deployed computational infrastructure of the Worldwide LHC Computing Grid (WLCG) [2]. Geographically distributed more than 8000 scientists from 200 universities and laboratories in 42 countries analyze the LHC data searching for new discoveries. The four LHC experiments (ALICE, ATLAS, CMS and LHCb) lead WLCG resource usage in the number of completed jobs, processed data volume, and in the core-hours used for High Energy Physics (HEP) and Nuclear Physics (NP) experiments. The resources consumed by the experiments are distributed over more than 200 WLCG centers and amount to 1 million of x86 CPU cores, 1 EB of data storage (50:50 disk/tape). The data transfers between centers use network lines with the capacity up to 100 Gbps.

Since the start of LHC data taking, the experiments operate under conditions in which contention for computing resources among high-priority physics activities happens routinely. Scientific priorities in HEP present Big Data challenges requiring state-of-the-art computational approaches, and therefore, serve as drivers of an integrated computer and data infrastructure. These priorities include investigating properties of Higgs boson candidates in an attempt to better understand the origin of mass and search for new laws of physics [3]. To avoid potential shortfalls in projected grid resources, LHC experiments are actively using supercomputers and cloud computing resources as an important supplement to keep up with the rapid pace of data collection and to produce simulated events for the experiments which are too complex and require enormous computing resources to produce them on the WLCG.

During the preparation stage and the first two LHC data taking periods (2011-2018) the experiments have developed various software tools and packages to ease the everyday data analysis activities. Those tools included:

- Workflow Management – software tools to translate physicist requests into production tasks. ProdSys2 [4] is the one developed by ATLAS experiment
- Workload Management – tools for submission and scheduling of jobs and tasks (PanDA [5] by ATLAS, ALIEN [6] by ALICE and DIRAC [7] by LHCb, also used by Belle-II experiment
- Monitoring and Analytics – tools for infrastructure performance monitoring and accounting of production jobs and tasks, shares, users (BigPanDA [8], Kibana [9], MonaLisa, PerfSonar [10])
- Data Management – tools for data replication, bookkeeping and handling (Rucio [11])
- Information System – tools for queues and resources description (AGIS/CRIC [12])

This list illustrates that some efforts and technical decisions were duplicated among the experiments and it shows the lack of universality and the absence of common solutions that could be applied in different experiments.
The LHC Run1 (2009–2013) and Run2 (2015–2018) have convinced physicists that their developed software needs fundamental re-engineering to address the realities of future commodity, highly parallel processors, and that, if they can achieve this re-engineering, they could emerge with codes ready to exploit High Performance Computing facilities quite well. Explicitly targeting supercomputing in the re-engineering efforts is a tactic that will trickle down to benefit the Ethernet cluster approach that is used for the most HENP computing in the past two decades. LHC experiments have successfully combined High Performance Computers (HPC) and Leadership Class Facilities (LCF) with distributed computing and used HPCs for more than 5 years. Currently HPC facilities are integrated via different technologies because of the unique nature of HPCs and because of local/regional physics groups preferences. The U.S. Department of Energy, Office of Science, High Energy Physics and Advanced Scientific Computing (DOE ASCR and HEP) funded the BigPanDA project [13] which has provided the first important demonstration of the capabilities that a workload management system (WMS) can have on improving the uptake and utilization of supercomputers from both application and systems points of view.

The EuroHPC program has been launched in 2019 by the European Commission [14]. The EC plans to develop and to reinforce the European high-performance computing and data processing capabilities to achieve exascale capabilities by 2021-2023 and post-exascale facilities by 2026-27. Five petascale and three pre-exascale consortia have been built in Europe. The growth of HPC facilities in Europe, USA and Asia will increase the pool of users and competition for HPC resources. The next generation of HPCs is evolving from pure computational facilities to resources that can be used for extreme data processing (BigData, High Performance Data Analytics and Artificial Intelligence – Machine and Deep Learning). Those machines will also provide multi 100Gb/s external connectivity and data processing capacities of several TB/s.

There are a lot of commonalities and functional goals between different scientific domains:

- Converge cloud, grid and HPC services in one infrastructure;
- Overcome fragmentation between scientific domains and countries

The new HPCs can be coupled with scientific data lakes and it opens a new avenue for data processing and analysis. In addition there is a hot topic for HEP experiments related to non-x86 Intel architecture usage. HENP has already proven itself to be a successful early adopter of ML techniques at small scales. The recent boost in using machine learning algorithms in HENP for particle and event identification, energy estimation and pile-up suppression led to a huge interest from physicists to new architectures (such as GPU and TPU, both have been extremely useful for in speeding the training of complex deep learning models, Hyper Parameter Optimization and active learning) and to High Performance Computing (resource-rich many-core processors such as RISC, GPUs, and TPUs are vital to the optimization of the training time of most modern machine learning algorithms, including deep neural networks, generative adversarial networks, autoencoders, etc). The High Luminosity LHC run will begin operations in 2027/28 with expected data volumes to increase by at least an order of magnitude as compared with present systems. ATLAS computing and storage needs estimation are presented in figure 1 and figure 2, respectively. HL-LHC needs for ATLAS and CMS are above the expected hardware technology evolution (15% to 20%/yr). Extrapolating from existing trends in disk and compute pricing, and assuming fixed infrastructure budgets, the implications for end-user analysis of the data are significant. The LHC experiments CPU needs can be partially (opportunistically) covered by HPCs and commercial cloud computing providers, but for storage there is no solution like this, the risk is that the current system architecture and tools will not be able to provide the abstractions and data management capabilities to scale to meet the expected growth. This challenge cannot be solved by simply extending the current LHC computing model. New state-of-the-art technologies need to be applied and potentially developed, leveraging the investments
and research already being conducted in the commercial sector. As it is seen from the plots, only a set of ‘aggressive’ R&D projects may help to solve a shortage in storage and CPU needs for the HL-LHC era. In this paper we will describe just one of them, a “scientific data lake” R&D project.

![Graph showing projected CPU requirements of ATLAS experiment between 2020 and 2034.](image)

**Figure 1.** Projected CPU requirements of ATLAS experiment between 2020 and 2034 based on 2020 assessment. Three scenarios are shown, corresponding to an ambitious (“aggressive”), modest (“conservative”) and minimal (“baseline”) development program. The black lines indicate annual improvements of 10% and 20% in the computational capacity of new hardware for a given cost, assuming a sustained level of annual investment. The blue dots with the brown lines represent the 3 ATLAS scenarios following the present LHC schedule. The red triangles indicate the Conservative R&D scenario under an assumption of the LHC reaching in average 200 primary vertexes per one bunch crossing ($\mu$) in Run4 (2028-2030).

When developing a model for managing, processing and storing such volumes of scientific data, many factors should be taken into account – in particular, the availability and capabilities of high-speed networks, the development of microprocessor technologies and storage systems. The fundamental problem is the creation of a geographically distributed infrastructure – a data lake based on supercomputer centers, high-performance computing resource centers (grids) and cloud centers – that allows to deploy an environment in which the computing infrastructure looks logically unified for the end user. Until recently, the grid model was used as such an infrastructure, which was well suited for working with data that has volumes of up to tens of petabytes, but new algorithms and applications are required to process, analyze, and store exabytes of scientific experiments results (possibly based on machine learning methods – e.g. for automatic distribution of data among media), as well as non-relational databases for metadata storage.

One of the possible data processing and storage models is the federation of computing resources, or the so called data lake model.
Figure 2. Projected disk storage requirements of ATLAS between 2020 and 2034 based on 2020 assessment. Three scenarios are shown, corresponding to an ambitious (“aggressive”), modest (“conservative”) and minimal (“baseline”) development program. The black lines indicate annual improvements of 10% and 20% in the storage capacity of new hardware for a given cost, assuming a sustained level of annual investment. The blue dots with the brown lines represent the 3 ATLAS scenarios following the present LHC schedule. The red triangles indicate the Conservative R&D scenario under an assumption of the LHC reaching $\mu=200$ in Run4 (2028-2030).

3. High-Luminosity LHC R&D computing projects

The overarching common challenge for data intensive experiments is distributed data handling. The evolution of the computing facilities and the way storage will be organized and consolidated will play a key role in how the possible shortage of resources will be addressed by the LHC experiments. Technologies that will address the HL-LHC computing challenges may be applicable for other scientific communities (SKAO, DUNE, LSST, Belle-II, JUNO, etc.) to manage large-scale data volumes. To address HL-LHC distributed data handling challenge the experiments together with WLCG have launched several R&D projects. The major projects are listed below with a brief description.

**Data Lake.** The aim is to consolidate geographically distributed data storage systems connected by fast network with low latency. The Data Lake approach, under development and evaluation by DOMA, ESCAPE, Russian Data Lake [15] and others, foresees consolidating storage at a few federated regional centres, with multi-level caching (using e.g. Xcache) providing a content delivery network down to the CPUs. The Data Lake model as an evolution of the current infrastructure bringing reduction of the storage and operational costs.

**Data Carousel.** An ultimate goal is to use tape more effectively and actively in distributed computing context. By ‘data carousel’, we also mean an orchestration between workflow/workload management (WFMS), data management (DDM) and data archiving services whereby a bulk data processing campaign with its inputs resident on tape, is executed by staging and promptly processing a sliding window of X% of inputs onto buffer disk, such that only X% of inputs are pinned on disk at any one time.

**Data management across Hot/Cold storage.** HEP needs to manage its data samples across
a variety of mediums based on the types of data and its uses: from tape (or other cold storage technologies) to disk (spinning and ssd), to caches (including world wide accessed data in clouds and “data lakes”) to content delivery networks. Data placement and data access decisions can dramatically impact computing efficiency and costs. Data management across hot/cold storage is one of the tracks of a joint HEP-Google R&D project [16].

HEP Data formats and Input/Output optimization for distributed object storage. Optimized Input/Output is a topic of interest across the sciences. US ATLAS, CERN and Google conduct a R&D to explore whether HEP experiments may benefit from data stored as objects, and not necessarily ROOT trees or ROOT objects [17]. For example HDF5, widely used for numerical datasets, is itself developing a cloud and HPC service that works with Kubernetes and is a cloud/HPC agnostic. A key issue here is that object storage systems have high latency for access, making HEP data analysis use cases inefficient.

Intelligent Data Delivery Service (iDDS). The intelligent data delivery system will deliver events as opposed to delivering bytes. This allows an edge service to prepare data for production consumption, the on-disk data format to evolve independently of applications, and decrease the latency between the application and the storage. The first implementation of iDDS has been demonstrated by ATLAS in April-May 2020 for Data Carousel and active ML workflows.

Operations Intelligence. Reduce the HEP experiments computing operations effort by exploiting anomaly detection, time series and classification techniques to help the operators in their daily routines, and to improve the overall system efficiency and resource utilization.

4. Data lake prototype in Russia
In this section we will present results of Data Lake conceptual design. In particular the Data Lake concept implementation as a geographically distributed storage system using resources of three Russian research centers shown on figure 3: JINR (Dubna), PNPI (Gatchina) and MEPhI (Moscow). A peculiarity of this study is the absence of a dedicated high-speed network backbone between these centers. Each center is connected to Wide Area Network (WAN) with a 10 Gbps line and by different network providers, it introduces a significant fluctuations in the characteristics of network bandwidth.

We have addressed two different approaches when we built a distributed storage infrastructure. In the first approach a group of computing sites uses a shared storage system located at one of them and a particular data caching system deployed at other sites. Data

![Figure 3. Russian Data Lake Phase 2.](image)
Caches should optimize read operations performance. This approach is targeted on read-oriented workloads, such as physics analysis (figure 4).

The second approach examines a geographically distributed shared storage system with disk pools at each site and a common management server. This approach is targeted on write-oriented workloads, such as data processing or derivation production (figure 5).

**Figure 4.** Data Lake with site-local caching.

**Figure 5.** Distributed storage system on geographically dispersed pools.
The primary goal of this project is to build a storage system prototype to cope with the requirements for high-throughput data processing in a distributed computing environment for modern scientific research, such as experiments at the HL-LHC and other planned research in the framework of megascience-class projects. As part of this study, workable prototypes of distributed storage systems are created using dedicated but limited resources of the participating institutions. Then a number of tests are carried out on deployed prototypes, both synthetic, developed specifically for this kind of research, and real-life using production software of LHC experiments together with the HammerCloud testing system developed within the WLCG project. These tests are targeted at measuring the increased efficiency of using the computing resources of sites with reduced data I/O overhead. For read-oriented caches three possible ways of cache organization were examined:

- A dedicated cache server with shared access from all compute nodes
- A shared cache built on disk resources of compute nodes with a dedicated management node
- Individual non-shared caches on each compute node

Obtained results have demonstrated that the efficiency improvement depends on the organization of the cache, and it was observed for the cases when a common cache is used for all compute nodes, both on the basis of a dedicated server, and for the variant with a shared cache built on the resources of compute nodes. Individual non-shared caches on each compute node did not show any significant increase in efficiency.

The next stage of this study is building a prototype of a distributed storage system based on EOS technology where disk pools will be deployed on each participating site and combined into a common storage space. The possibility of distributing files across disk pools in the most efficient way in terms of data access will be investigated.

Acknowledgments
We would like to thank our colleagues I. Bird, S. Campana, X. Espinal and S. Jezequel for valuable contributions and fruitful discussions during our work on this project.

This project in Russia was supported by the Russian Science Foundation grant No. 19-71-30008 (the research is conducted in the Plekhanov Russian University of Economics in Moscow).

References
[1] Evans L and Bryant P 2008 J. Instrum. 3 S08001
[2] Shiers J 2007 Comput. Phys. Commun. 177 219
[3] Aad G et al. 2012 Phys. Lett. B 716 1
[4] Borodin M et al. 2015 J. Phys. Conf. Ser. 664 062005
[5] Maeno T 2008 J. Phys. Conf. Ser. 119 062036
[6] Saiz P et al. 2003 Nucl. Instr. Meth. A 502 437
[7] Tsaregorodtsev A et al. 2008 J. Phys. Conf. Ser. 119 062048
[8] Padowski S et al. 2018 J. Phys. Conf. Ser. 1085 032043
[9] Alekseev A et al. 2018 J. Phys. Conf. Ser. 1015 032003
[10] Vicinanza D et al. 2012 J. Phys. Conf. Ser. 396 042060
[11] Barisits M et al. 2019 Comput. Softw. Big Sci. 3 11
[12] Anisenkov A et al. 2019 EPJ Web Conf. 214 03003
[13] Klimentov A et al. 2015 J. Phys. Conf. Ser. 608 012040
[14] URL: https://www.etp4hpc.eu/pujades/files/Blueprint$%$20document$%$20190904.pdf
[15] Klimentov A, Kiryanov A and Zarochentsev A 2018 Open Systems, DBMS 26
[16] Barisits M et al. (ATLAS and Google R&D Collaborations) 2019 EPJ Web Conf. 214 04020
[17] Lassnig M, De K and Klimentov A 2017 Preprint ATL-SGFT-PUB-2017-002