BNL-114326-2017-CP

BigData and Computing Challenges in High Energy and Nuclear Physics

A. Klimentov

Submitted to the Journal of Instrumentation

September 2017

Physics Department

Brookhaven National Laboratory

**U.S. Department of Energy
USDOE Office of Science (SC),
High Energy Physics (HEP) (SC-25)**

Notice: This manuscript has been authored by employees of Brookhaven Science Associates, LLC under Contract No. DE- SC0012704 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.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or any third party's use or the results of such use of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof or its contractors or subcontractors. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

INSTR17. INSTRUMENTATION FOR COLLIDING BEAM PHYSICS
FEB 27 - MAR 3, 2017
NOVOSIBIRSK, RUSSIA

BigData and Computing Challenges in High Energy and Nuclear Physics

A.Klimentov^{a,b,1} M.Grigorieva^{b,c} A.Kiryanov^{b,d} and A.Zarochentsev^{b,e}

^a*Brookhaven National Laboratory,
Upton USA*

^b*National Research Center "Kurchatov Institute",
Moscow, Russia*

^c*Tomsk National Polytechnic University,
Tomsk, Russia*

^d*Petersburg Nuclear Physics Institute of NRC KI,
Gatchina, Russia*

^e*St.Petersburg State University,
St.Petersburg, Russia*

E-mail: alexei.klimentov@cern.ch

ABSTRACT: In this contribution we discuss the various aspects of the computing resource needs experiments in High Energy and Nuclear Physics, in particular at the Large Hadron Collider. This will evolve in the future when moving from LHC to HL-LHC in ten years from now, when the already exascale levels of data we are processing could increase by a further order of magnitude. The distributed computing environment has been a great success and the inclusion of new super-computing facilities, cloud computing and volunteering computing for the future is a big challenge, which we are successfully mastering with a considerable contribution from many super-computing centres around the world, academic and commercial cloud providers. We also discuss R&D computing projects started recently in National Research Center "Kurchatov Institute"

KEYWORDS: Computing, Data processing Methods, Software Architectures

¹Corresponding author.

Contents

1	Scale of Computing Needs for High Energy and Nuclear Physics	1
1.1	Importance of Supercomputers for High Energy and Nuclear Physics Scientific Program	3
2	HENP Computing R&D Projects	4
2.1	Federated Data Storage	4
2.2	Data Knowledge base for High Energy and Nuclear Physics Experiments	5

1 Scale of Computing Needs for High Energy and Nuclear Physics

The largest scientific instrument in the world - the Large Hadron Collider (LHC) [1] operates at the CERN Laboratory in Geneva, Switzerland. The experiments at the LHC explore the fundamental nature of matter and the basic forces that shape our universe. To address an unprecedented multi-petabyte data processing challenge, experiments are relying on the deployed computational infrastructure of the Worldwide LHC Computing Grid (WLCG) [2]. More than 6000 scientists from 200 universities and laboratories in 45 countries analyze the LHC data in search of new discoveries. ATLAS and ALICE experiments [3][4] lead WLCG resource usage in the number of jobs completed, processed data volume, and in the core-hours used for High Energy and Nuclear Physics. Since the start of LHC data taking, ATLAS and ALICE operate under conditions in which contention for computing resources among high-priority physics activities happen routinely. Both experiments will collect a factor of 10-100 more data during the next 3-5 years. Scientific

Figure 1. LHC computing needs.

priorities in High Energy (HEP) and Nuclear Physics (NP) present Big Data challenges requiring state-of-the art computational approaches, and therefore, serve as drivers of integrated computer

and data infrastructure. For HEP, 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 [5]. For NP, these priorities include investigating properties of the Quark Gluon Plasma. To avoid potential shortfalls in projected LHC Grid resources, ATLAS and ALICE are actively evaluating supercomputing- scale resources as an important supplement to keep up with the rapid pace of data collection and to produce simulated events for LHC experiments which are too complex and require enormous computing resources to produce them on the WLCG.

The LHC Run1 (2009-2013) and the first years of LHC Run2 (2015-2018) have convinced physicists that their codes need 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 and Supercomputing 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 now used for most HENP computing. The DOE ASCR and HEP funded the BigPanDA project [6] as 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. ATLAS and ALICE experiments have expressed an interest in the common projects how supercomputers can be used and how a new federated resources paradigm can be set up for geographically distributed centers. The ATLAS experiment at the CERN LHC uses a geographically distributed grid of approximately 150,000 cores continuously (250 000 cores at peak), (over 1,000 million core-hours per year) to simulate, and analyze its data (today's total data volume of ATLAS is more than 260 PB). After the early

Figure 2. Relative CPU consumption on the Grid of the main computational activities in ATLAS during LHC Run1.

success in discovering a new particle consistent with the long awaited Higgs boson, ATLAS is starting the precision measurements necessary for further discoveries that will become possible by much higher LHC collision energy and rates from Run2. The need for simulation and analysis will overwhelm the expected capacity of WLCG computing facilities unless the range and precision of physics studies will be curtailed. Even today, important analyses that require large simulated samples, are expected to have to wait for many months before WLCG existing Grid resources will be able to provide the simulated data they need. In addition, some physical processes of interest at LHC are nearly impossible to calculate on traditional grid resources because of extremely large computing requirements. Supercomputers offers unique opportunity to produce such datasets

through massive parallelization. Furthermore, generated events can be shared with theoreticians and scientific collaborations worldwide, because they are of general interest and not ATLAS or ALICE specific. Figure 1 shows LHC experiments computing needs for the High-Luminosity runs. Moving forward, this type of backlog is expected to grow larger and must be addressed for continued good progress of the field and responsible exploitation of the accelerator and detector. Figure 2 shows that Monte-Carlo (MC) simulation consumes 42% of ATLAS Grid resources [7], MC simulation will be a very well-suited supercomputers payload for all HENP experiments. It is the largest fraction of ATLAS and ALICE computing, is critically important for physics results, and well suited for opportunistic execution on supercomputers.

1.1 Importance of Supercomputers for High Energy and Nuclear Physics Scientific Program

It is now accepted that ATLAS and ALICE simulation and analysis applications must undergo a transformation to be able to make effective use of current and future computer architectures. Only by using processing hardware efficiently will experiments be able to achieve the massive computing throughput that will be needed in the next decade. As an example of the approach, the decomposition of an event simulation into tens of thousands of small steps will provide an opportunity to map the code on to a wide range of architectures. Supercomputers are a particular target because of the massive computing power they can bring to bear, and because they provide the lowest possible communications overhead for massively decomposed tasks. Fully optimized use of current and emerging supercomputing facilities by HENP applications is a long-term goal, certainly requiring work of five or more years. Early access to supercomputers and the demonstration described earlier will provide an important stimulus to this work. Access to facilities coupled with collaborative help in the transformation of HENP code would be a major scientific contribution to the physics discoveries of the next ten years. All of these challenges require large scale computing capacities. Looking for new signatures of fundamental physics requires comparison between observation and predictions of new theories: between ensembles of real events and corresponding ensembles of simulated Monte-Carlo events under various theoretical models. In addition, extensive simulation of all known physics processes are required to draw statistical conclusions. Because of the need for simulations, ATLAS is often limited not by its measurement abilities, but by the ability to analyze the collected data. Additionally, there are limitations on our ability to simulate comparison events: limitations on the total number of events, on their complexity, and in some cases limitations in simulating fully the theoretical models.

ALICE physicists are investigating multiple properties of the Quark Gluon Plasma, it is one of the top priorities of Nuclear Physics. In pursuit of this goal, Monte Carlo production dominates ALICE offline computing resources requirements. ALICE Geant4 - based Monte Carlo production of Pb+Pb collisions and p+Pb collisions can be produced on supercomputers at scale. To integrate supercomputers with the WLCG distributed computing we developed and implemented a new capability to collect information about unused worker nodes on supercomputer Titan (OLCF - Oak Ridge Leadership Class facilities), and based on that information adjust workload parameters to fill free resources. Proof-of-concept tests of this mechanism, executed over a few days, achieved increased system utilization levels from 90% to 93% (14.3% of free cycles) during the performance period of this test, as well as provided short wait times to ATLAS and ALICE for jobs submitted to Titan; all with no negative impact on OLCF's ability to schedule large, leadership-class jobs. Titan

was fully integrated with the ATLAS production system and now the ATLAS experiment routinely runs Monte-Carlo simulation tasks there. All operations, including data transfers to and from Titan, are transparent to the ATLAS Computing Operations team and physicists.

To conclude, HENP applications not only require large amounts of CPU-intensive calculations, but also new capabilities that only the supercomputing facilities can provide. Storage of ATLAS data, currently at over 260 petabytes of active data worldwide, increasingly strains against the limits of manageable and analyzable data. From the usage information above, it becomes clear that supercomputing contributions of the order of 10 Million or more core hours per year become important and valuable. LHC computing can also be a close to ideal "crack-filling" application. The ATLAS production management system is being upgraded to make it aware of dynamically changing resources [8] and thus able to exploit groups of processors that become available for relatively short times. Storage capacity on the petabyte scale will be very valuable when available, but simulation, the largest user of CPU power, can be effectively performed on facilities offering only transient storage.

2 HENP Computing R&D Projects

To address LHC computing challenges beyond Run2, we've started several R&D projects namely Federated Data Storage to provide a consistent access and to federate isolated grid storage elements, and Data Knowledge Base, to bring a consistency to HENP experiments meta-data handling.

2.1 Federated Data Storage

Figure 3. Federated data storage centers and technology choice

In 2015 in the framework of the Laboratory "BigData Technologies for mega-science class projects" in NRC "Kurchatov Institute" a work has begun on the creation of a united disk resource federation for geographically distributed data centres, located in Moscow, St.Petersburg, Dubna, Gatchina (all above centres are part of the Russian Data Intensive Grid - RDIG of WLCG) and Geneva, its integration with existing computing resources and provision of access to this resources for applications running on both supercomputers and high throughput distributed computing systems (Grid). The objective of these studies was to create a federated storage system with a single

access endpoint and an integrated internal management system. With such an architecture, the system looks like a single entity for the end user, while in fact being put together from geographically distributed resources. The system should possess the following properties:

- fault tolerance through redundancy of key components;
- scalability, with the ability to change the topology without stopping the entire system;
- security with mutual authentication and authorization for data and metadata access;
- optimal data transfer routing, providing the user direct access to the closest data location;
- universality, which implies validity for a wide range of research projects of various sizes, including, but not limited to the LHC experiments.

So far as EOS and dCache storage systems are tested and used by all four major LHC experiments, the authors decided to start with these two as a software platforms. The first prototype of geographically distributed federated data storage comprising of CERN and RDIG centres has been set up in 2016. It was demonstrated that such storage system can be efficiently used for data processing and analysis by the LHC scientific applications. We have performed synthetic and real-life application tests from ATLAS and ALICE experiments, which have shown a reasonably high performance of the federation mostly limited by the local disk speeds and the bandwidth of network connections. Synthetic tests have also shown the expected dependency of metadata access speed on the speed of access to the federation entry point that already speaks in favour of the proposed system, since metadata access times are not bound by the location and availability of the storage servers. Also, file I/O speeds correlate reasonably with the throughput of the network channels. Further research is being carried on to properly compare the performance and features of both exploited storage systems, as well as a more ambitious experiment of mixing different storage systems in a single federation.

2.2 Data Knowledge base for High Energy and Nuclear Physics Experiments

Data Knowledge Base for HENP (DKB) project was started in 2016 by joint team from Kurchatov Institute, Tomsk Polytechnic University and ATLAS experiment at CERN. The motivation of this project was the idea of providing a holistic coherent view on the whole process of physics analysis: from physicist idea, to available data, and the development the knowledge-based infrastructure, providing fast access to relevant scientific information, which is currently scattered around different services and data sources. To verify this assumption, a series of ATLAS collaboration metadata studies have been conducted. The results of these studies led to the conclusion that the main goal of the DKB will be to organize metadata in ATLAS, so as to provide a holistic view on physics topics, including integrated representation of all ATLAS documents (papers, drafts, supporting documents, conference notes, Indico meetings, Twiki pages, etc) and corresponding data samples (real data, Monte-Carlo datasets, containers). Though the initial motivation was for ATLAS, the similar approach is applicable for HENP experiments in general. There are three primary metadata sources for ATLAS data:

1. Physics Topics - metadata, describing some physics topic attributes, like project, data and MC production campaigns, LHC energy and luminosity, data taking periods, software releases and other production configuration attributes.
2. Documents - metadata sources, related to ATLAS documents, including papers, drafts, supporting documents, conference notes, meetings notes. These metadata are stored in CERN Document Server, GLANCE, Indico, Twiki pages, HEP InSPIRE.
3. Data Samples Catalogs - are stored in Distributed Data Management system - Rucio, ATLAS Metadata Database - AMI, Production System database.

Figure 4. Data Knowledge Base architecture prototype.

All these data sources exist independently. Some metadata are integrated, but in general scientists need to obtain cross relations among metadata within each category and among categories by themselves. DKB is designed to fill this gap in metadata integration and is considered to look for cross references among the metadata, stored in various data sources.

Current DKB architecture (figure 4) is based on Semantic Web and hybrid metadata integration. Hybrid integration is supposed to combine consolidation of unchangeable metadata about some global objects within Virtuoso RDF storage, with the virtual integration of changing and auxiliary metadata. The global integrating data schema for DKB is designed in OWL format and contains the following main classes: Documents, Data Samples, Physics Topics and Authors. Documents are organized in chains with Scientific Paper in the beginning. Each Paper has underlying Supporting Notes, from which the datasets and other necessary meta information can be extracted by means of parsing the full text of the document. This allows to expand the document's chain by adding references to the corresponding Data Samples/Containers and Physics Topics attributes.

Metadata consolidation is provided by a set of tools for metadata extraction, conversion and import, executed in a certain sequence, organized as a dataflow. Automation of this dataflow is supported by Apache Kafka Streams.

To provide virtual metadata integration of various metadata sources, each of them must be adapted

to Semantic Web data representation using SPARQL endpoints, which will allow executing federated requests via SPARQL requests.

To visualize integrated metadata DKB uses a JavaScript library Ontodia, that allows to visualize, navigate and explore the data in the form of an interactive graph based on underlying data sources. Implementation of the DKB is closely related to the datasets discovery and is supposed to be capable of automatic acquisition of knowledge from heterogeneous and distributed sources, including archives of scientific papers, research groups wiki pages, tutorials, conference and workshop information, and link these information with well-structured technical data about the experiment (datasets, analysis code, various metadata information on all used signal and background samples).

Acknowledgments

This work was funded in part by the Russian Ministry of Science and Education under Contract No. 14.Z50.31.0024 and the U. S. Department of Energy, Office of Science, High Energy Physics and Advanced Scientific Computing under Contracts No. DE-AC02-98CH10886 and DE-AC02-06CH11357, the federated storage project is funded by the Russian Fund for Fundamental Research under contract 15-29-0794_2 ofi_m and the data knowledge base project is funded by the Russian Science Foundation under contract N16-11-10280. 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] LHC - The Large Hadron Collider. <http://lhc.web.cern.ch/lhc/>.
- [2] Jamie Shiers. The worldwide LHC computing grid (worldwide LCG). Computer physics communications, 177:219-223, 2007.
- [3] ATLAS Collaboration, G Aad, et al. The ATLAS experiment at the CERN large hadron collider. J. Instrum, 3:S08003, 2008.
- [4] Kenneth Aamodt, A Abrahantes Quintana, R Achenbach, S Acounis, D Adamova, C Adler, M Aggarwal, F Agnese, G Aglieri Rinella, Z Ahammed, et al. The ALICE experiment at the CERN LHC. Journal of Instrumentation, 3(08):S08002, 2008.
- [5] Georges Aad, T Abajyan, B Abbott, J Abdallah, S Abdel Khalek, AA Abdelalim, O Abdinov, R Aben, B Abi, M Abolins, et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. Physics Letters B, 716(1):1-29, 2012.
- [6] A.Klimentov et al. Next generation workload management system for big data. In 16th International Workshop on Advanced Computing and Analysis Techniques in Physics Research. plenary talk. September 2014. Prague, Czech Rep.
- [7] G.Stewart. Evolution of computing and software at LHC : from run2 to HL-LHC. Computing in High Energy and Nuclear Physics Conference. April 2015, Okinawa, Japan., 2015.
- [8] M.Borodin, K.De, J.Garcia Navarro, D.Golubkov, A.Klimentov, T.Maeno and A.Vaniachine, Scaling up ATLAS production system for the LHC Run 2 and beyond : project ProdSys2. J.Phys.Conf.Ser. 664 (2015) no.6, 062005.