



REPORT ON HEP APPLICATION AND IMPLEMENTATION

TECHNICAL DESCRIPTION OF HIGH ENERGY PHYSICS APPLICATION AND IMPLEMENTATION

Document Filename: **BG-DNA3.3-v1.1-VU-HEPApplicationImplementation.doc**
Activity: **NA3**
Partner(s): **VU, IFJ PAN, NICPB, ITPA**
Lead Partner: **VU**
Document classification: **PUBLIC**

Abstract: The BalticGrid deliverable DNA3.3 presents the results achieved by implementing project's milestone MNA3.7 "Test site for HEP application operational". The document aims to give a technical description of applications implemented, namely: Compact Muon Solenoid application and Large Hadron Collider Beauty application (both applications following experiments of CERN). These applications are followed by statistical Monte Carlo data analysis and have the task to validate and demonstrate a successful scientific use of grid technology in HEP area. The implementation of them gives a powerful tool for scientists from Baltic States. The plan for future involves constant support and development of such applications, adopting software and algorithms according to changes in EGEE, introducing suitable graphical user interfaces and performance tools, establishing necessary new VOs.





REPORT ON HEP APPLICATION AND IMPLEMENTATION
Technical description of application and plans for support and
Special Interest Groups tasks

Document review and moderation

	Name	Partner	Date	Signature
Released for moderation to				
Approved for delivery by				





Document Log

Version	Date	Summary of changes	Authors
0	28/10/2006	Draft version	Algimantas Juozapavicius, Mario Kadastik
1	30/01/2006	Draft version	Algimantas Juozapavicius, Mario Kadastik, Mariusz Witek, Michal Krasowski, Tadas Meskauskas, Eduardas Kutka, Daiva Kaukeniene
1.1	31/01/2006	Final version	Algimantas Juozapavicius, Mario Kadastik, Mariusz Witek, Janis Dzerins, Eduardas Kutka



CONTENTS

1	INTRODUCTION.....	5
1.1	PURPOSE OF THE DOCUMENT	5
1.2	GOALS OF APPLICATIONS.....	5
1.3	ABBREVIATIONS.....	5
2	EXECUTIVE SUMMARY.....	7
3	HIGH ENERGY PHYSICS APPLICATIONS IN BALTICGRID.....	9
3.1	CMS (COMPACT MUON SOLENOID) APPLICATION	10
3.1.1	<i>Simulation of CMS events</i>	<i>10</i>
3.1.2	<i>The structure of CMS software.....</i>	<i>12</i>
3.1.3	<i>Implementation of CMS jobs.....</i>	<i>12</i>
3.2	LHCb (LARGE HADRON COLLIDER BEAUTY) APPLICATION.....	18
3.2.1	<i>Requirements and computing procedure.....</i>	<i>18</i>
3.2.2	<i>Prerequisites and software.....</i>	<i>18</i>
3.2.3	<i>Statistics</i>	<i>19</i>
3.2.4	<i>Statistical data analysis using Monte Carlo methods</i>	<i>21</i>
3.2.5	<i>Importance and Requirements</i>	<i>21</i>
3.2.6	<i>Computing procedure and software.....</i>	<i>22</i>
4	CONCLUSIONS AND FUTURE WORK.....	24
4.1	ACHIEVEMENTS:.....	24
4.2	ACTIONS PLANNED FOR THE FUTURE	24
4.2.1	<i>Constant support of HEP applications implemented</i>	<i>24</i>
4.2.2	<i>Strategy for establishing new VOs</i>	<i>25</i>



1 INTRODUCTION

The Baltic Grid project aims i) to develop and integrate the research and education computing and communication infrastructure of the Baltic States into emerging European Grid infrastructure, ii) to bring the knowledge in Grid technologies and use of Grids in the Baltic States to a level comparable to that in EU member states with a longer experience in the development, deployment and operation of Grids, and iii) to further engage the Baltic States in policy and standards setting activities. The integration of the Baltic States into the European Grid infrastructure will primarily focus on extending the EGEE to the Baltic States.

The Baltic Grid project implements applications from High Energy Physics, as pilot ones. This is of high strategic importance for the Baltic States and it is designed to give a rapid involvement in EGEE grid infrastructure, enabling the new member states to participate faster in the European Research Area.

The Baltic Grid is taking advantage of design and implementation of such most suitable applications as well as integrating computing infrastructure of academic institutions of the Baltic States. The goal of this deliverable is to present suitable applications developed in the Baltic Grid project, and to provide analysis of technical aspects of such applications.

1.1 PURPOSE OF THE DOCUMENT

The purpose of this document is to present results of conceptual and technical analysis done for applications from HEP, while implementing milestone MNA3.7 “Test site for HEP applications operational”. This description helps to identify the most suitable usage and services for users of BalticGrid, adjusting them to computing infrastructure in the best possible way. It presents also a plan for applications’ support and for adaptation and development of them in the future.

1.2 GOALS OF APPLICATIONS

The main objective of HEP applications “CMS application” and “LHCb application” in Baltic Grid project is to help scientists and other users to be involved in the research work in elementary particle physics by using grid technologies, giving them wide area of computing power and tools, adopting applications to the level of tools of their partners in more advanced EU countries. Realization of these goals lead to formation of suitable VOs, as well as to close collaboration between application developers and grid experts, providing analysis for the deployment and run of applications, providing graphical user interface and performance analysis tools.

1.3 ABBREVIATIONS



AOD – Data Format for Event Content Definition

BG – Baltic Grid

BI – Bioinformatics

CMS – Compact Muon Collider

ESD – file format of Event Summary Data

G-PM – Grid Performance Measurement tool

HEP – High-Energy Physics

LHC – Large Hedron Collider

LHCb – Large Hadron Collider beauty

MD – Migrating Desktop

OCM-G – OMIS-Compliant Monitor for the Grid

SUP – Application Support



2 EXECUTIVE SUMMARY

The BalticGrid project implements pilot applications from HEP, having the task to validate and demonstrate a successful scientific use of grid technology established and to create a powerful tool for scientists from Baltic States, namely:

- CMS application (following the Compact Muon Solenoid experiment from CERN),
- LHCb application (based on Large Hadron Collider Beauty experiment, CERN).

These tasks were decided to select and implement because of the critical importance of Grids to the community of scientists from Baltic States, and because of their relative maturity. HEP applications are well developed and most matured within EGEE, and are critical for success of scientists from Baltic countries participating in CERN experiments or doing research in elementary particle physics, as well as in other related areas.

It follows from the analysis, made to study research and computing needs of Baltic States' scientists in the time period of writing the proposal for BalticGrid project and while the work in the first months of the project (deliverable DNA3.1), that the experiments of [CMS](#) (the Compact Muon Solenoid Experiment) and [LHCb](#) (The Large Hadron Collider Beauty Experiment), two from the four different LHC experiments, are of special value to BG community.

The CMS application is of interest for scientists from Baltic States, already involved in this experiment, and/or to scientists going to join this experiment very soon. Compact Muon Solenoid applications are used to generate Monte Carlo datasets, to simulate events with the particle detector, to reconstruct the particles and to analyze sequences of events and particles. The software was implemented and installed by developing the procedure of joining the corresponding CMS VO of EGEE.

The LHCb experiment is well suited for the purposes of BalticGrid infrastructure testing. Like CMS application, it helps to simulate particle events in Large Hadron Collider too. Full simulation software for LHCb experiment is producing Monte Carlo data in the same form as a real collider detector and processing them by a reconstruction program. The output data is a collection of events written in the format as expected from detector electronics (Raw Data) or in the processed form of Event Summary Data. While it deals with such specific data analysis methods and data flow/streaming procedures, it gives a possibility to formulate and use quite understandable measure for evaluating the efficiency of computing aspects of BalticGrid infrastructure.

Statistical data analysis is specialized for each experiment and is based mainly on the Monte Carlo technique. It is applied to estimate the expected errors or values of the real measurement by generating large numbers of simplified experiments and by looking to the



distribution of final results. The procedure requires submitting a large number of CPU intensive jobs, each consuming a few hours on an average CPU unit. It is the grid technology that enables to provide the answer in a reasonable period of time. The final application generates files containing input parameters, submitting jobs for each input file, checking if jobs are finished, collecting results and producing the final processing.

The BalticGRID deliverable DNA3.3 presents the results achieved by implementing project's milestone:

- MNA3.7 "Test site for HEP applications operational" - to implement and test applications suggested.

The HEP applications are implemented by supporting the existing VOs in EGEE, such as the CMS and LHCb VOs. The statistical Monte Carlo data analysis, following these applications, especially LHCb, has a potential to be developed further, and to attract additional users. If a suitable analysis area introduced, it has a potential to be transformed into a separate SIG. The implementation and operational of pilot HEP applications gives following benefits to the users from Baltic States:

- they introduce a knowledge to Baltic users, which allows them to interact efficiently with computing and research infrastructure of other European countries,
- they target a wide range of computing needs,
- they give large data transfer and throughput computing needs of the LHC experiments (of special value are CMS experiments for users from the Baltic States, as well as LHCb experiments for Baltic region users)

The technical and operating environment of HEP applications have the following development features:

- most of applications are using open source code, which is common practice in many grids
- programming platforms in use are C++ and Fortran, together with popular subroutine libraries
- many applications are indifferent to the computer architecture (32-bit and/or 64-bit), like statistical software, nevertheless some applications, like CMS and LHCb experiments, still need x86 32-bit computer architecture

There are plans for constant support of HEP applications in the future. Scientists from the Baltic States will take major advantages by running their applications in a distributed computing environment, if the applications will have always a status up-to-date, their software and architecture will be constantly renewed and updated.

The application support expert group will maintain a close communication between application developers and grid experts to speed-up application adaptation, providing analysis for the deployment/run of applications in a grid environment.



3 HIGH ENERGY PHYSICS APPLICATIONS IN BALTICGRID

The BalticGrid project, in the first year of operation, initiated two pilot applications with the task to validate and demonstrate a successful scientific use of grid technology established and to create a powerful tool for scientists from the Baltic States. These pilot applications are related to topics from high-energy physics (HEP), namely Compact Muon Solenoid and Large Hadron Collider Beauty, experiments from CERN, as well as topics of statistical analysis of data of nuclear and sub-nuclear physics, production of Monte Carlo samples and distributed data analysis.

High Energy Physics is a natural user domain for Grid computing, due to the very large amounts of data it produces and the dispersed communities who need to analyse it. Therefore, the HEP community was selected as pilot application area for BalticGrid, guiding the implementation of the evolving EGEE grid infrastructure, and remains one of the key user groups running jobs on the BG infrastructure.

The Baltic States involvement with HEP community of CERN, and more precisely with EGEE is important for scientists, engaged in fundamental scientific research. It is important for their future involvement with ERA too. The research and development interest comes through the computing needs of the Large Hadron Collider (LHC), a new particle accelerator under construction by CERN, the European Organization for Nuclear Research, and scheduled to start operation in 2007. This massive machine, set to be the largest scientific instrument in the world, will produce an unprecedented 15 petabytes of useful data per year, equivalent to a stack of CDs 22 kilometers high (more than 23 million CDs).

The experiments of [CMS](#) (the Compact Muon Solenoid Experiment) and [LHCb](#) (the Large Hadron Collider Beauty Experiment), being two from the four LHC experiments, are of special value.

There is growing number of scientists in the Baltic States, already involved in CMS experiment. Nowadays Compact Muon Solenoid applications are used to generate Monte Carlo datasets, to simulate these events with the detector, to reconstruct particles and to analyze results of reconstruction. CMS application is implemented by joining the corresponding VO of EGEE. It helps scientists from Baltics to stay in the frontier of nowadays achievements and to use most modern tools in their fundamental research.

The LHCb experiment, beside fundamental research, is at the moment used for the purposes of testing BalticGrid infrastructure, – it is well suited for it. Technically full simulation software for LHCb experiment is producing Monte Carlo data in the same form as a real detector and processing them by a reconstruction program. The output data is a collection of events written in the format as expected from detector electronics (Raw Data) or in the processed form of Event Summary Data. While it deals with such specific data analysis methods and data flow/streaming procedures, it gives a possibility to formulate and use quite



understandable measure for evaluating the efficiency of computing aspects of BalticGrid infrastructure.

Statistical data analysis is following both experiments and is based mainly on the Monte Carlo technique, of different kind for each experiment. It is applied to estimate the expected errors or values of the real measurement by generating large numbers of simplified experiments and by looking to the distribution of final results. The procedure requires submitting a large number of CPU intensive jobs, each consuming a few hours on an average CPU unit. It is the grid technology that enables to provide the answer in a reasonable period of time. The final application generates files containing input parameters, submitting jobs for each input file, checking if jobs are finished, collecting results and producing the final processing.

3.1 CMS (COMPACT MUON SOLENOID) APPLICATION

A new revolution is pending in high-energy physics. With the launch of the LHC accelerator at CERN in 2007, there will be a flood of new data. No single computing centre can aspire to process it all. It is important to a scientists and researchers of the Baltic States to partake in reaping the fruits of this cornucopia of data, in order to stay on the front-line. High-energy physics is fundamental of our understanding of the physical world.

There are several hints of new physics beyond the Standard Model. The discovery of the neutrino mass is a clear indication that the Standard Model is not a complete theory of particle physics. In the Standard Model, neutrinos are massless; we need new physics to provide the masses and the data from the LHC machine to help determine this physics experimentally.

Also, the LHC is needed to discover the scalar Higgs doublet, the particle whose vacuum expectation value obtained in electroweak symmetry breaking gives masses to weak vector bosons and quarks.

In addition, supersymmetric particles such as gluino or wino, for example, could be discovered in the LHC. Supersymmetry is an elegant symmetry relating fermions to bosons needed for gauge coupling unification and string theory.

3.1.1 Simulation of CMS events

As CMS is one of the four large experiments at the LHC, then it is quite well suited to the Grid as the Grid has been originally designed and deployed for exactly such applications. As the collisions happening in experiment are continuously happening and are independent of each other, then every single collision can be studied separately. The amount of collisions per year is estimated to be around 10^7 , which in turn means that the tasks, which CMS physicists face, are highly parallelizable.

The requirement of a single job is to have access to 1 GB of memory and access to stored datasets. An average analyzable chunk size is ca 500-2000 events, which constitutes to approximately 2 GB data files depending on the contents. Minimum bias (collisions without producing new particles) simulated datasets are smaller and signal datasets as well as final detector response will be bigger. The total amount to be processed and analyzed on the Grid per year should be in the order of 5 PB.

The huge amount of computing power needed for the scientific task comes from the probabilistic nature of the quantum world. Using Monte Carlo techniques it is needed to simulate one physically interesting processes several million times to get information of the process. The whole cycle of Monte Carlo simulation consists of : generation of collision event using known laws of physics and presumable new physics; simulating the propagation of millions of particles, that come into being after the collision, through the complex body of detector taking account of all interactions between particles and material of the detector; finding the detector response and the digital signal that can be read out in the case of working process of the real CMS detector and reconstructing the information back to the particles which were traveling through it; finally, interpreting all reconstructed particles and their specific signatures through complex analysis. Of course in the case of real working LHC experiment the nature makes the first three steps of described tasks, but it is crucial first to simulate the detector behavior to know how the detector works and to be able to interpret the complex signatures of particle to find whether new physics exists or not.

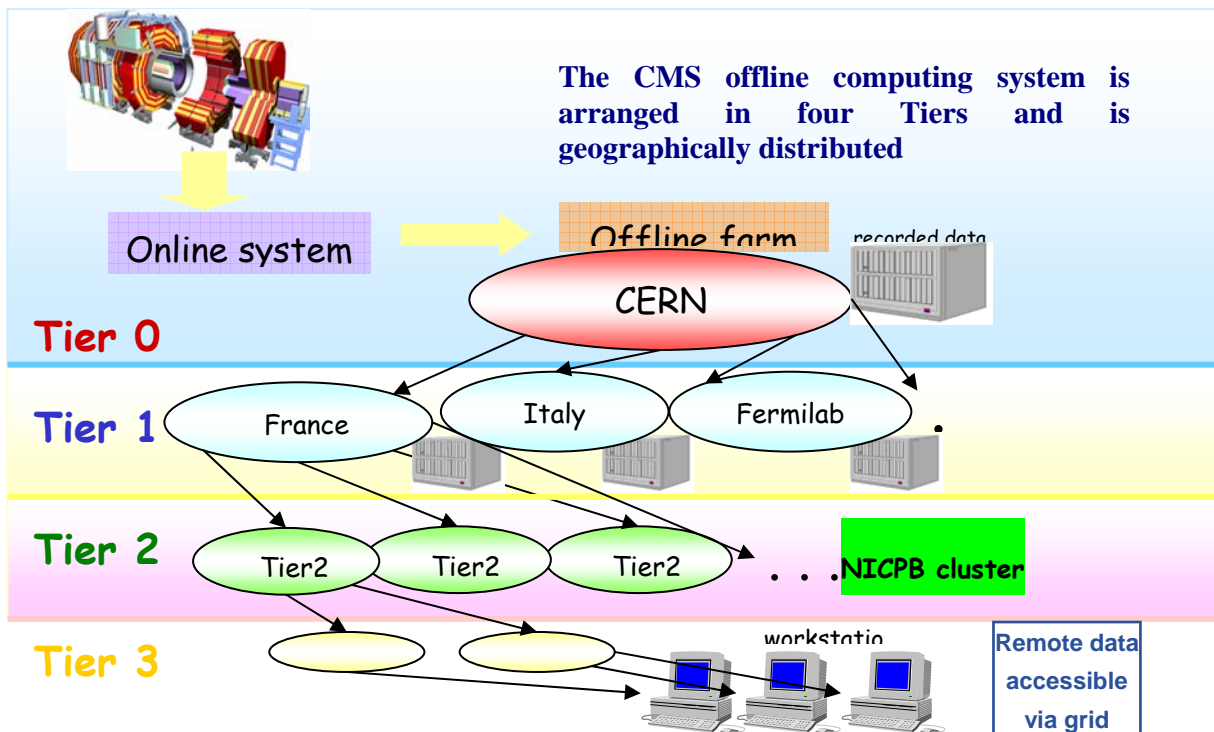


Fig. 1. The architecture of data distribution within CMS experiment simulation



3.1.2 The structure of CMS software

In addition to the requirement that CMS software (CMSSW) should be preinstalled in a computing element, the respective Tier 2 centers which contribute resources to CMS has also to run additional services like PhEDEx (CMS data management and transfer tool), Frontier (calibration data cache) etc. These tools are only needed at participating sites and ordinary simpler jobs like monte carlo production can be run on any Grid resource which has CMS software installed.

The application itself (CMSSW) is distributed as a versioned package. CMS uses remote Grid installations to unify the installed software base across all sites. The first installation initiates a separate RPM (RedHat Package Manager) package repository and uses the apt package management software to install the required software and all its dependencies. The multitude of dependencies will not be described here, but their separate installation is not required as CMS keeps track on what versions work together and always distributed to the sites versions, which are interoperable. Starting from CMSSW, which was released during summer 2006, CMS uses a single software release and unified execution methodology to perform all of the possible tasks (MC production, detector simulation, reconstruction as well as analysis).

To enable a cluster to support CMS VO requires from the cluster only that CMS VO is granted access with enough disk space in the VO specific software area. A single release is ca 2-4GB and usually at least 4-5 releases are maintained at sites although with overlapping software. The actual requirement is in the order of 10 GB per computing element. To support purely the Monte Carlo production jobs the worker nodes must comply with minimum requirement of 1GB of memory per job and outbound access to allow CMS jobs to send monitoring information during job execution as well as stage out produced files to a central location. SRM client software is also needed to be installed at the worker nodes. For software installations rpm build operating system utility is required as some packages are rebuilt on the worker node to best match to the running environment.

3.1.3 Implementation of CMS jobs

For Grid users the implementation of CMS jobs is easy through a special application called CRAB. Users specify their own analysis code and configuration files as well as which datasets they want to run on and how they would like the dataset to be split (number of events per job etc). CRAB then performs the resource location through available datasets as well as clusters supporting all of the requirements, splits the jobs according to user specifications and submits them. In the first year of BalticGrid project there was a steady transformation of the architecture of CRAB software, from the old form to a very recent one, as it is shown in fig. 2 below.

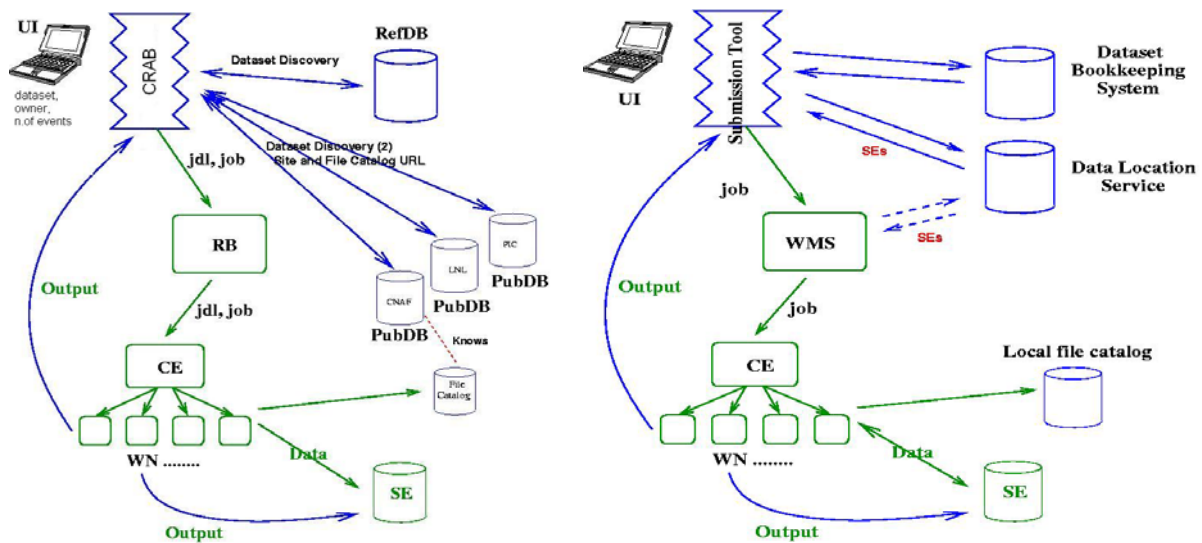


Fig. 2. The architecture of CRAB at the start of BG (left graph) and in recent days (right graph).

CRAB is a user-friendly tool whose aim is to simplify the work of users with no knowledge of grid infrastructure to create, submit and manage job analysis into grid environments:

- written in python and installed on UI (grid user access point)

Users have to develop their analysis code in a interactive environment and decide which data to analyse. They have to provide to CRAB:

- Dataset name, number of events
- Analysis code and parameter card
- Output files and handling policy

CRAB handles data discovery, resources availability, job creation and submission, status monitoring and output retrieval.

The actual results are also managed by the same application which brings the resulting output files to a central location allowing the user to just specify the configuration and then forget about the whole Grid integration part. For bigger Monte Carlo production activities as well as data reconstruction or re-reconstruction, they are performed centrally by a limited number of people who are competent and have special tools for such tasks.

During the first year of BalticGrid project CMS has been running a lot of Monte Carlo production as well as some analysis jobs on BalticGrid. During this period there was also bigger service maintenance when migration from old software to new software was performed and in which time CMS didn't run any significant number of jobs on Grid anywhere. Starting from June/July CMS started the analysis and data production challenge "Service Challenge 4



(SC4)” in which one of the BalticGrid partners (NICPB) took part. The challenge lasted until end of September. During that challenge almost 17 000 analysis jobs were run in T2_Estonia (NICPB) as can be seen in fig. 3:

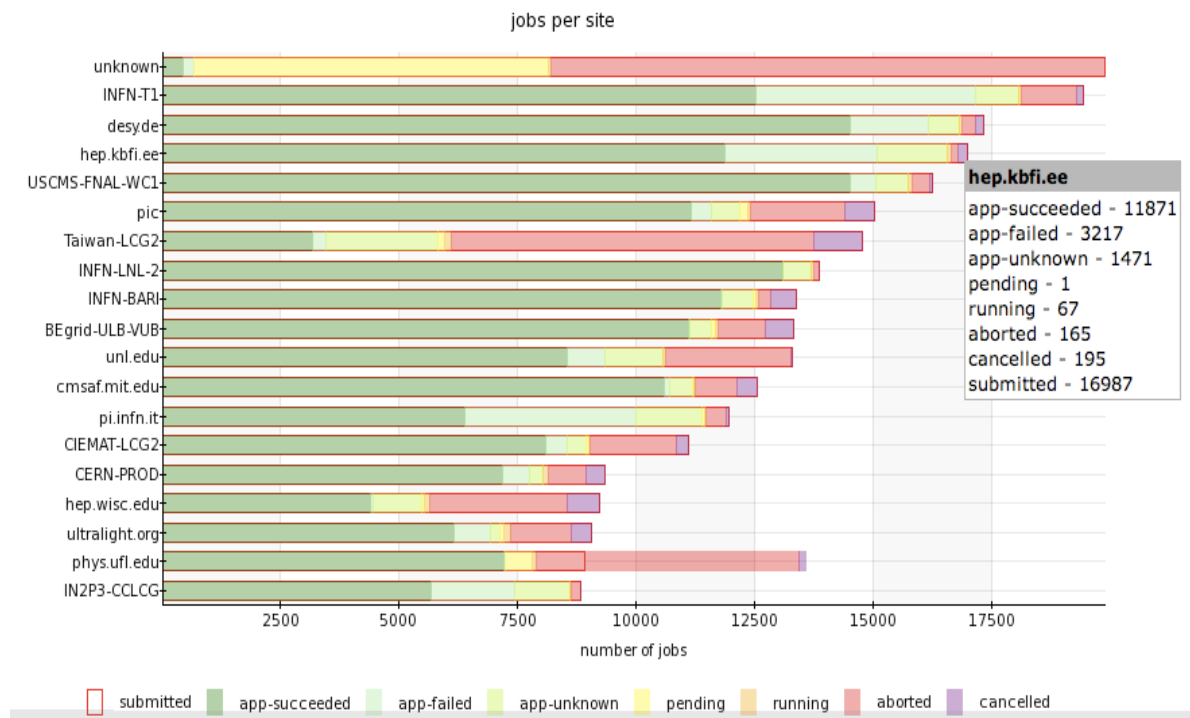


Fig. 3. Analysis' jobs of CMS on T2_Estonia cluster

In addition CMS produced 66Mio events out of which a fraction was produced in Estonia. Also some users of CMS VO were running their jobs on other clusters, which are tested in accordance to EGEE rules (clusters in Lithuania and Latvia). During SC4 also the transfer quality and possibility was tested to verify CMS computing model as well as individual computing centers and their connectivity. In that time T2_Estonia transferred to it's location 160TB of data and exported a bit more than 90 TB of data as can be seen on the following figures:

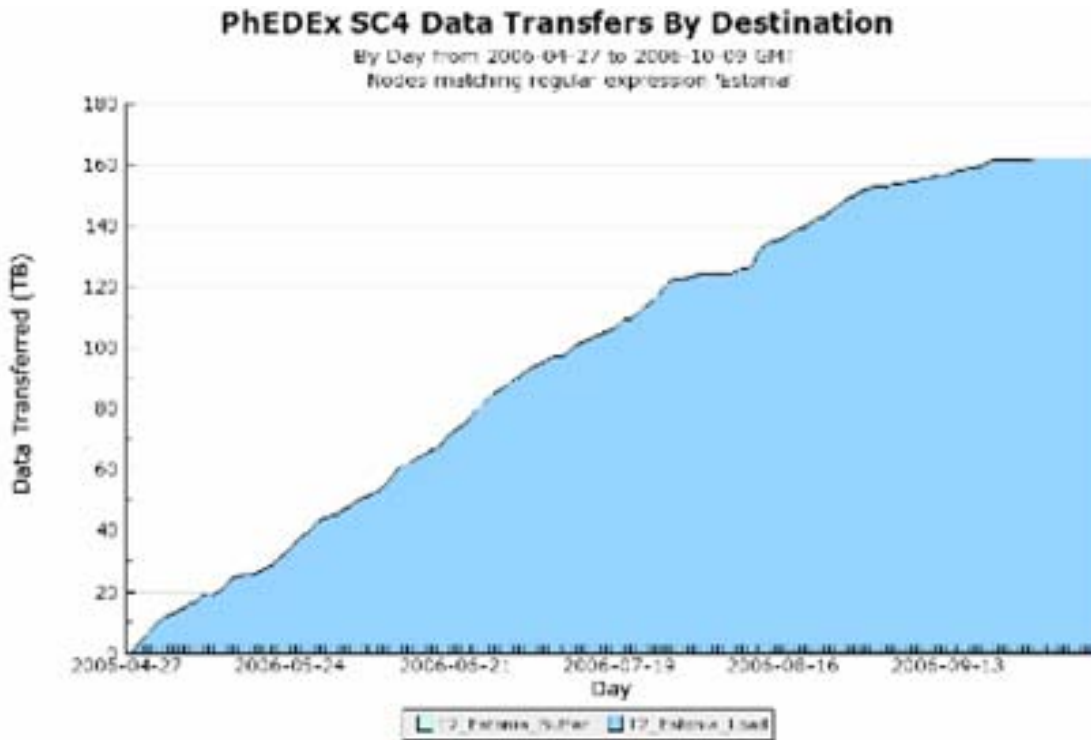


Fig. 4. PhEDEx data transfers by destination

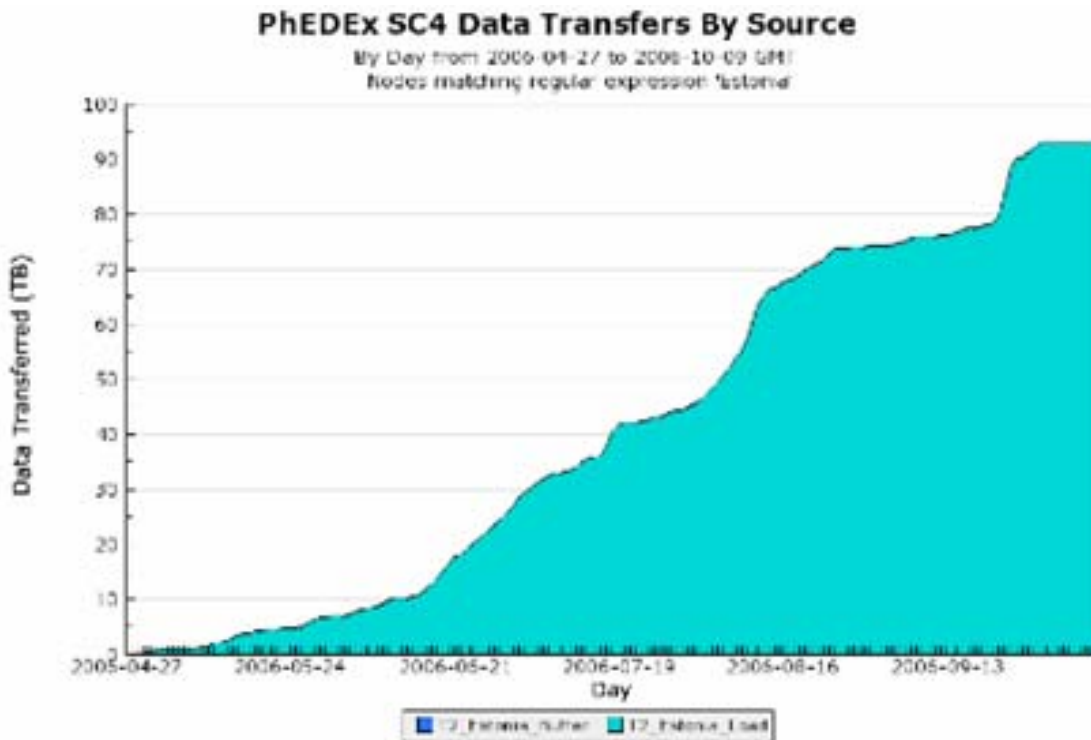


Fig. 5. PhEDEx data transfers by source



The average transfer rate of 20MB/s which was the target for SC4 was kept for majority of the challenge with some slower periods and with peaks to almost 50MB/s

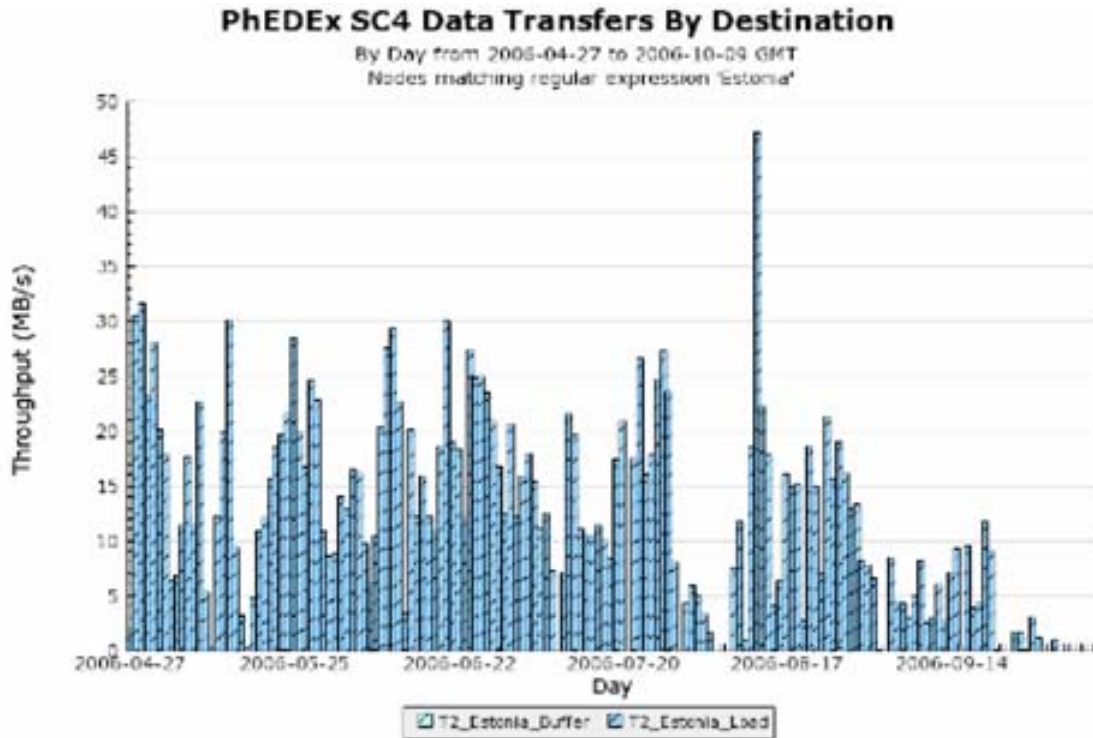


Fig. 6. PhEDEx SC4 data transfers by destination

In October 2006 CMS started the next challenge CSA06 (Computing Software and Analysis 2006), which is to demonstrate 25% capacity running of the whole computing model. This includes data reconstruction and processing at a constant pace at CERN, exporting of data to Tier 1 centers and from there to Tier 2 centers. Also analysis of re-construction results is to be performed. To the date of writing of this document only the data export and transfer part of the test has been in running and has been running as expected. Tier 0 has been performing over the estimates (100% uptime) and Tier 1 and Tier 2 centers are capable of keeping the pace of incoming data. There have been tests (both intentional and unintentional) of site downtimes and elimination of backlog of transfers. The average rate NICPB has been able to keep over the whole 3 weeks of CSA06 has been in the excess of 20MB/s including downtimes and has been able to keep over 40MB/s the past few days:



3.2 LHCb (LARGE HADRON COLLIDER BEAUTY) APPLICATION

The main goal of this application, among other goals, is to test and validate the BalticGrid infrastructure. The means for such test are processes related to CERN experiments. Tests aim to show, how BG operational infrastructure corresponds to the EGEE one. Namely, the efforts in the project were taken to generate processes of high energy physics, to simulate detectors' response and to perform the re-construction task. These efforts are followed by the additional procedures: to store results in form of AOD, to install the LHCb or similar software components and to test BG infrastructure in real data production and data analysis tasks.

The application implemented is to simulate large number of events using the full simulation program, from LHCb experiment of CERN. A single event simulated contains the information about products of proton-proton collision. It is presented in the form which is expected to come from the real detector. Then the data are used to research experimental aspects, in a full scale. The main aspects of such study are:

- to analyze the detector's performance,
- to optimize the detector's design,
- to develop algorithms capable to run on the on-line farm, during real data taking (Trigger algorithms),
- to prepare off-line event reconstruction and analysis.

3.2.1 Requirements and computing procedure

A single LHCb job produces an output file typically corresponding to 500 events of proton-proton collision at the LHC nominal energy. The application requires relatively modern hardware: computers must be equipped with CPUs faster than 1.5 GHz and having at least 1 GB of RAM memory. The main platform for software development has to include i386 compatible processors and Linux operating system. Linux flavour, currently supported is SLC3. Another platform corresponds to MS Windows with MS Visual C++ compiler. Also LHCb VO has to be supported by each site running this application.

The validation of the EGEE and BalticGrid infrastructure is restricted to SLC3 systems only.

Production job consists of a few steps executed on a given working node (WN). Each output of a given step is used as an input for the next step and is processed subsequently. Most files from intermediate steps are deleted. The final output file is usually transferred to the Tier-1 center (by default now is CERN) and stored on a tape.

3.2.2 Prerequisites and software

Jobs are submitted centrally to an execution site, employing LHCb production managers. The execution site has to pass standard LCG tests as well as specific LHCb tests. Dedicated jobs, sent to the site by the LHCb production manager, are responsible for the LHCb software installation. The site is registered in a list of active sites known to the LHCb resource broker.

The main part of application is based on the code developed by the LHCb Collaboration. The general framework is called GAUDI. It incorporates several HEP packages that are used at various stages of data processing.

Application consists of two main modules. The first one, called GAUSS, involves three main steps:

- physics' event generation at the level of particle four-momenta (PYTHIA+EVTGEN),
- transport of particles throughout the detector material (GEANT4),
- generation of an response of electronics, and building the raw data structure of the event (Raw Buffer).

Output file from a single job contains typically about 500 Raw Buffers, one Raw Buffer per event. The second module, called BRUNEL, reads Raw Buffers and performs event reconstruction in various detector systems, e.g. in tracking, in particle identification, in calorimeters or in muon system. The records of output file in ESD format are recorded onto disk and then transferred to the high capacity data center, where they are stored on tapes.

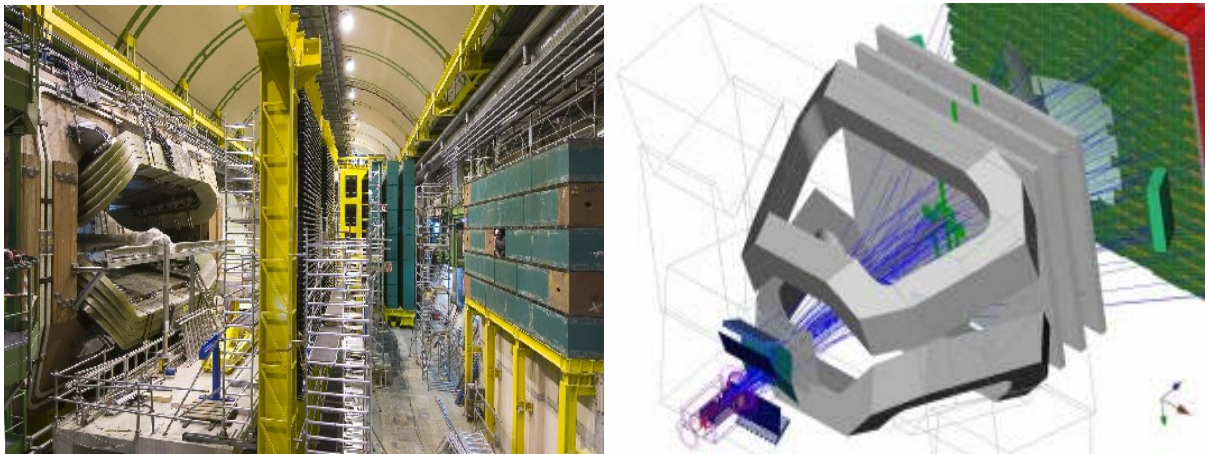


Fig. 8. LHCb detector under the construction (left) and its simulated counterpart (right). On the right picture the blue lines correspond to products of proton-proton collision.

3.2.3 Statistics

MC production corresponding to LHCb jobs has been running on the LCG infrastructure all over the world for several years. In particular, about 2% of the MC production was performed



in Cracow and Warsaw during 2006. Daily average is about 40 jobs running in Tier-2 in Cracow.

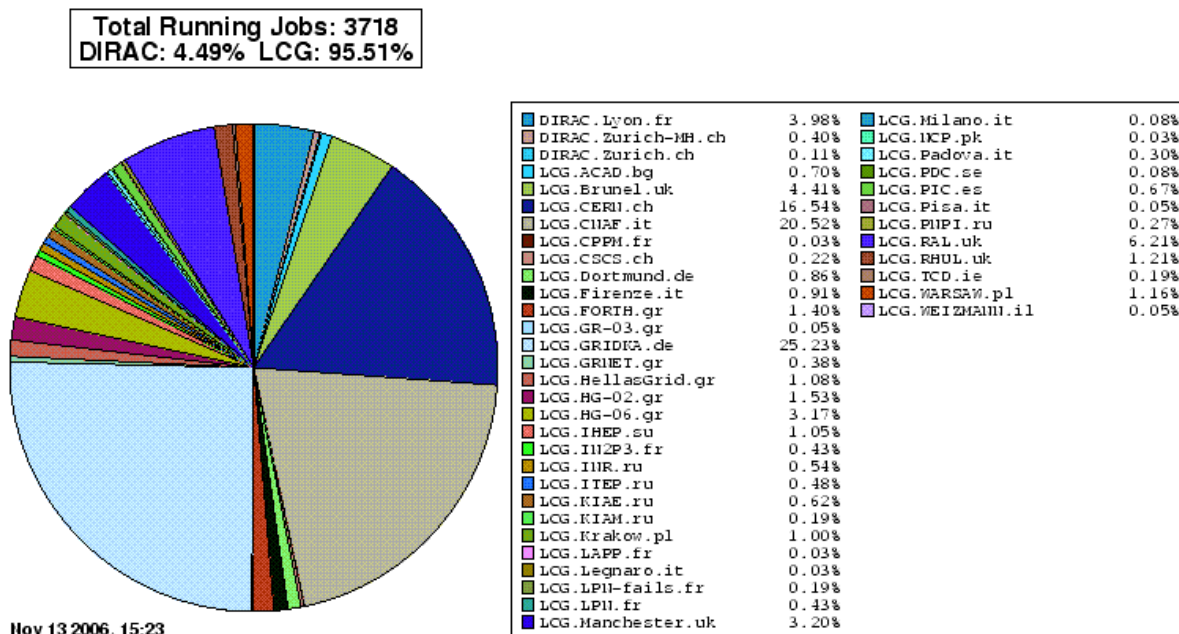


Fig. 9. LHCb monitoring system and snapshot of running jobs for sites participating in the MC production.

Three BalticGrid sites in Lithuania, Latvia and Estonia are to be prepared to use the LHCb MC production for tests of their EGEE functionality. The sites are currently in a preparation stage to support the LHCb VO. After completion they will be registered into the list of the LHCb production sites. Results of the tests will be reported in the next reporting period.



3.2.4 Statistical data analysis using Monte Carlo methods

The principle purpose of this application is to provide a framework for solving statistical problems, employing CPU intensive Toy Monte Carlo methods. The application is being primarily developed in order to study both sensitivity and systematic biases of the CP violation measurements. It can be used to solve other statistical problems as well.

The CP, which transforms matter into anti-matter particle, stands for combination of C – charge conjugation, and P – parity. The violation of CP symmetry implies that the behaviour of matter and anti-matter is different. It is one of the three conditions necessary to explain why the visible Universe is overwhelmingly made of matter.

The CP measurements come as a result of complicated procedure. *B* mesons are produced in hadronic environment of proton-proton collisions. The *B* meson production is over 100 times smaller than the normal one. Moreover, *B* meson decays that are interesting for CP measurements, are relatively rare ranging from 10^{-4} down to 10^{-9} of all *B* decays. Extraction of tiny signal out of the huge background requires sophisticated algorithms, operating already at the level of on-line data taking (reduction from 40 million down to 200 events per second). The data are then reconstructed off-line and CP-violation phenomena are studied, for more than 50 different *B* meson decay modes. In the final step, essential physics parameters are determined by applying the fitting procedure to the data.

3.2.5 Importance and Requirements

The application is used by physicists from the IFJ PAN LHCb group and concerns the analyses of CP violation in B meson decays. The framework of the application is generic and may be exploited in other scientific fields covered by the NA3 task of the BalticGrid.

Full functionality of the application is provided by the ROOT package. The ROOT system provides a set of object oriented frameworks with all the functionality needed to handle and analyse large amounts of data in a very efficient way (see fig. 8). Having the data defined as a set of objects, specialised storage methods are used to get direct access to the separate attributes of the selected objects, without having to touch the bulk of the data. Included are histogramming methods in 1, 2 and 3 dimensions, curve fitting, function evaluation, minimisation, graphics and visualization classes to allow the easy setup of an analysis system that can query and process the data interactively or in batch mode.

Because of the built-in CINT C++ interpreter the command language, the scripting, or macro, language and the programming language are all in C++. The interpreter allows the fast prototyping of the macros since it removes the time consuming compile/link cycle. It also provides a good environment to learn C++. If more performance is needed the interactively developed macros can be compiled using a C++ compiler.

The system has been designed in such a way that it can query its databases in parallel on MPP machines or on clusters of workstations or high-end PC's. ROOT is an open system that can be dynamically extended by linking external libraries. This makes ROOT a premier platform suitable to build data acquisition, simulation and data analysis systems. The ROOT package distribution is available on most of popular platforms: Linux, UNIX and Windows. While running on the BalticGrid clusters, the Linux version is used.

3.2.6 Computing procedure and software

A Toy Monte Carlo approach is commonly used in HEP experiments. HEP detectors are very complex apparatus, moreover, investigated phenomena are very subtle. The required precise understanding of both the whole chain of data acquisition as well as the analysis is provided by the full simulation program. The full simulation is a very CPU intensive task, thus only limited statistics MC samples can be produced, insufficient to study measurements precisions. Toy Monte Carlo technique is commonly used in such cases. The main idea is to prepare a simplified model of the measurement employing parton distribution functions (PDF), acceptance functions etc., that are derived from the full simulation program, to submit a large number of jobs which execute the procedure with different initial parameters on a GRID and to analyse distributions of outcomes. The final result depends on many other parameters, some of them determine the PDF or a variable essential for final results extraction.

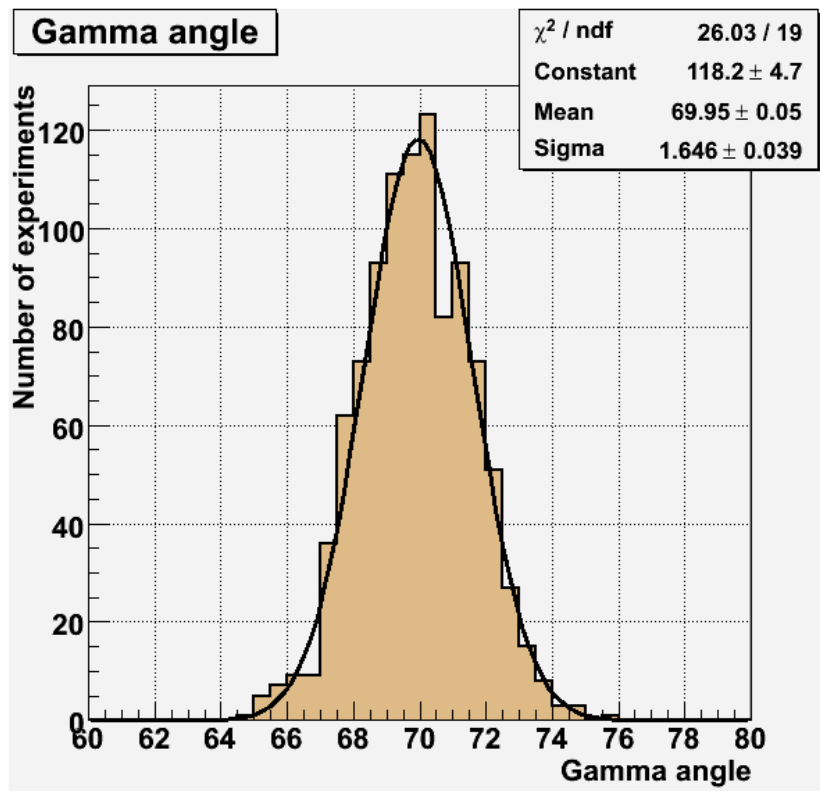


Fig. 10. The typical distribution of results from a bunch of 1000 jobs. The width of the distribution is an estimation of the measurement uncertainty.



Dedicated package which provides tools for building Toy MC program is called RooFit. It was developed for the BaBar experiment and is a part of the ROOT framework. RooFit contains many advanced utilities like high quality pseudo-random number generators, a large set of probability density functions, including conditional probability densities, etc.

The structure of application consists of three main parts:

- **framework part** – based on ROOT, it is responsible for providing the site with ROOT environment. If no installation is found, the ROOT sources are imported and compiled. This is done once, when first job arrives to a given cluster. Thus, the application can be executed on any Linux flavor and some Unix platforms
- **user part** – in the form of ROOT script (C++), although the user script is specific to HEP topics, it can be easily adopted to solve other problems of BalticGrid users
- **grid part** – responsible for submitting jobs and retrieval of results, currently implemented in the form of a set of shell scripts that prepare and submit series of jobs followed by collecting the outputs and final analysis. It is foreseen to integrate the application into Migrating Desktop. The execution of few thousands of jobs is needed in the full production mode to solve one problem. A single job may require several hours of CPU time, depending on complexity level of the problem. The production procedure may require several repetitions.

User script can be executed in two modes:

- interactive mode (C++ interpreter) during the development phase on private workstation
- batch mode for production on the GRID employing compiled version of user code, that results in a faster execution than for an interpreted code.

Statistics. A few thousands of jobs have been executed on BaltigGrid sites in total. Initially, several pilot mini productions has been submitted in bunches of a hundred jobs to verify the ROOT installation and correctness of main steering script. Then the simple RooFit scripts were attached and executed, and the results from different sites have been compared. Finally, a simple model of the CP violation measurement has been implemented and extensive productions have been launched. Preliminary physics results have been obtained.



4 CONCLUSIONS AND FUTURE WORK

4.1 ACHIEVEMENTS:

1. The High Energy Physics applications in BalticGrid, namely: “CMS application”, “LHCb application”, following by statistical Monte Carlo data analysis, are well developed and implemented in the BG infrastructure, they correspond also to the applications, developed or under development in EGEE. The results of the development and implementation of applications correspond to the project’s milestone “MNA3.7 HEP application operational”, this milestone is achieved.
2. BalticGrid implemented these applications by supporting the existing VOs in EGEE, such as the CMS and LHCb VOs.
3. The application “Statistical Data Analysis using Monte Carlo Methods” has a potential to be developed further, for other areas of research. In that case it will attract additional users: scientists and specialists from the computer modelling of live sciences area. It has a potential to be transformed into a separate SIG.
4. The implementation and operational of pilot HEP applications gives following benefits to the users from Baltic States:
 - they introduce a knowledge to Baltic users, which allows them to interact efficiently with computing and research infrastructure of other European countries,
 - they target a wide range of computing needs,
 - they give large data transfer and throughput computing needs of the LHC experiments (of special value are CMS experiments for users from the Baltic States, as well as LHCb experiments for Baltic region users)
5. The technical and operating environment of HEP applications have the following development features:
 - most of applications are using open source code, which is common practice in many grids
 - programming platforms in use are C++ and Fortran, together with popular subroutine libraries
 - many applications are indifferent to the computer architecture (32-bit and/or 64-bit), like statistical software, nevertheless some applications, like CMS and LHCb experiments, still need x86 32-bit computer architecture.

4.2 ACTIONS PLANNED FOR THE FUTURE

4.2.1 Constant support of HEP applications implemented



Computational grids are becoming increasingly common, promising ultimately to be ubiquitous and thereby changing the way global resources are accessed and used. At the same time grids are seen as crucial technology for scientific research. The dynamic, heterogeneous and distributed nature of grids dramatically increases the complexity of grid application development.

Scientists from the Baltic States will take major advantages by running their applications in a distributed computing environment, like the Baltic Grid infrastructure, if the applications will have always a status up-to-date, their software and architecture will be constantly renewed and updated.

The application support expert group will maintain a close communication between application developers and grid experts to speed-up application adaptation, providing analysis for the deployment/run of applications in a grid environment.

4.2.2 Strategy for establishing new VOs

Currently the general VO “BalticGRID” is being established and installed. Other VOs, serving some applications from Material Science (GAMESS and DALTON) as well as some regional users (LitGrid VO) are also implemented. As soon as the group of active users of statistical Monte Carlo data analysis will be identified, from the same domain or using the same software package, the new VO will be arranged for them.