Romanian Reports in Physics, Vol. 65, No. 1, P. 122–132, 2013

BUILDING AND TESTING AN INFRASTRUCTURE FOR STUDYING PROTON-PROTON COLLISIONS AT TeV SCALE

M. CUCIUC[1, 2], M. CIUBANCAN[1, 2], V. TUDORACHE[1, 2], A. TUDORACHE[1, 2], R. PAUN[1, 2], G. STOICEA[1], C. ALEXA[1]

[1] "Horia Hulubei" National Institute for Physics and Nuclear Engineering, Str. Reactorului no.30, P.O. BOX MG-6, RO-077125, Măgurele, Romania,
[2]University of Bucharest, Faculty of Physics, Str. Atomiştilor 405, P.O. BOX MG-11, RO–077125, Măgurele, Romania
E-mail: Constantin.Mihai.Cuciuc@cern.ch, E-mail:Mihai.Ciubancan@cern.ch,
E-mail: Valentina.Tudorache@cern.ch, E-mail: Alexandra.Tudorache@cern.ch,
E-mail: Remus.Andrei.Paun@cern.ch, E-mail:Gabriel.Stoicea@cern.ch,
E-mail: Calin.Alexa@cern.ch

Received August 2, 2012

Abstract. The motivation and means of defining an infrastructure for Monte Carlo production is presented. The system was built using available software packages for generating events, simulating the detector response and reconstruction of the physics objects. A master-slave architecture computer cluster was design and implemented to provide storage and data processing resources. Tests were performed to check the performances of the developed system using both well-understood physics channels, such as top quark pair production and decay, as well as beyond the Standard Model physics signals. Data processing tests based on Condor batch system and PROOF architecture for parallel data analysis will also be presented.

Key words: framework, generation, simulation, analysis.

1. INTRODUCTION

Studying particle accelerator physics involves vast phase-spaces and large numbers of processes that are known only to some extent. The adopted analysis method is simulating large numbers of events according to theoretical models by Monte Carlo generators and checking measurable quantities obtained from these events against similar measurable quantities acquired by detectors from real collisions. Given the cost of a high-energy physics experiment it is not a surprise that many theoretical models are tested in simulated environments long before they are confirmed by real data. Furthermore, in order to obtain quantities similar to those obtained from detectors, more simulations are performed in order to estimate the results of the particles' interaction with matter, which is also an intrinsically statistical process.

2. HARDWARE SETUP

The software infrastructure is based on a non-homogeneous computer cluster that was designed to provide storage and data processing resources for Monte Carlo simulations and data analysis. The master-slave architecture cluster contains a number of 40 cores where the master-node is a server with 8 cores and 7 TB of storage. There are also 5 slave nodes, 3 with 8 core and 2 with 4 cores. The network connection between the machines is done *via* a 1 Gbps private network providing a very good level of network security, the master-node being configured as a gateway for the rest of the servers.

3. SOFTWARE SETUP

In order to perform the simulation and analysis, a chain of software packages has been put together. This allows one to generate events using a Monte Carlo generator, simulate the detector response for these events and perform analysis on them. Available packages have been used wherever these were available for the required tasks, in order to lower the effort put into the development of the platform and focus on the testing and usage of the system. The next sections describe each of the modules in detail.

3.1. EVENT GENERATION

The Monte Carlo generators that have been used in testing this framework are PYTHIA 6.4 [1], PYTHIA 8.1 [1,2] and POWHEG [3]. All these have different advantages, a summary description of the three being given in Table 1.

Table 1

Comparison of the specific features of Monte Carlo Generators

Generator	Diagram Precision	Showering	Programming language
PYTHIA 6.4	LO	Yes	FORTRAN
PYTHIA 8.1	LO	Yes	C++
POWHEG	NLO	No	FORTRAN

The fact that POWHEG is a Next-to-Leading Order (NLO) generator distinguishes it from the other two, making its output preferable. However, the Leading Order (LO) generators are equally suited for studies involving reconstructed events, as most differences are ironed out by the uncertainties introduced by the detector simulation. One advantage that both PYTHIA packages have over POWHEG is that they include the showering mechanism. POWHEG provides the particles resulted from the interaction (albeit having calculated their properties with NLO precision) and does not decay or hadronize them. The data sets that were produced with POWHEG were showered by PYTHIA 8.1.

A key point regarding simulating hadron-hadron collisions is the description of the internal structure of the interacting hadrons. This description is provided by the LHAPDF [4] library, for which event generators have an interface to. All studies that employ the PYTHIA 8.1 or POWHEG event generators in this paper use the CT10 [5] parton distribution function provided by LHAPDF. The PYTHIA 6.4 samples have used the CTEQ 5L [6] PDF.

All Monte Carlo generators output their showered particle data using the HepMC [7] event record, while POWHEG outputs an intermediate file of unshowered events using the Les Houches Accord [8] event record.

3.2. DETECTOR SIMULATION

While the event generators provide relevant information themselves, allowing for detailed studies of the involved processes, in order to be able to compare the simulations to the real data one needs a detector model and its response to the involved particles. Ideally, the simulation step would transport the generated particles through the detector and calculate each particle's interaction with both active and passive materials in the detector. While this is an accurate process, it is both time consuming and highly detector-specific, allowing for little configurability. The Delphes [9] package, which is used for this purpose, uses an approach where the particles' paths are intersected with the components of the detector and the energy deposits are estimated from these intersections. In order to simulate the noise and limited resolution of real hardware there is an additional step that smears the obtained results. For object reconstruction, Delphes implements a tracking system in a magnetic field, calorimeters and a muon system. The package is able to describe complex detectors and outputs physics analysis objects, such as photons, leptons, missing transverse energy and jets. For the purpose of this analysis it has been modified to also include the calorimeter constituents for each jet.

3.3. DATA FLOW

Once the Monte Carlo generators provide the output data this has to be further processed by the reconstruction package and the user analysis. The data flow for this scheme is presented in Fig. 1.

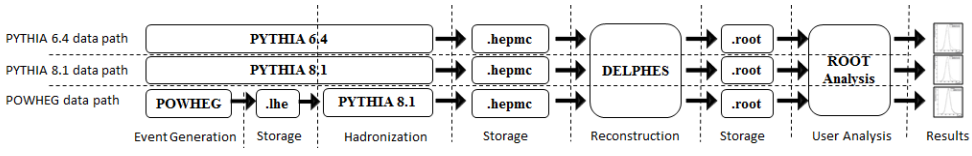

Fig. 1 – Architecture and data flow for each of the three Monte Carlo Generators.

It is worth mentioning that the output provided by the Delphes package contains copies of all particles produced by the Monte Carlo generators, allowing for both parton-level and reconstruction-level analyses to be performed simultaneously, on a single file.

3.4. COMPUTER CLUSTER SOFTWARE SETUP

The computer cluster was designed to run Monte Carlo simulations using the Condor [10] batch system and data analysis using the Parallel ROOT Facility, PROOF [11]. Both applications are using the same head node.

From the Condor batch system point of view, the master-node is running COLLECTOR, MASTER, NEGOTIATOR and SCHEDULER services, basically controlling the resources, managing the job's distribution and recovering the output of the jobs. While the other nodes are configured as worker nodes, were the simulations are running. The user and storage space is distributed among the worker-nodes using NFS protocol.

In order to perform parallel data analysis using the PROOF framework we configured one server as master-node while keeping others as slave-nodes.

The master-node configuration file includes the necessary declarations of the storage pool, which is used for data storage transferred with the xrootd [12] protocol.

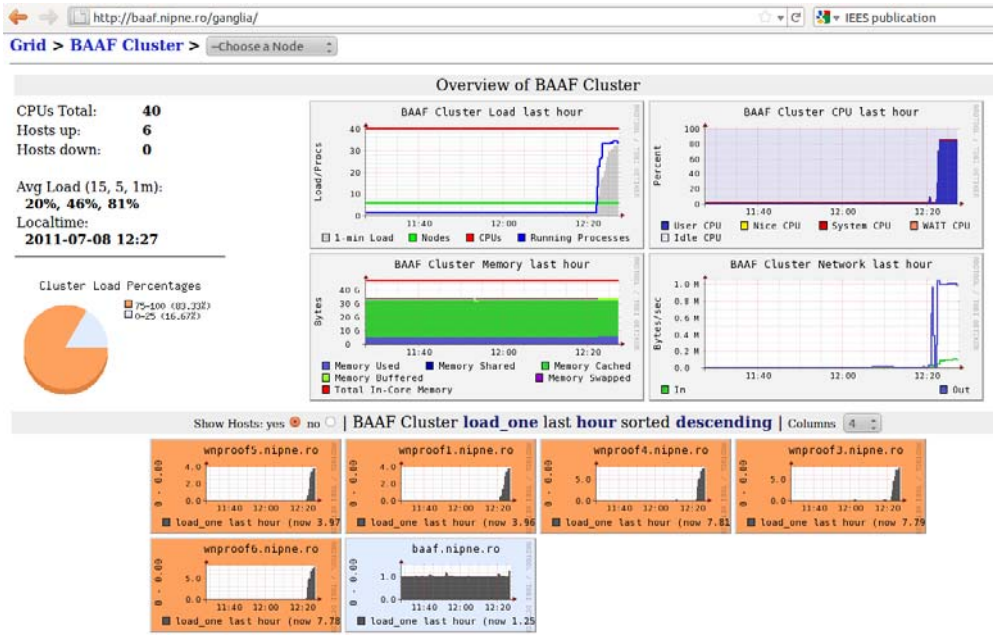

Fig. 2 – Ganglia monitoring web page: the load of the computer cluster during tests.

For the parallel data analysis infrastructure we have chosen xrootd because is a fully generic suite for fast, low latency and scalable data access, which can natively serve any kind of data. During our tests we have noticed more stable connections having better transfer speeds using xrootd when compared to the NFS protocol. When compared to secure shell (scp), there was a 5 fold increase in transfer speed when using xrootd. Using xrootd allowed the system to use the full network bandwidth.

The computer cluster is monitored using the Ganglia Monitoring System [13] through a password protected web interface. This gives the user the ability to monitor key parameters, such as the process distribution throughout the cluster and memory, CPU, network usage statistics.

Figure 2 shows some of the available parameters shortly after a 10^9 events job was submitted.

4. TESTS AND ANALYSES

Having the generated events, as well as the reconstructed objects, the analysis consists of an application that has been written based on the ROOT [14] framework which provides easily accessible open-source libraries for high-energy physics analyses. A number of studies were performed for testing the whole platform, from generators to analysis.

Several tests have been performed to check the functionality of the batch system and of the parallel data processing infrastructure.

4.1. COMPUTER CLUSTER TESTS

In order to run simulation jobs using the Condor batch system one needs to prepare the job by providing a submit script file that instructs the batch system how to proceed. This script file contains, among others, information about the input file, output file and executable which is to be run. Passing this file as a parameter to the condor_submit command completes the job submission process.

4.2. PROOF TESTS RESULTS

Preparing the data analysis for PROOF framework takes some additional steps. First, one has to declare the datasets using commands similar to these:

```
root[.] TDSet *set = new TDSet("TTree", "pythiaEvents")
root[.] set->Add("/pool/data/test/proof_test._00001.root")
root[.] set->Add("/pool/data/test/proof_test._00002.root")
```

The analysis skeleton script is generated automatically by the framework:

```
root[.] pythiaEvents->MakeSelector("Analysis")
```

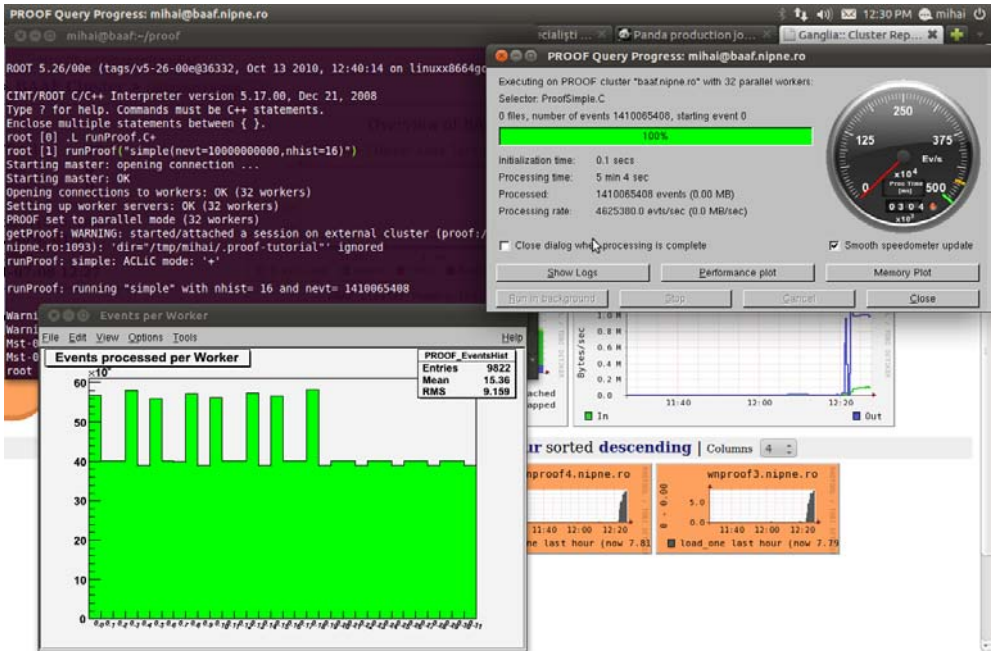

Fig. 3 – Ganglia monitoring web page: the load of the computer cluster during tests.

The framework initializes the two analysis skeleton files, Analysis.h and Analysis.C which the user can then customize to operate on the input data in a meaningful way, while also creating its output data. The actual job submission is executed, in our case:

root[.] set->Process("Analysis.C")

During testing of the PROOF setup we have noticed a 6 fold increase in job completion time when using the full cluster, compared to the time it took a single dual core machine to complete the test. Figure 11 shows the results of one such test.

4.3. TOP-ANTITOP EVENTS

Top quarks events produced in p-p collisions $\left(pp \to t\bar{t} \to bW^+ \;\; bW^-\right)$ can be characterized by the way the W-bosons decay. The top-antitop events are divided in three channels [9]: the fully leptonic (dileptonic) channel, the fully hadronic channel and the semi-leptonic channel.

4.3.1. Semi-leptonic channel

The top quark pairs produced decay in 4/9 of events in the semi-leptonic channel [9], one W is decaying leptonically and the other W is decaying

hadronically. The final state is composed of four jets, a lepton and missing transverse energy. This channel is both easy to reconstruct and able to give the top quark mass for each event.

The selection criteria were chosen according to the event topology, and these are: at least 4 jets with large transverse momentum, exactly one lepton with sufficient transverse momentum and missing transverse energy. Each event that satisfies the selection criteria is processed: the three jets which, combined yield the largest p_T are considered to come from the top quark that decays hadronically. An additional selection is applied requiring that the combination of two of the three jets provides a mass consistent within ± 20 GeV/c^2 of the W mass. Several samples have been generated, both for background as well as for the signal. These are summarized in Table 2, along with the selection efficiencies for subsequent selection criteria applied on the data samples.

Table 2

Selection efficiency for signal and background

Sample	Cross-section [pb]	Generator	% events for $n_{\text{jets}} \geq 4$ with $p_T > 40$ GeV	% events for one lepton with $p_T > 20$GeV	% events with $E_T^{miss} > 20$GeV
$t\bar{t}$	120.6	PYTHIA8.1	45.9	6.8	6.1
W+jets	36670	PYTHIA8.1	0.2	$74 \cdot 10^{-6}$	$68 \cdot 10^{-6}$
Z+jets	1207	PYTHIA8.1	0.3	0.091	0.022
WW	28.77	PYTHIA6.4	2.2	0.2	0.2
WZ	10.67	PYTHIA6.4	2.7	0.2	0.2
ZZ	4.434	PYTHIA6.4	3.2	0.2	0.1

Even though one can notice a good level of background rejection, the large cross-section of the W + jets channel is sufficient to affect the analysis. Figure 2 shows the reconstructed masses for both top quarks and W bosons for signal and background processes.

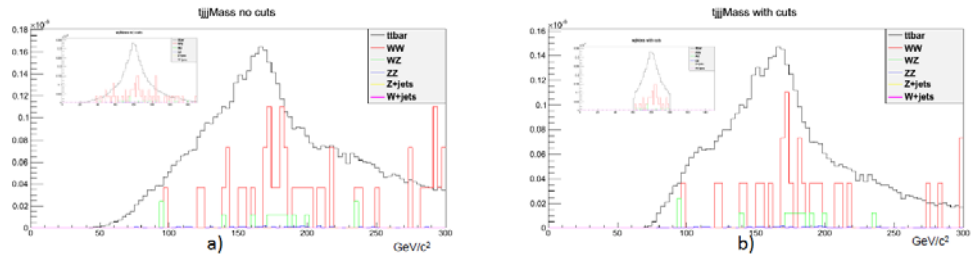

Fig. 4 – Comparison between signal and background: a) reconstructed top quark mass and associated W boson mass (small plot); b) similarly, after the application of the constraint that the reconstructed W boson mass be within ±20 GeV/c2 of the W mass.

Summing both signal and background in order to obtain an image compatible with the physical results one would expect from an experiment, we obtain the plots shown in Fig. 5.

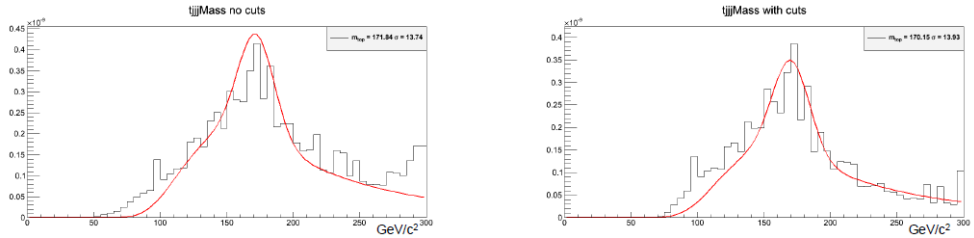

Fig. 5 – Mass fit after the background has been added to the signal. Right plot has the additional constraint that the reconstructed W boson mass be within ±20 GeV/c2 of the W mass.

It can be seen that although the plot shape has been heavily affected by the background, the top quark mass can still be obtained.

4.3.2. Dileptonic channel

The fully leptonic (dileptonic) channel is representing 1/9 of the $t\bar{t}$ events; both W bosons are decaying into a lepton-neutrino pair, resulting in an event with two charged leptons, two neutrinos and two b quarks. This channel has two undetectable neutrinos making the reconstruction of the top mass very difficult.

Using the described infrastructure, detailed analyses have been performed to study the influence of initial state and final state radiations on the reconstruction of the top quark mass in the dileptonic channel [10].

To illustrate the functionality of the cluster we will show here only the transverse momentum of the leptons obtained with PYTHIA and POWHEG, when ISR and FSR were switched on and off (Fig. 6).

4.4. CHARGED HEAVY LEPTON PAIR PRODUCTION

Sequential charged heavy lepton (CHL) pair production in proton-proton collisions at $\sqrt{s} = 14\text{TeV}$ has been studied using PYTHIA 6 Monte Carlo generator [1]. The existence of new heavy leptons is predicted by various models with an extended gauge sector. Both the Drell-Yan and the gluon - gluon production mechanisms were considered.

Assuming Standard Model couplings, $L^{\pm} \to (e,\mu)^{\pm} Z^0 \to (e,\mu)^{\pm} - dijet$ decay channel as a function of heavy lepton mass M_L and Z' boson mass $M_{Z'}$ was investigated.

A comparative study between signal and background was performed for several kinematic variables. We will present here only two plots showing jet multiplicity, (Fig. 7), and leading jet p_T distribution (Fig. 8).

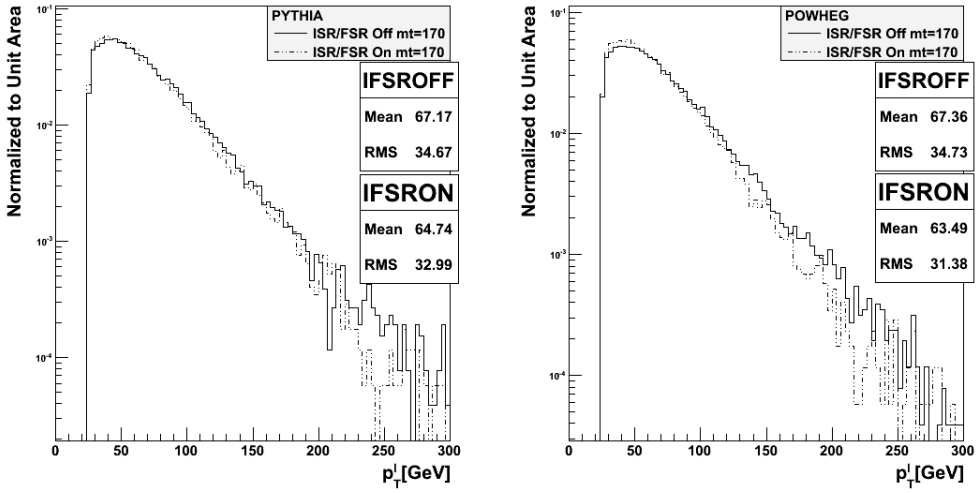

Fig. 6 – Electrons transversal momentum distribution obtained with two MC generators, POWHEG and PYTHIA, for ISR/FSR switched on and off.

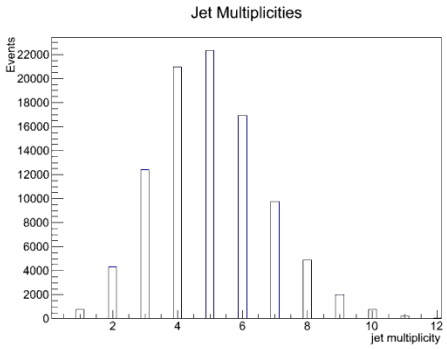

Fig. 7 – Jet multiplicity distribution.

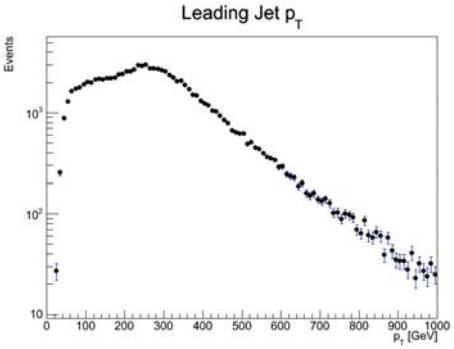

Fig. 8 – Leading jet P_T distribution.

A detailed analysis of this CHL pair production channel will be presented in a dedicated article which is under preparation.

4.5. HEAVY QUARKS PAIR PRODUCTION

A study of pair-production of a sequential 4-th generation quark followed by their decays to a W boson and a b type quark $t'\bar{t}' \to W^+bW^-\bar{b}$ in proton-proton

collisions at $\sqrt{s} = 14\text{TeV}$ has been performed. We considered that the mixing between the *t'* quark and the b type quark is $B.R.\left(t' \to W^{+} b\right) = 1$. We present here two plots showing the P_T distribution for *t'* quark (Fig. 9), and b quark (Fig. 10).

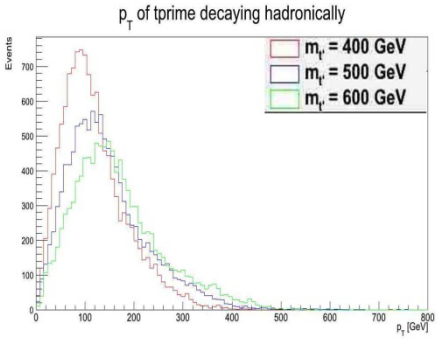

Fig. 9 – The *t'* type quark P_T distribution.

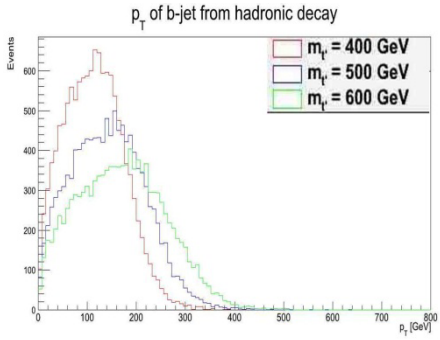

Fig. 10 – The b type quark P_T distribution.

A detailed analysis of this 4th generation quark pair production channel will be presented in a dedicated article which is under preparation.

5. SUMMARY AND CONCLUSIONS

- An infrastructure for proton-proton collisions study at ultra-relativistic energies was built using available software packages for generating events, simulating the detector response and reconstruction of the physics objects.
- In order to provide high-performance capabilities for storage and data processing a master-slave architecture computer cluster was designed and implemented.
- Tests were performed to check the performances of the developed system using both well-understood physics channels, such as top quark pair production and decay, as well as beyond the Standard Model physics signals.
- Results of data processing tests made for Condor batch system and for parallel data analysis PROOF architecture were presented.

The positive results that have been obtained for both the software and the hardware components have persuaded the authors to scale up the computer cluster system by increasing the storage capacity as well as the number of processors.

Acknowledgements. This work was supported by POSDRU/88/1.5/S/56668 and POSDRU/107/1.5/S/82514 grants. We also acknowledge the support of the Romanian National Authority for Scientific Research (ANCS) through ATLAS - Capacities Module III RO-CERN and PN 09370101 contracts.

REFERENCES

1. T. Sjostrand, S. Mrenna, and P. Z. Skands, JHEP, **0605**, 026 (2006).
2. T. Sjöstrand, S. Mrenna and P. Skands, Comput. Phys. Comm., **178**, 852 (2008).
3. P. Nason, JHEP, **0411**, 040 (2004), hep-ph/0409146; S. Frixione, P. Nason and C. Oleari, JHEP, **0711**, 070 (2007), arXiv:0709.2092; S. Alioli, P. Nason, C. Oleari and E. Re, JHEP, **1006**, 043 (2010), arXiv:1002.258.
4. *** LHAPDF, hep-ph/0508110; http://hepforge.cedar.ac.uk/lhapdf/
5. *** CT10 http://arxiv.org/pdf/1007.2241.pdf
6. *** CTEQ 5L, http://www.phys.psu.edu/~cteq/CTEQ5Table/cteq5l.tbl
7. M. Dobbs and J.B. Hansen, Comput. Phys. Commun., **134**, 41 (2001); http://lcgapp.cern.ch/project/simu/HepMC/
8. E Boos et al., hep-ph/0109068 ; J. Alwall et al., Comput. Phys. Commun., **176**, 300 (2007).
9. Delphes, http://arxiv.org/abs/0903.2225
10. Todd Tannenbaum, Derek Wright, Karen Miller, and Miron Livny, *Condor - A Distributed Job Scheduler*, in Thomas Sterling, editor, *Beowulf Cluster Computing with Linux*, The MIT Press, 2002; http://research.cs.wisc.edu/condor/publications.html
11. Maarten Ballintijn, Marek Biskup, René Brun, Philippe Canal, Derek Feichtinger, Gerardo Ganis, Günter Kickinger, Andreas Peters and Fons Rademakers, Nuclear Instruments and Methods in Physics Research A **559**, 13–16 (2006).
12. *** XRootD, http://xrootd.slac.stanford.edu/docs.html
13. *** The Ganglia Distributed Monitoring System: Design, Implementation, and Experience. Matthew L. Massie, Brent N. Chun, and David E. Culler, Parallel Computing, **30**, *7*, (2004); http://sourceforge.net/apps/trac/ganglia
14. *** ROOT: R. Brun, F. Rademakers, Nucl. Inst. & Meth. in Phys. Res. A **389**, 81–86 (1997).
15. The Review of Particle Physics, K. Nakamura et al. (Particle Data Group), J. Phys. G **37**, 075021 (2010).
16. V. Tudorache, M. Cuciuc, *Initial and final state radiation studies for top quark mass reconstruction in dileptonic channel for √s=7 TeV P-P collisions*, Romanian Reports in Physics, **64**, 945–956 (2012).