GRID TECHNOLOGIES IN SPbSU LONG-RANGE CORRELATIONS ANALYSIS AND MC SIMULATIONS FOR ALICE

I.G. Altsybeev, G.A. Feofilov, M.V. Kompaniets, V.N. Kovalenko, V.V. Vechernin, I.S. Vorobyev, A.K. Zarochentsev

Laboratory of Ultra-High Energy Physics, St. Petersburg State University
198504 Ulianovskaya st., 1, Petrodvorets, St. Petersburg, Russia
feofilov@hiex.phys.spbu.ru

Studies of long-range correlations within the physics program of ALICE [1] require a high accuracy statistical analysis of experimental data. In this report, examples of Grid technologies used for the analysis of several types of correlations in proton-proton and Pb-Pb collisions in ALICE are presented. The main stages of software development and debugging on the basis of the AliAnalysis Manager [2] platform are described, allowing one to make calculations with both local and distributed computing systems (PROOF [3, 4], Grid). Examples of Monte-Carlo calculations are also given. A distributed storage and data processing system (ALICE Environment Grid [5, 6]) providing access to experimental data and results of modeling of proton-proton and Pb-Pb collisions is described. The analysis of large data (tens of Tb) allowed one to obtain results with statistics of more than 20 million events as well as to investigate a topological structure of long-range correlations.

Introduction

ALICE experiment [1] at the Large Hadron Collider gives us a unique opportunity to study heavy ion and proton collisions at high energies. In such collisions nuclear matter is believed to undergo a phase transition to quark-gluon plasma, a state of matter which is thought to have existed just moments after the Big Bang. Since the start of operation in 2008 ALICE collects data with rate of about 4 PB a year, which means that one has to develop powerful and adaptable computing system to handle with it. Grid technologies and distributed computing are an integral part of any modern high energy physics experiment and ALICE is not an exception.

Physical program

Our physics motivation is based on the String-Fusion Model predictions [7] where the effects of string interaction are taken into account in the form of fusion or percolation. As quark-gluon string is an extended object in the rapidity space, i.e. contributes by fragmentation to wide rapidity range, one can expect appearance of the correlations between observable quantities such as charged particles multiplicity n, transverse momentum pt, net charge q etc., in distant rapidity intervals. The SPbSU team now actively searches for the long-range n-n, pt-n, pt-pt and qq correlations. Such types of correlations are investigated in pp and AA collisions, string fusion effect is also investigated in Monte Carlo calculations using Grid and PROOF facilities. The presence of such long-range correlations can be also considered as one of the signatures of the early stages of quark-gluon plasma formation. Experimental studies of the long-range correlations in pp and AA collisions are proposed for ALICE at LHC and NA61 at SPS experiments, CERN [8, 9].

Distributed data processing in PROOF and Grid

The main goal of our analysis is the event-by-event processing of reconstructed or simulated ALICE experimental data in ESD format (Event Summary Data). To complete physical program with high statistics of the data Grid calculations are essential. In order to get the reliable results we have to deal with analysis at the level of about not less than 10^6 high-multiplicity events. Meanwhile Grid computing takes a lot of time and it is very unpractical to wait for hours for each debugging run. The most convenient case for code debugging is local analysis, but it implies processing of very small

number of events, especially in case of AA collisions. PROOF calculations take an intermediate place. It is suitable for an intermediate statistics and allows working in an interactive mode. So the 3-stages data processing strategy "local → PROOF → Grid" was accepted. The key issue here is AliAnalysisManager class [2] in AliROOT, which provides the unified access to all three ways of analysis. It makes possible to run one code containing physics logic in all execution systems transparently, one just needs to change execution scripts. The same benefit appears in Monte-Carlo calculations which were also performed in our analysis. Dwell on each stage of analysis more detailed.

Stage 1 – Local analysis

At this first stage a new code is created and a first debugging is performed in order to compile and run the code. Local analysis is very convenient for debugging due to the absence of latency, providing real-time feedback. Most of code writing is done at this stage; however, it is not possible to debug the physics analysis code using the local PC, especially in case of AA collisions. One can't debug physics logic using the statistics of tens or hundreds events.

Stage 2 – data analysis using PROOF

When the primary debugging of code is done, we need to ensure that physical ideas are implemented correctly. The second stage should be applied with the debugging output in several runs using a significant statistics dataset. These runs are needed to check if the program is calculating the results in line with the initial physics ideas. There could be always some errors leading to non realistic results, and to see those errors one needs a statistics that can provide at least preliminary conclusions. Analysis using the parallel ROOT facility (PROOF) [3, 4] suits these requirements perfectly. PROOF is a software system enabling ROOT-based analysis and processing in parallel on distributed resources. This system optimizes the execution time by implementing data parallelism at event-level and provides real-time feedback, which is very important for code developing. At the same time PROOF allows one to get results with statistics up to 10^6 events on different runs, which is quite enough to get preliminary results. At this stage the tuning and checking of main parameters is also performed (cutting thresholds, histogram limits etc.). Additional software required by the algorithm to be run can be loaded directly to the system in optimized way.

Nevertheless, the PROOF analysis has some drawbacks, and unfortunately main of them is insufficient stability of operation. At this stage it is harder to debug code, and some PROOF specific errors can appear like merging problems. Also there is still insufficient statistics to get final results.

List of ALICE PROOF clusters, their operation status, current workload and other parameters can be found at http://alimonitor.cern.ch/stats?page=PROOF/list

Stage 3 – Grid analysis

At this stage one performs last checks of physics logic and gets final results using full statistics of data available in AliEn system.

AliEn [5, 6] – ALICE Environment – is a set of middleware tools and services that implement a Grid infrastructure. The development of AliEn started in 2000 by ALICE collaboration, and it was deployed for distributed Monte Carlo productions on several remote computing sites. From 2005 AliEn has been used both for data production and end-user analysis. In spite of the fact that Grid technologies were rapidly evolving, AliEn has very successfully served its primary goal of hiding from the end user the complexity and heterogeneity of all underlying Grid services. The main AliEn components are as follows:

- File Catalogue
  - – UNIX-like file system interface
  - – Entries are retrieved by LFN (Logical File Name) or GUID (Globally Unique IDentifier)
  - – Mapping to physical files
  - – Powerful metadata catalogue
  - – Automatic Storage Element selection
  - – 4 storage technologies: CASTOR2, dCache, DPM, Scalla

– Multiple storage protocols: Xrootd, torrent, srm, file
• Authentification & authorization
• ROOT interface
• VO-box system
• Monitoring based on MonALISA software [10], package management
• Workload management using TaskQueue, Job Agents & pull model

AliEn system provides access to all Grid data and allows us to get the final reliable results with statistics up to 10^7 events. On the other hand usage of AliEn system implies high latency and queues. Sometimes Grid specific errors appear, mainly validation and execution problems for some jobs, which are very difficult to trace and debug.

ALICE Computing in SPbSU

Cluster RU-SPbSU is in stable operation since 2004 running AliEn middleware. In 2005 it was registered in RDIG and started its operation in the Large Hadron Collider Grid. In order to use the site as storage of Tier-2 level, the Xrootd middleware was installed. After the significant upgrades according to EGEE program in 2007, 2008 and 2010 RU-SPbSU cluster supports 4 Virtual Organizations (ATLAS, CMS, LHCb, ALICE). In 2011 control servers and then working nodes were moved to the virtual VMware Vsphere machines to form a part of University cloud. The POD (Proof On Demand) cluster is being prepared on the same cloud for ALICE experimental data processing by SPbSU researchers.

By 2012 cluster provides ALICE 62 TB of disk space and 146 computing cores, 96 of those are also available for 3 other VOs. Both the cluster and the LCG middleware demonstrate stable performance and functioning in the Worldwide LHC Computing Grid (WLCG).

Experimental data processing results

Correlations n-n, pt-n and pt-pt were investigated in pp collisions in ALICE at the energies of 0.9 and 7 TeV. Experimental results on the correlation coefficient behavior against the width of pseudorapidity windows and the gap between them are shown in Fig. 1 (plots were approved as ALICE Preliminary). Non-zero p_T-N and p_T-p_T correlations coefficients can be considered as a signature of quark-gluon strings fusion and QGP formation on the early stages of proton-proton collisions.

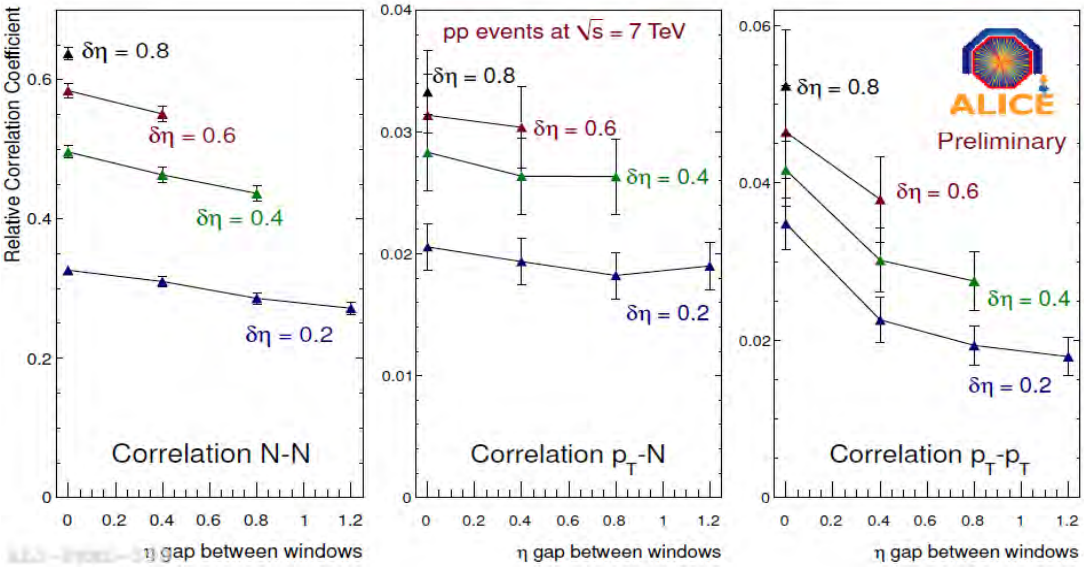

Fig. 1: The dependence of long-range N-N, p_T-N and p_T-p_T correlations on the pseudorapidity gap between windows in pp collisions at 7 TeV, measured for different widths δη of the observation windows [11]. Normalized observables. Lines are drawn to guide the eye

Additional studies in the Grid framework of long-range correlations were performed in azimuthally separated windows in order to obtain cleaner information that is less affected by short-range effects. The dependence of correlation coefficients on azimuthal configuration of windows is shown in Fig. 2.

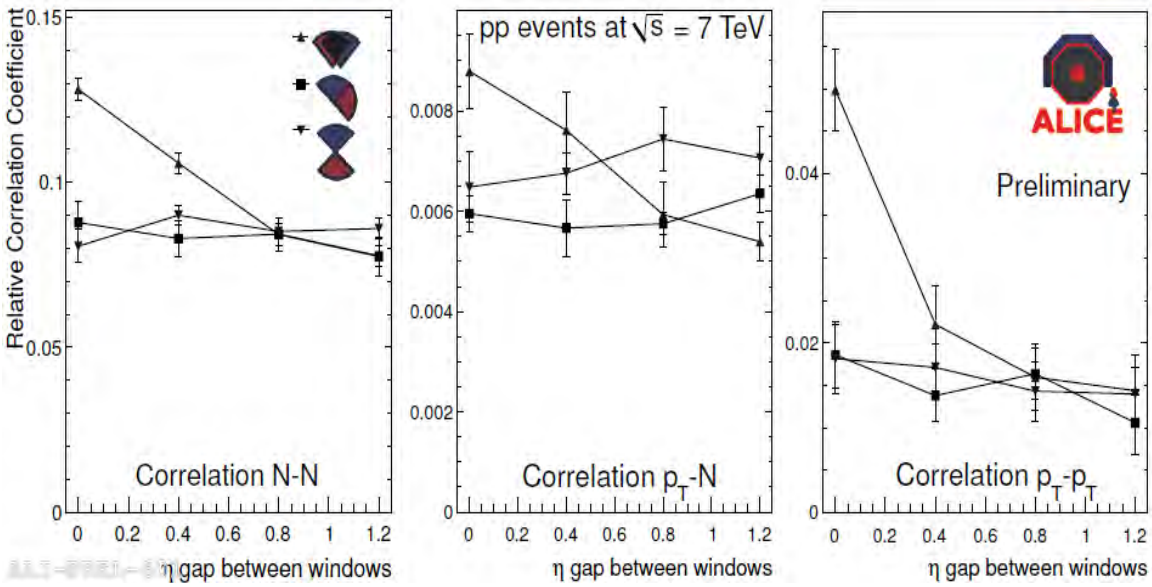

Fig.2: The same as in Fig.1 but for different configurations of backward and forward π/2 azimuth sectors (for the width of the pseudorapidity windows – δη=0.2) [11]. Relative orientations of sectors are marked by color

Monte-Carlo calculation results

The general scheme of the AliAnalysis framework is applied also to Monte Carlo calculations. It implies quick developing, testing and debugging of Monte Carlo algorithms. It permits us to obtain quickly the results for the different Monte Carlo models of heavy ion collisions, try different parameters and assumptions. All this makes it possible to get a better physical understanding of the processes studied.

We implemented the model with color string formation and fusion for pp and AA collisions at the LHC energies and performed event-by-event simulations for pp and PbPb collisions.

In order to perform event-by-event Monte Carlo calculation, three abstraction levels of the algorithms were taken:

1) The core of Monte Carlo generator, which covers the simulation of a collision and corresponding mathematical procedures.
2) AliAnalysis task class, that runs the generator events, collects the simulated data, performs initial statistical processing
3) Final data processing and plotting the results.

In addition, there are several configuration and run scripts.

Monte Carlo calculations for heavy ion collisions with high statistics demand computational power comparable to one of used for data analysis. In order to submit tasks to the Grid (WLCG) or perform PROOF analysis in the AliAnalysis framework some datasets should be provided, but during calculations the actual data are ignored. As a result pure MC simulations are performed in both PROOF and Grid mode.

In Fig. 3 examples of N-N and p_T-N correlation functions are shown for pp collisions at 7 TeV. The plots demonstrate decrease of N-N correlation strength (the slope) and non-zero p_T-N correlation with taking into account the string fusion.

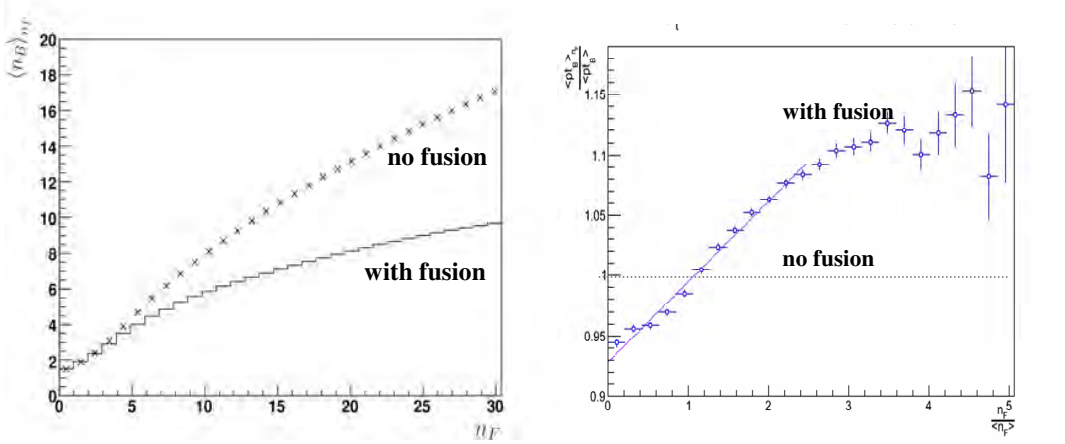

Fig. 3: Examples of N-N (left) and p_T-N (right) correlation functions calculated in Monte-Carlo simulations for pp collisions at 7 TeV. Configuration of pseudorapidity windows is (–0.8, 0) (0, 0.8)

Summary

1) Grid technologies allows us to perform an event-by-event long-range correlations analysis with statistics up to 10^7 events and thereby to get the reliable physical results.

2) Our experience with the ALICE data analysis shows importance of combination of Grid with other interactive High Performance Computing technologies such as PROOF, especially for debugging reasons.

3) Analysis framework can be used also for our own Monte Carlo calculations with high statistics.

References

[1] ALICE Collaboration, "ALICE: Physics Performance Report" - Volume 2, CERN/LHCC 2005-030; ALICE PPR Volume II, 5 December 2005; CERN. (J.Phys. G32 (2006) 1295-2040 [Section: 6.5.15 - Long-range correlations, p.1749-1751])

[2] The Analysis Framework http://aliweb.cern.ch/Offline/Activities/Analysis/AnalysisFramework/index.html

[3] The Parallel ROOT Facility, PROOF http://root.cern.ch/drupal/content/proof

[4] Data Analysis with PROOF, G Ganis, J. Iwaszkiewicz, F. Rademakers, Proceedings of ACAT 2008 Conference

[5] ALICE Environment: Open Source GRID Framework http://alien2.cern.ch/

[6] AliEn: ALICE environment on the GRID, S. Bagnasco, L. Betev, P. Buncic, F. Carminati, C. Cirstoiu, C. Grigoras, A. Hayrapetyan, A. Harutyunyan, A.J. Peters, P. Saiz; CERN. International Conference on Computing in High Energy and Nuclear Physics (CHEP'07); 2008 J. Phys.: Conf. Ser. 119 062012

[7] M. A. Braun and C. Pajares, Phys. Lett. B 287, 154 (1992); Nucl. Phys. B 390, 542, 549 (1993).

[8] ALICE Collaboration, "ALICE: Physics Performance Report" - Volume 2, CERN/LHCC 2005-030; ALICE PPR Volume II, 5 December 2005; CERN. (Journ.Phys.G: Nuclear and Particle Phys., 2006, 733 pages)

[9] SHINE/NA61 proposal: Study of hadron production in hadron nucleus and nucleus nucleus collisions at the CERN SPS. By NA49–future Collaboration (N. Antoniou et al.). CERN–SPS – 2006–034, CERN–SPSC–P–330.

[10] MonALISA: An Agent Based, Dynamic Service System to Monitor, Control and Optimize Grid based Applications, I.C. Legrand, H.B. Newman, R. Voicu, C. Cirstoiu, C. Grigoras, M. Toarta, C. Dobre, CHEP 2004, Interlaken, Switzerland.

[11] G.Feofilov (for the ALICE Collaboration), "Long-Range (Forward-Backward) pt and Multiplicity Correlations in pp collisions at 0.9 and 7 TeV", poster report at QM-2011, 23-28 May 2011, Annecy, France.