

# ALICE Collaboration

---



# Contents

---

<b>6</b>	<b>Data analysis</b>	<b>6</b>
6.1	Introduction . . . . .	6
6.1.1	The analysis activity . . . . .	6
6.2	Organization of the data analysis . . . . .	7
6.3	Infrastructure tools for distributed analysis . . . . .	8
6.3.1	gShell . . . . .	8
6.3.2	PROOF – the Parallel ROOT Facility . . . . .	8
6.4	Analysis tools . . . . .	9
6.4.1	Statistical tools . . . . .	9
6.4.2	Calculations of kinematics variables . . . . .	9
6.4.3	Geometrical calculations . . . . .	9
6.4.4	Global event characteristics . . . . .	10
6.4.5	Comparison between reconstructed and simulated parameters . . . . .	10
6.4.6	Event mixing . . . . .	10
6.4.7	Analysis of the High-Level Trigger (HLT) data . . . . .	11
6.4.8	Visualization . . . . .	11
6.5	Existing analysis examples in AliRoot . . . . .	11
	<b>References</b>	<b>15</b>



# 6 Data analysis

---

## 6.1 Introduction

The analysis of experimental data is the final stage of event processing and it is usually repeated many times. Analysis is a very diverse activity, where the goals of each particular analysis pass may differ significantly.

The ALICE detector is optimized for the reconstruction and analysis of heavy-ion collisions. In addition, ALICE has a broad physics programme devoted to pp and pA interactions.

The main points of the ALICE heavy-ion programme can be summarized as follows [1]:

- **global event characteristics:** particle multiplicity, centrality, energy density, nuclear stopping;
- **soft physics:** chemical composition (particle and resonance production, particle ratios and spectra, strangeness enhancement), reaction dynamics (transverse and elliptic flow, HBT correlations, event-by-event dynamical fluctuations);
- **hard probes:** jets, direct photons;
- **heavy flavours:** quarkonia, open charm and beauty production.

The pp and pA programme will provide, on the one hand, reference points for comparison with heavy ions. On the other hand, ALICE will also pursue genuine and detailed pp studies. Some quantities, in particular the global characteristics of interactions, will be measured during the first days of running exploiting the low-momentum measurement and particle identification capabilities of ALICE.

### 6.1.1 The analysis activity

We distinguish two main types of analysis: scheduled analysis and chaotic analysis. They differ in their data access pattern, in the storage and registration of the results, and in the frequency of changes in the analysis code. The detailed definition is given in Section 6.2.

In the ALICE Computing Model the analysis starts from the Event Summary Data (ESD). These are produced during the reconstruction step and contain all the information for the analysis. The size of the ESD is about one order of magnitude lower than the corresponding raw data. The analysis tasks produce Analysis Object Data (AOD) specific to a given set of physics objectives. Further passes for the specific analysis activity can be performed on the AODs, until the selection parameter or algorithms are changed.

A typical data analysis task usually requires processing of selected sets of events. The selection is based on the event topology and characteristics, and is done by querying the tag database (see Chapter ??). The tags represent physics quantities which characterize each run and event, and permit fast selection. They are created after the reconstruction and contain also the unique identifier of the ESD file. A typical query, when translated into natural language, could look like “Give me all the events with impact parameter in <range> containing jet candidates with energy larger than <threshold>”. This results in a list of events and file identifiers to be used in the consecutive event loop.

The next step of a typical analysis consists of a loop over all the events in the list and calculation of the physics quantities of interest. Usually, for each event, there is a set of embedded loops on the reconstructed entities such as tracks,  $V^0$  candidates, neutral clusters, etc., the main goal of which is to select the signal candidates. Inside each loop a number of criteria (cuts) are applied to reject the background combinations and to select the signal ones. The cuts can be based on geometrical quantities

such as impact parameters of the tracks with respect to the primary vertex, distance between the cluster and the closest track, distance of closest approach between the tracks, angle between the momentum vector of the particle combination and the line connecting the production and decay vertices. They can also be based on kinematics quantities such as momentum ratios, minimal and maximal transverse momentum, angles in the rest frame of the combination. Particle identification criteria are also among the most common selection criteria.

The optimization of the selection criteria is one of the most important parts of the analysis. The goal is to maximize the signal-to-background ratio in case of search tasks, or another ratio (typically  $\text{Signal}/\sqrt{\text{Signal} + \text{Background}}$ ) in case of measurement of a given property. Usually, this optimization is performed using simulated events where the information from the particle generator is available.

After the optimization of the selection criteria, one has to take into account the combined acceptance of the detector. This is a complex, analysis-specific quantity which depends on the geometrical acceptance, the trigger efficiency, the decays of particles, the reconstruction efficiency, the efficiency of the particle identification and of the selection cuts. The components of the combined acceptance are usually parametrized and their product is used to unfold the experimental distributions or during the simulation of some model parameters.

The last part of the analysis usually involves quite complex mathematical treatments, and sophisticated statistical tools. Here one may include the correction for systematic effects, the estimation of statistical and systematic errors, etc.

## 6.2 Organization of the data analysis

The data analysis is coordinated by the Physics Board via the Physics Working Groups (PWGs). At present the following PWG have started their activity:

- detector performance;
- global event characteristics and soft physics (including proton–proton physics);
- hard probes: jets and direct photons;
- heavy flavours.

### Scheduled analysis

The scheduled analysis typically uses all the available data from a given period, and stores and registers the results using Grid middleware. The tag database is updated accordingly. The AOD files, generated during the scheduled analysis, can be used by several subsequent analyses, or by a class of related physics tasks. The procedure of scheduled analysis is centralized and can be considered as data filtering. The requirements come from the PWGs and are prioritized by the Physics Board taking into account the available computing and storage resources. The analysis code is tested in advance and released before the beginning of the data processing.

Each PWG will require several sets of AOD per event, which are specific for one or a few analysis tasks. The creation of the AOD sets is managed centrally. The event list of each AOD set will be registered and the access to the AOD files will be granted to all ALICE collaborators. AOD files will be generated via Grid tools at different computing centres and will be stored on the corresponding storage elements. The processing of each file set will thus be done in a distributed way on the Grid. Some of the AOD sets may be quite small and would fit on a single storage element or even on one computer; in this case the corresponding tools for file replication, available in the ALICE Grid infrastructure, will be used.

### Chaotic analysis

The chaotic analysis is focused on a single physics task and typically is based on the filtered data from the scheduled analysis. Each physicist also may access directly large parts of the ESD in order to search for rare events or processes. Usually the user develops the code using a small subsample of data, and changes the algorithms and criteria frequently. The analysis macros and software are tested many times on relatively small data volumes, both experimental and Monte Carlo. The output is often only a set of histograms. Such a tuning of the analysis code can be done on a local data set or on distributed data using Grid tools. The final version of the analysis will eventually be submitted to the Grid and will access large portions or even the totality of the ESDs. The results may be registered in the Grid file catalogue and used at later stages of the analysis. This activity may or may not be coordinated inside the PWGs, via the definition of priorities. The chaotic analysis is carried on within the computing resources of the physics groups.

## 6.3 Infrastructure tools for distributed analysis

### 6.3.1 gShell

The main infrastructure tools for distributed analysis have been described in Chapter 3. The actual middleware is hidden by an interface to the Grid, gShell [2], which provides a single working shell. The gShell package contains all the commands a user may need for file catalogue queries, creation of sub-directories in the user space, registration and removal of files, job submission and process monitoring. The actual Grid middleware is completely transparent to the user.

The gShell overcomes the scalability problem of direct client connections to databases. All clients connect to the gLite [3] API services. This service is implemented as a pool of preforked server daemons, which serve single-client requests. The client-server protocol implements a client state which is represented by a current working directory, a client session ID and time-dependent symmetric cipher on both ends to guarantee client privacy and security. The server daemons execute client calls with the identity of the connected client.

### 6.3.2 PROOF – the Parallel ROOT Facility

The Parallel ROOT Facility, PROOF [4] has been specially designed and developed to allow the analysis and mining of very large data sets, minimizing response time. It makes use of the inherent parallelism in event data and implements an architecture that optimizes I/O and CPU utilization in heterogeneous clusters with distributed storage. The system provides transparent and interactive access to terabyte-scale data sets. Being part of the ROOT framework, PROOF inherits the benefits of a performing object storage system and a wealth of statistical and visualization tools. The most important design features of PROOF are:

- transparency – no difference between a local ROOT and a remote parallel PROOF session;
- scalability – no implicit limitations on number of computers used in parallel;
- adaptability – the system is able to adapt to variations in the remote environment.

PROOF is based on a multi-tier architecture: the ROOT client session, the PROOF master server, optionally a number of PROOF sub-master servers, and the PROOF worker servers. The user connects from the ROOT session to a master server on a remote cluster and the master server creates sub-masters and worker servers on all the nodes in the cluster. All workers process queries in parallel and the results are presented to the user as coming from a single server.

PROOF can be run either in a purely interactive way, with the user remaining connected to the master and worker servers and the analysis results being returned to the user's ROOT session for further

analysis, or in an ‘interactive batch’ way where the user disconnects from the master and workers (see Fig. ?? on page ??). By reconnecting later to the master server the user can retrieve the analysis results for that particular query. This last mode is useful for relatively long running queries (several hours) or for submitting many queries at the same time. Both modes will be important for the analysis of ALICE data.

## 6.4 Analysis tools

This section is devoted to the existing analysis tools in ROOT and AliRoot. As discussed in the introduction, some very broad analysis tasks include the search for some rare events (in this case the physicist tries to maximize the signal-over-background ratio), or measurements where it is important to maximize the signal significance. The tools that provide possibilities to apply certain selection criteria and to find the interesting combinations within a given event are described below. Some of them are very general and are used in many different places, for example the statistical tools. Others are specific to a given analysis.

### 6.4.1 Statistical tools

Several commonly used statistical tools are available in ROOT [5]. ROOT provides classes for efficient data storage and access, such as trees and ntuples. The ESD information is organized in a tree, where each event is a separate entry. This allows a chain of the ESD files to be made and the elaborated selector mechanisms to be used in order to exploit the PROOF services. Inside each ESD object the data is stored in polymorphic containers filled with reconstructed tracks, neutral particles, etc. The tree classes permit easy navigation, selection, browsing, and visualization of the data in the branches.

ROOT also provides histogramming and fitting classes, which are used for the representation of all the one- and multi-dimensional distributions, and for extraction of their fitted parameters. ROOT provides an interface to powerful and robust minimization packages, which can be used directly during some special parts of the analysis. A special fitting class allows one to decompose an experimental histogram as a superposition of source histograms.

ROOT also has a set of sophisticated statistical analysis tools such as principal component analysis, robust estimator, and neural networks. The calculation of confidence levels is provided as well.

Additional statistical functions are included in `TMath`.

### 6.4.2 Calculations of kinematics variables

The main ROOT physics classes include 3-vectors and Lorentz vectors, and operations such as translation, rotation, and boost. The calculations of kinematics variables such as transverse and longitudinal momentum, rapidity, pseudorapidity, effective mass, and many others are provided as well.

### 6.4.3 Geometrical calculations

There are several classes which can be used for measurement of the primary vertex: `AliITSVertexerZ`, `AliITSVertexerPPZ`, `AliITSVertexerIons`, `AliITSVertexerTracks`. A fast estimation of the  $z$ -position can be done by `AliITSVertexerZ`, which works for both lead–lead and proton–proton collisions. An universal tool is provided by `AliITSVertexerTracks`, which calculates the position and covariance matrix of the primary vertex based on a set of tracks, and also estimates the  $\chi^2$  contribution of each track. An iterative procedure can be used to remove the secondary tracks and improve the precision.

Track propagation to the primary vertex (inward) is provided in `AliESDtrack`.

The secondary vertex reconstruction in case of  $V^0$  is provided by `AliV0vertexer`, and in case of cascade hyperons by `AliCascadeVertexer`. An universal tool is `AliITSVertexerTracks`, which can

be used also to find secondary vertices close to the primary one, for example decays of open charm like  $D^0 \rightarrow K^- \pi^+$  or  $D^+ \rightarrow K^- \pi^+ \pi^+$ . All the vertex reconstruction classes also calculate distance of closest approach (DCA) between the track and the vertex.

The calculation of impact parameters with respect to the primary vertex is done during the reconstruction and the information is available in `AliESDtrack`. It is then possible to recalculate the impact parameter during the ESD analysis, after an improved determination of the primary vertex position using reconstructed ESD tracks.

#### 6.4.4 Global event characteristics

The impact parameter of the interaction and the number of participants are estimated from the energy measurements in the ZDC. In addition, the information from the FMD, PMD, and T0 detectors is available. It gives a valuable estimate of the event multiplicity at high rapidities and permits global event characterization. Together with the ZDC information it improves the determination of the impact parameter, number of participants, and number of binary collisions.

The event plane orientation is calculated by the `AliFlowAnalysis` class.

#### 6.4.5 Comparison between reconstructed and simulated parameters

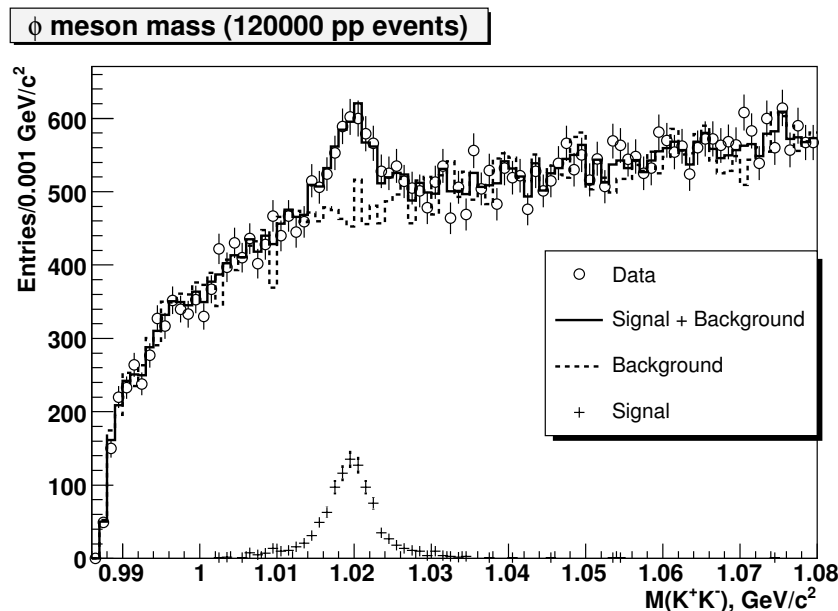
The comparison between the reconstructed and simulated parameters is an important part of the analysis. It is the only way to estimate the precision of the reconstruction. Several example macros exist in AliRoot and can be used for this purpose: `AliTPCComparison.C`, `AliITSComparisonV2.C`, etc. As a first step in each of these macros the list of so-called ‘good tracks’ is built. The definition of a good track is explained in detail in the ITS [6] and TPC [7] Technical Design Reports. The essential point is that the track goes through the detector and can be reconstructed. Using the ‘good tracks’ one then estimates the efficiency of the reconstruction and the resolution.

Another example is specific to the MUON arm: the `MUONRecoCheck.C` macro compares the reconstructed muon tracks with the simulated ones.

There is also the possibility to calculate directly the resolutions without additional requirements on the initial track. One can use the so-called track label and retrieve the corresponding simulated particle directly from the particle stack (`AliStack`).

#### 6.4.6 Event mixing

One particular analysis approach in heavy-ion physics is the estimation of the combinatorial background using event mixing. Part of the information (for example the positive tracks) is taken from one event, another part (for example the negative tracks) is taken from a different, but ‘similar’ event. The event ‘similarity’ is very important, because only in this case the combinations produced from different events represent the combinatorial background. Typically ‘similar’ in the example above means with the same multiplicity of negative tracks. One may require in addition similar impact parameters of the interactions, rotation of the tracks of the second event to adjust the event plane, etc. The possibility for event mixing is provided in AliRoot by the fact that the ESD is stored in trees and one can chain and access simultaneously many ESD objects. Then the first pass would be to order the events according to the desired criterion of ‘similarity’ and to use the obtained index for accessing the ‘similar’ events in the embedded analysis loops. An example of event mixing is shown in Fig. 6.1. The background distribution has been obtained using ‘mixed events’. The signal distribution has been taken directly from the Monte Carlo simulation. The ‘experimental distribution’ has been produced by the analysis macro and decomposed as a superposition of the signal and background histograms.



**Figure 6.1:** Mass spectrum of the  $\phi$  meson candidates produced inclusively in the proton–proton interactions.

### 6.4.7 Analysis of the High-Level Trigger (HLT) data

This is a specific analysis which is needed in order to adjust the cuts in the HLT code, or to estimate the HLT efficiency and resolution. AliRoot provides a transparent way of doing such an analysis, since the HLT information is stored in the form of ESD objects in a parallel tree. This also helps in the monitoring and visualization of the results of the HLT algorithms.

### 6.4.8 Visualization

The visualization classes give the possibility for prompt inspection of the simulation and reconstruction results. The initial version of the visualization is available in the `AliDisplay` class. Another more elaborated module `DISPLAY` is under development.

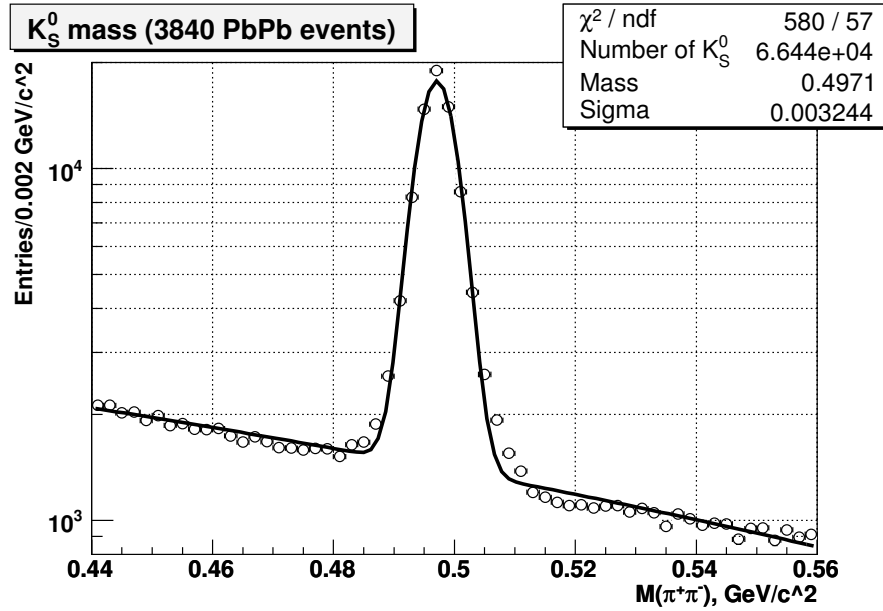
## 6.5 Existing analysis examples in AliRoot

There are several dedicated analysis tools available in AliRoot. Their results were used in the Physics Performance Report and described in ALICE internal notes. There are two main classes of analysis: the first one based directly on ESD, and the second one extracting first AOD, and then analysing it.

- **ESD analysis**

- **$V^0$  and cascade reconstruction/analysis**

The  $V^0$  candidates are reconstructed during the combined barrel tracking and stored in the ESD object. The following criteria are used for the selection: minimal-allowed impact parameter (in the transverse plane) for each track; maximal-allowed DCA between the two tracks; maximal-allowed cosine of the  $V^0$  pointing angle (angle between the momentum vector of the particle combination and the line connecting the production and decay vertices); minimal and maximal radius of the fiducial volume; maximal-allowed  $\chi^2$ . The last criterion requires the covariance matrix of track parameters, which is available only in `AliESDtrack`. The reconstruction is performed by `AliV0vertexer`. This class can be used also in the analysis. An example of reconstructed kaons taken directly from the ESDs is shown in Fig.6.2.



**Figure 6.2:** Mass spectrum of the  $K_S^0$  meson candidates produced inclusively in the Pb–Pb collisions.

The cascade hyperons are reconstructed using the  $V^0$  candidate and ‘bachelor’ track selected according to the cuts above. In addition, one requires that the reconstructed  $V^0$  effective mass belongs to a certain interval centred in the true value. The reconstruction is performed by `AliCascadeVertexer`, and this class can be used in the analysis.

### Open charm

This is the second elaborated example of ESD analysis. There are two classes, `AliD0toKpi` and `AliD0toKpiAnalysis`, which contain the corresponding analysis code. The decay under investigation is  $D^0 \rightarrow K^- \pi^+$  and its charge conjugate. Each  $D^0$  candidate is formed by a positive and a negative track, selected to fulfil the following requirements: minimal-allowed track transverse momentum, minimal-allowed track impact parameter in the transverse plane with respect to the primary vertex. The selection criteria for each combination include maximal-allowed distance of closest approach between the two tracks, decay angle of the kaon in the  $D^0$  rest frame in a given region, product of the impact parameters of the two tracks larger than a given value, pointing angle between the  $D^0$  momentum and flight-line smaller than a given value. The particle identification probabilities are used to reject the wrong combinations, namely  $(K, K)$  and  $(\pi, \pi)$ , and to enhance the signal-to-background ratio at low momentum by requiring the kaon identification. All proton-tagged tracks are excluded before the analysis loop on track pairs. More details can be found in Ref. [8].

### Quarkonia analysis

Muon tracks stored in the ESD can be analysed by the macro `MUONmassPlot_ESD.C`. This macro performs an invariant-mass analysis of muon unlike-sign pairs and calculates the combinatorial background. Quarkonia  $p_t$  and rapidity distribution are built for  $J/\psi$  and  $\Upsilon$ . This macro also performs a fast single-muon analysis:  $p_t$ , rapidity, and  $\theta$  vs  $\phi$  acceptance distributions for positive and negative muon tracks with a maximal-allowed  $\chi^2$ .

- **AOD analysis**

Often only a small subset of information contained in the ESD is needed to perform an analysis. This information can be extracted and stored in the AOD format in order to reduce the computing resources needed for the analysis.

The AOD analysis framework implements a set of tools like data readers, converters, cuts, and other utility classes. The design is based on two main requirements: flexibility and common AOD particle interface. This guarantees that several analyses can be done in sequence within the same computing session.

In order to fulfil the first requirement, the analysis is driven by the ‘analysis manager’ class and particular analyses are added to it. It performs the loop over events, which are delivered by an user-specified reader. This design allows the analyses to be ordered appropriately if some of them depend on the results of the others.

The cuts are designed to provide high flexibility and performance. A two-level architecture has been adopted for all the cuts (particle, pair and event). A class representing a cut has a list of ‘base cuts’. Each base cut implements a cut on a single property or performs a logical operation (and, or) on the result of other base cuts.

A class representing a pair of particles buffers all the results, so they can be re-used if required.

### **Particle momentum correlations (HBT) – HBTAN module**

Particle momentum correlation analysis is based on the event-mixing technique. It allows one to extract the signal by dividing the appropriate particle spectra coming from the original events by those from the mixed events.

Two analysis objects are currently implemented to perform the mixing: the standard one and the one implementing the Stavinsky algorithm [9]. Others can easily be added if needed.

An extensive hierarchy of the function base classes has been implemented facilitating the creation of new functions. A wide set of the correlation, distribution and monitoring functions is already available in the module. See Ref. [10] for the details.

The package contains two implementations of weighting algorithms, used for correlation simulations (the first developed by Lednicky [11] and the second due to CRAB [12]), both based on an uniform interface.

### **Jet analysis**

The jet analysis [13] is available in the module JETAN. It has a set of readers of the form `AliJetParticlesReader<XXX>`, where `XXX = ESD, HLT, KineGoodTPC, Kine`, derived from the base class `AliJetParticlesReader`. These provide an uniform interface to the information from the kinematics tree, from HLT, and from the ESD. The first step in the analysis is the creation of an AOD object: a tree containing objects of type `AliJetEventParticles`. The particles are selected using a cut on the minimal-allowed transverse momentum. The second analysis step consists of jet finding. Several algorithms are available in the classes of the type `Ali<XXX>JetFinder`. An example of AOD creation is provided in the `createEvents.C` macro. The usage of jet finders is illustrated in `findJets.C` macro.

### **$V^0$ AODs**

The AODs for  $V^0$  analysis contain several additional parameters, calculated and stored for fast access. The methods of the class `AliAODv0` provide access to all the geometrical and kinematics parameters of a  $V^0$  candidate, and to the ESD information used for the calculations.

### **MUON**

There is also a prototype MUON analysis provided in `AliMuonAnalysis`. It simply fills several histograms, namely the transverse momentum and rapidity for positive and negative muons, the invariant mass of the muon pair, etc.

The analysis framework is one of the most active fields of development. Many new classes and macros are in preparation. Some of them are already tested on the data produced during the Physics Data Challenge 2004 and will become part of the ALICE software.



# References

---

## Chapter 6

- [1] CERN/LHCC 2003-049, ALICE Physics Performance Report, Volume 1 (7 November 2003); ALICE Collaboration: F. Carminati *et al.*, J. Phys. G: Nucl. Part. Phys. **30** (2004) 1517–1763.
- [2] [http://alien.cern.ch/download/current/gClient/gShell\\_Documentation.html](http://alien.cern.ch/download/current/gClient/gShell_Documentation.html).
- [3] <http://glite.web.cern.ch/glite/>.
- [4] <http://root.cern.ch/root/PROOF.html>.
- [5] <http://root.cern.ch>.
- [6] CERN/LHCC 99-12.
- [7] CERN/LHCC 2000-001.
- [8] A. Dainese, PhD Thesis, University of Padova, 2003, [arXiv:nucl-ex/0311004].
- [9] A. Stavinsky *et al*, NUKLEONIKA **49** (Supplement 2) (2004) 23–25;  
[http://www.ichtj.waw.pl/ichtj/nukleon/back/full/vol49\\_2004/v49s2p023f.pdf](http://www.ichtj.waw.pl/ichtj/nukleon/back/full/vol49_2004/v49s2p023f.pdf).
- [10] P.K. Skowroński for ALICE Collaboration, [arXiv:physics/0306111].
- [11] R. Lednický and V.L. Lyuboshitz, Sov. J. Nucl. Phys. **35** (1982) 770.
- [12] <http://www.nscl.msu.edu/pratt/freecodes/crab/home.html>.
- [13] C. Loizides, PhD Thesis, University of Frankfurt, 2005, [arXiv:nucl-ex/0501017].

