Journal of Physics: Conference Series **513** (2014) 032102 doi:10.1088/1742-6596/513/3/032102

PROOF-based analysis on the ATLAS Grid facilities: first experience with the PoD/PanDa plugin

E. Vilucchi[3], A. De Salvo[1], C. Di Donato[2], R. Di Nardo[3], A. Doria[2], G. Ganis[4], A.Manafov[5], G. Mancini[3], S. Mazza[6], F. Preltz[6], D. Rebatto[6], A. Salvucci[7], A. R. Sanchez Pineda[3]

[1] INFN sezione di Roma1, piazzale Aldo Moro, 00146 Roma (RM), Italy
[2] INFN sezione di Napoli, Complesso universitario di Monte SantAngelo, via Cinthia, 80126 Napoli (NA), Italy
[3] INFN Laboratori Nazionali di Frascati, via E. Fermi 40, 00044 Frascati (RM), Italy
[4] CERN, CH - 1211 Geneva 23, Switzerland
[5] GSI Helmholtzzentrum für Schwerionenforschung GmbH, Planckstr. 1, 64291 Darmstadt, Germany
[6] INFN sezione di Milano, via Celoria 16, Milano (MI)
[7] Radboud University Nijmegen and Nikhef, The Netherlands

E-mail: elisabetta.vilucchi@lnf.infn.it

Abstract. In the ATLAS computing model Grid resources are managed by PanDA, the system designed for production and distributed analysis, and data are stored under various formats in ROOT files. End-user physicists have the choice to use either the ATHENA framework or directly ROOT, that provides users the possibility to use PROOF to exploit the computing power of multi-core machines or to dynamically manage analysis facilities. Since analysis facilities are, in general, not dedicated to PROOF only, PROOF-on-Demand (PoD) is used to enable PROOF on top of an existing resource management system.
In a previous work we investigated the usage of PoD to enable PROOF-based analysis on Tier-2 facilities using the PoD/gLite plug-in interface. In this paper we present the status of our investigations using the recently developed PoD/PanDA plug-in to enable PROOF and a real end-user ATLAS physics analysis as payload. For this work, data were accessed using two different protocols: XRootD and file protocol. The former in the site where the SRM interface is Disk Pool Manager (DPM) and the latter where the SRM interface is StoRM with GPFS file system. We will first describe the results of some benchmark tests we run on the ATLAS Italian Tier-1 and Tier-2s sites and at CERN. Then, we will compare the results of different types of analysis, comparing performances accessing data in relation to different types of SRM interfaces and accessing data with XRootD in the LAN and in the WAN using the ATLAS XROOTD storage federation infrastructure.

1. Introduction

In the ATLAS computing model [1] the sites providing computing facilities are organized in a hierarchical four-level tiered structure, with a Tier-0 distributed at CERN and at Wigner (HU), and 10 regional sites, Tier-1s, each one defining a cloud with a set of Tier-2s and Tier-3s. The Grid resources are managed by the PanDA system [2], a data-driven workload management system designed for production and distributed analysis based on a pilot paradigm, separating the pure job management issues from the infrastructure interaction. Data are stored under various formats in ROOT files and end-user physicists have the choice to use either the ATHENA framework or directly ROOT, that provides users the possibility to use PROOF [3] to exploit the computing power of multi-core machines or analysis facilities.

As discussed in [4], a convenient way to enable PROOF on top of not dedicated resources is provided by PoD (PROOF on Demand) [5]. In the study presented we used the recently developed PoD/PanDA plug-in which allows to start PROOF clusters on PanDA-managed resources, testing the system with three different analyses. The first two analyses presented (referred to as *analysis 1* and *analysis 2* in the following) are real analysis examples that use PROOF/PoD for the whole workflow. They use the same type of input dataset, but with a different approach for the code. The third analysis (called *analysis 3*), still a real analysis, uses a different sample that affects storage access performances, therefore representing a very different use case of the PROOF cluster, mixed with the Grid resources accessed in a "standard" way. Depending on the adopted storage resource manager (SRM), data are accessed using two different protocols: XRootD for Disk Pool Manager (DPM, [6]) and file protocol for StoRM [7] with GPFS underling file system. Both storage systems are accessed on the LAN and remotely on the WAN exploiting the FAX infrastructure (Federated ATLAS storage systems using XRootD [8, 9]). All tests were made on the Tier-1 and Tier-2s of the ATLAS Italian cloud and on the CERN analysis facilities.

2. The ATLAS Italian cloud: access protocols in different storage systems

The ATLAS Italian cloud is composed of the Tier-1 at CNAF (Bologna), four Tier-2s (Frascati, Milano, Napoli and Roma1) and eleven Tier-3s. There are differences in the storage solutions that affect the result of this work. The Tier-2s of Frascati, Napoli and Roma1 adopt DPM as Storage manager. DPM natively offers a XRootD daemon allowing direct file access with the XRoot protocol, so DPM sites could easily join the Federated ATLAS XRootD system. FAX is a storage federation aimed at presenting the Tier-1, Tier-2s and Tier-3s storage systems as a large single system and at providing data access via a single entrance, using XRootD's redirection technology. The HTTPS/WebDav protocol is implemented in the latest DPM releases.
The Tier-1, the Tier-2 of Milano and most of the grid-enabled Tier-3s use StoRM as SRM server over a GPFS file system, allowing running jobs to access data with standard POSIX I/O function calls. Moreover, if an XRootD access is also needed - e.g. to join FAX - the standard XRootD daemon can be deployed over GPFS, as it happened at the Tier-1 that is part of FAX for the Italian cloud. We also tested PoD on a sample stored in the CERN storage system managed with EOS, an exabyte scale storage system, XRootD based, adopted at CERN and designed for analysis-style data access [11].

3. Enabling PROOF on Tier-2 using PoD via PanDA

3.1. Enable a PROOF cluster via Panda

As in the previous work [3], PROOF was enabled with PoD. As a reminder, PoD is a tool-set which allows users to dynamically sets up a PROOF cluster at a user's request on any resource management system (RMS). It provides a plug-in based system in order to use different job submission front-ends. In [3] we used the gLite plug-in, the only one available at that time to cope with Grid resources. In the latest PoD release a new PanDA plug-in has been introduced in order to allow users to setup PROOF clusters on PanDa sites. The plug-in uses prun [10] to make a bulk submission in Panda of the required number of workers in the PanDA analysis queue. Prun is a PanDA-based tool for submitting grid jobs and, when jobs run on the workers, they start an xproofd daemon on the node, ready to accept PROOF connections from the master. The PROOF master is started from the PoD server on the user interface (UI) from which the user submits the PROOF cluster request. Because of the network topology of PROOF clusters set up by PoD - i.e. the PROOF master local or on a UI node - users must currently submit the requests to the site hosting the UI, otherwise the workers will not be able to connect back to the PoD server running on the master machine. A solution allowing request submission to any site will be available in the next PoD release.

3.2. Example of PoD at work

For illustration purposes we recall here the way PoD can submit to PanDA specifying the queue:

```
$  pod-submit -r panda -q ANALY_INFN-FRASCATI -n100.
```

The ATLAS user must have enabled PanDA and have obtained valid Grid credentials. A new feature of PoD, which was extensively used for this analysis, is the possibility to retrieve the startup latencies using the `pod-info` command:

```
$ pod-info -nl
8
worker pilatlas009@atlaswn050.lnf.infn.it:21001 (...) startup: 79s (...)
worker pilatlas009@atlaswn046.lnf.infn.it:21003 (...) startup: 82s (...)
worker pilatlas009@atlaswn042.lnf.infn.it:21005 (...) startup: 85s (...)
...
```

This information was used for the startup latency studies reported below. The PROOF analysis can start once a sufficient number of workers is available. Ideally, one would like to start the analysis just after the submission with the workers being automatically included as they become available. Support for this functionality has been introduced in PROOF only recently [11].

4. Start-up latency, calibration and analysis

4.1. Startup latency tests

These tests have been carried out to study the time required to allocate a certain number of job slots which, in the case of PoD, translate into worker nodes. This *startup latency* time depends on the status of the site, the priority of the user and the number of job slots requested. The number of required jobs was set to 100. A single test consists in a user requiring 100 workers and retrieving the startup time of each worker as mentioned above. Tests run by different users in different sites lead to results like those shown in Figure 1 (left) for the Italian Tier-1 site at CNAF. Figure 1 (left) shows that here is a waiting time for the first job, typically of the order

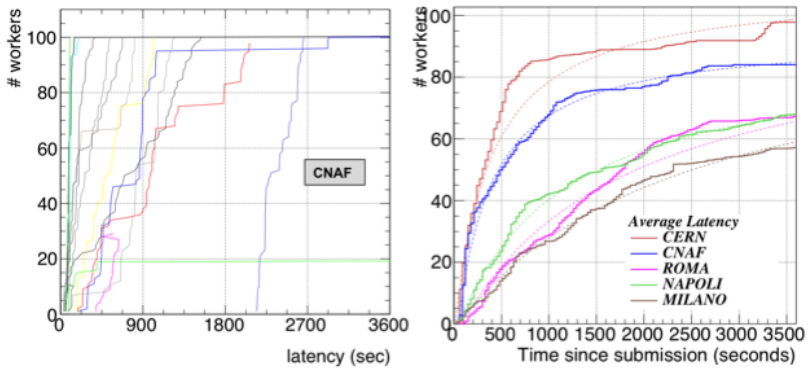

Figure 1. Startup latency at Tier-1/CNAF (left). Average startup latency (right).

of a few minutes, and that, once the request has received attention, the assignment rate is such that 100 jobs are typically available in the order of 10 minutes. However, as expected, the time distributions are broad, due to varying load conditions of the site. The other sites show similar behaviour, though the rates at which the jobs are assigned and the queuing time differs because of the different amount of available resources.

Figure 1 (right) shows the average startup latency for the five sites involved in the test. The larger amount of available resources at Tier-1 and CERN is reflected by the fact that the ramping

up slope is steeper than for the Tier-2 sites. Assuming that the rate at which job slots become available is proportional to the number of successful jobs and that the priority of assignment decreases with the number of jobs already started, we note that the curves can be described by the following formula:

$$n(t) = \frac{p_0 \cdot (t - t_0)}{1 + p_1 \cdot (t - t_0)} \quad (1)$$

with t_0 representing the queuing time for the first worker, p_0 being related to the rate of available slots and the ratio p_0/p_1 to the number of slots requested. The dotted lines in Figure 1 (right) show the result of an un-weighted fit to this formula. The results of the fits are given in Table 1. The fact that the simple formula above adapts reasonably well to the measured curves depends also on the fair-share assignment techniques adopted in the sites.

Site	t_0 [s]	p_0 [s^{-1}]	N_{jobs,p_0} [day]	$N_{jobs,PandA}$ [day]
CERN	15.4	0.282	24365	30321
CNAF	44.9	0.253	21859	18957
MILANO	45.7	0.041	3542	3728
NAPOLI	46.3	0.076	6566	6971
ROMA	79.0	0.054	4697	5630

Table 1. Numerical results of the fits shown in Figure 1 (right). The column N_{jobs,p_0} is the number of successful jobs per day derived from p_0. $N_{jobs,PanDA}$ is the number of successful jobs per day obtained from the PanDA monitoring system.

On average, in all cases the first worker starts within few minutes while the ramp time depends on the size of the site, with typical values for the Tier-2 sites and for Tier-1 and CERN sites. The number of estimated jobs per day is in good agreement with the ones retrieved from the PanDA monitoring.

4.2. Calibration tests

Before and after any analysis test we run calibration tests (disk access tests) performed in order to check the site efficiency and to obtain an upper limit on the performance in terms of MB/s as a function of the number of workers, using one of the most popular dataset formats in ATLAS. In addition the performance of different storage systems and different access protocols are also compared. We evaluate the access through the FAX infrastructure and some relatively new protocols like XRootD for GPFS and HTTP for DPM.

Calibration tests are run with PoD on the same datasets used for two of the three analysis: *analysis 1* and *analysis 2*. The calibration tool consists in a PROOF job running a dedicated TSelector reading from the files the whole content of each entry of the TTrees. The samples used for *analysis 1* and *analysis 2* are standard D3PDs and will be referred to as DS1 and DS2 respectively. DS1 is a dataset of ZZ MC sample, containing 90 files, corresponding approximately to 1 million events and with a total size of almost 100GB. DS2 is a MC sample of Standard Model Higgs boson decaying into photon pairs. The dataset contains 3 million 100KB-events composed of approximately 300 files totalizing 300 GB.

In Figure 2 we show the plot of the performance tests run at CNAF, Roma1 and CERN with DS1 and DS2. We measured the maximum input rate in MBytes/second using the PROOF statistics tools as a function of the number of workers. This quantity is derived from the number of bytes effectively read out from the files by the active workers divided by the total processing time; the latter includes event decompression and construction of the event information in memory, which is the only CPU load in this calibration test, giving a negligible effect to our study.

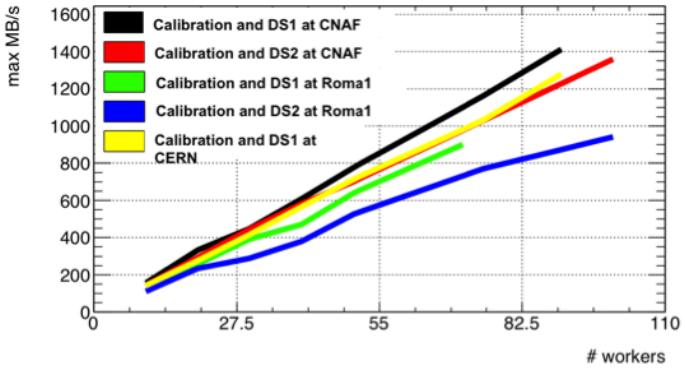

Figure 2. Max MB/s in function of number of allocated workers with input calibration DS1 and DS2 at CNAF, Roma1 and CERN.

The access protocol used at CNAF is GPFS that turns out to be more performing than XRootD on DPM at Roma1. CNAF and CERN have comparable performance. With these tests we show that the performance scales linearly up to 100 nodes, so it is reasonable to request up to 100 nodes to perform an analysis with PROOF in a Tier. Of course, performance in terms of MB/s depends on the storage system, the available protocols and the dataset structure.

4.3. Tests with analysis and access to the local storage

In this section we presents some of the tests performed running three different analyses in the Italian sites and at CERN, reading the input datasets from the local SRM. Analysis tests accessing input files from different sites will be discussed in section 4.4.

The first analysis is the Higgs boson selection in the four-lepton channel ($H \rightarrow ZZ^* \rightarrow 4l$). It is organized as a standalone C++ code without configuring the whole ATLAS software framework and RootCore. Since the number of branches contained in a D3PD TTree is huge (around 7500) and most of them are not used in this analysis, the code loads only the needed ones (around 500) in order to speed up and optimize the analysis process; in this way ROOT will not allocate memory for inactive branches and the cache and the reading process of the event is faster.

The second analysis is the selection of a Higgs decaying into two photons. *Analysis 2* is RootCore based with the EventLoop and D3PDReader packages. Using the D3PDReader package of RootCore, only the branches of the TTree used in the analysis are read from disk in an optimized way with respect to what it is done in the standard root TSelector. This reduces the I/O between the analysis and the data read from disk enhancing greatly the number of processed events per second. For this particular case, the RootCore based analysis allows a gain in speed close to a factor of 10 in terms of events processed per second with respect to the stand-alone version of the same analysis (using a standard TSelector) and for the same number of branches (100 branches).

The inputs files for *analysis 1* and *analysis 2* are the same datasets that we used for the calibration tests. Figure 3 (left) shows the curves of maximum MB/s as a function of the number of workers belonging to the PROOF clusters enabled at CNAF, Roma1 and at CERN to run *analysis 1*. The comparison of the results of the calibration with DS1 in Figure 2 with the first analysis in Figure 3 (left) shows that the weight of the additional event processing for this analysis is less than 20%. The plot in Figure 3 (right) shows the average of MB/s as a function of the number of workers at CNAF running *analysis 2*, and the performance is a consequence of the optimized file access. In fact, the rate of analyzed events for this analysis is higher that the rate of the first one. For example, with 75 workers the rate of *analysis 2* is 30,000 events

per second, while for *analysis 1* is 7,500 events per second. Results in Figure 3 (right) fluctuate

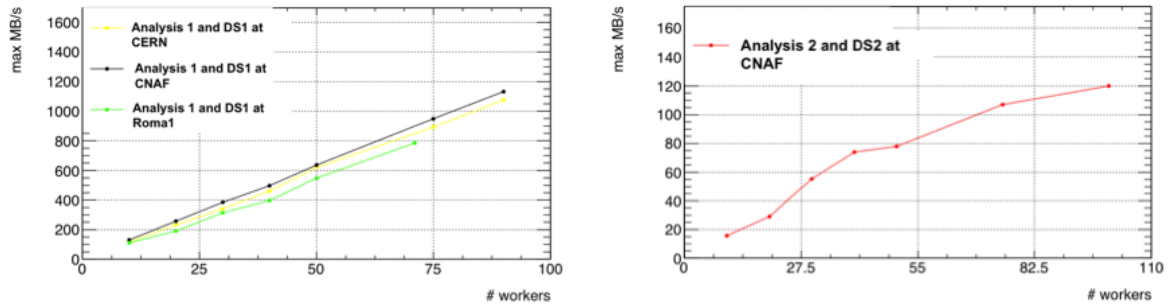

Figure 3. Max MB/s in function of number of workers running *analysis 1* in different sites (left). Average MB/s in function of number of workers at CNAF running *analysis 2* (right).

because of occasional problems while merging the processed dataset. Then, independently from the single performance of the two analyses in terms of megabytes per second, plots in Figure 3 show that I/O performance scale linearly with the number of workers and then there are no I/O limitation requesting a cluster of 100 workers to run a PROOF-based analysis. In Napoli's Tier-2 it has been also possible to use the http protocol to access the datasets running *analysis 2*. Some preliminary tests give the following performance results for 100 workers: 90 MB/s, 25567 Evts/s in average, then the http access methods seems to be slightly less performing than XRootD, but we do not have a case study large enough to make a conclusive statement.
Analysis 3 selects Higgs decaying in two leptons and two jets ($H \rightarrow ZZ \rightarrow llqq$), and it represents a different use case, because it uses the PROOF cluster to run over user-defined ntuples produced - with a "standard" Grid resources usage - in a previous step of the analysis. The input datasets used for *analysis 3*, DS3, is very different from DS1 and DS2, because it is a collection of small root ntuples (∼400) with an average size of 150MB/file and 0.2 KB per event. Then the rate of analyzed events per second is very high (around 1200000 Evts/s at CNAF with GPFS and 100 workers) and the plot in Figure 4 of the I/O scalability performance at CERN, CNAF, Roma1 and Napoli, running *analysis 3* , is a consequence of the files structure and the size. Nevertheless, Figure 4 shows that the performance is in agreement with expectation based on the number of workers and PoD can be successfully used also for this use case different from the first analyses.

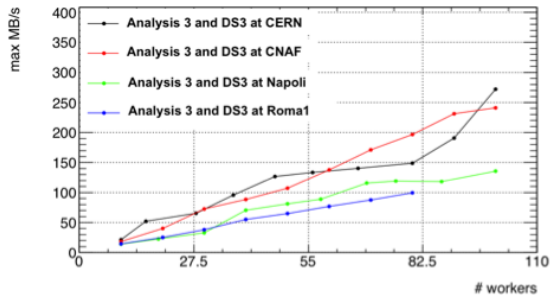

Figure 4. Max MB/s in function of number of workers running *analysis 3* at different sites.

4.4. Tests with analysis and remote storage access
We run analysis with PoD also accessing datasets from different sites, both with direct XRootD access (i.e. without using the redirection technology) and with a federated access through the XRootD redirector for the Italian cloud hosted at CERN. I/O performance results running *analysis 1* and *3*, with input data access over the WAN, are shown in plots of Figure 5 and depend on the network connections. We have also run the analysis with input datasets spread across different sites and accessed through the central FAX redirector. FAX works smoothly

and requests are forwarded to the appropriate storage elements, selecting the local SRM when possible. This federated access is transparent for the user.

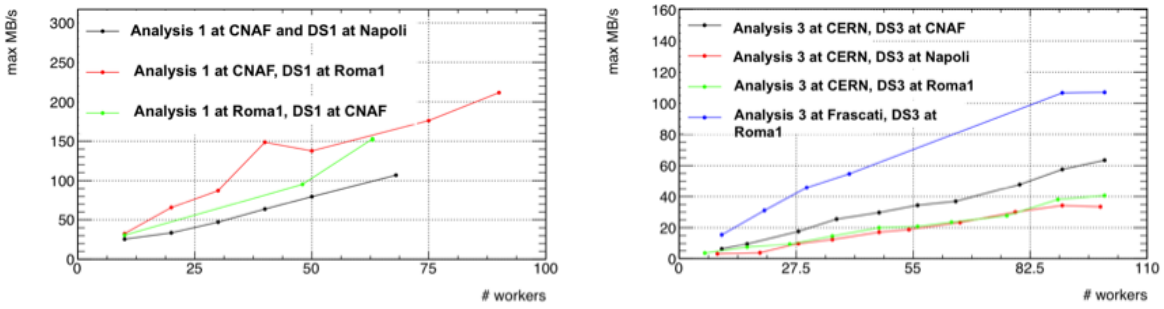

Figure 5. Max MB/s in function of the number of workers running *analysis 1* with DS1 over WAN (left). Max MB/s in function of the number of workers running *analysis 3* with DS3 over WAN (right).

5. Conclusions and future work

In this study we have used all the main components of the ATLAS analysis chain to evaluate the potential of PROOF-based analysis on non-dedicated resources. We successfully run real analysis examples at the same time of usual activities of an ATLAS Tier-2, competing with other users for analysis resources and showing that, at least up to 100 nodes, there is no saturation of the storage resources. In terms of startup latency, all the sites involved in the test have shown good responsiveness which PROOF would fully exploit when the currently experimental feature of adding workers while processing will be consolidated. The storage systems showed good scalability for the typical ranges of workers per session. We also showed that FAX works smoothly and made some initial tests using the HTTP protocol. For the future work we propose to evaluate PROOF with dynamic worker addition, focusing in particular on the advantages of the pull-based approach work distribution implemented by PROOF with the push architecture featured by Grids. We will also evaluate data access with the HTTPS/WebDav protocol.

References

[1] The ATLAS Collaboration 2008 The ATLAS experiment at the CERN Large Hadron Collider *J. Inst. 3 S08003*
[2] Maeno, T. et al. 2008 PanDA: distributed production and distributed analysis system for ATLAS *J. Phys.: Conf. Ser. 119*
[3] PROOF: Parallel ROOT `http://root.cern.ch/drupal/content/proof`.
[4] Di Nardo, R. et al. 2012 Enabling data analysis la PROOF on the Italian ATLAS Tier-2s using PoD *CHEP 2012: J. Phys.: Conf. Ser. 396 032043*
[5] PoD: `http://pod.gsi.de`
[6] Alvarez, A. et al. 2012 DPM: Future Proof Storage *J. Phys.: Conf. Ser. 396 032015*
[7] Corso, E., et al. 2006 StoRM: a grid storage resource manager *Cracow Grid Workshop (CGW2006), Cracow, Poland (2006)*
[8] Dorigo, A., et al. XROOTD/TXNetFile: a highly scalable architecture for data access in the ROOT environment *4th WSEAS International Conference on Telecommunications and Informatics*
[9] Bauerdick, L. 2012 Using Xrootd to Federate Regional Storage *J. Phys.: Conf. Ser. 396 042009*
[10] Prun: `https://twiki.cern.ch/twiki/bin/viewauth/AtlasComputing/PandaRun`
[11] EOS: `https://eos.cern.ch`.
[12] Berzano, D. et al. 2013 PROOF as a Service on the Cloud: a Virtual Analysis Facility based on theCernVM ecosystem *in these Proceedings*

Corrigendum: PROOF-based analysis on the ATLAS Grid facilities: first experience with the PoD/PanDa plugin

E. Vilucchi[3], A. De Salvo[1], C. Di Donato[2], R. Di Nardo[3], A. Doria[2], G. Ganis[4], A. Manafov[5], G. Mancini[3], S. Mazza[6], F. Preltz[6], D. Rebatto[6], A. Salvucci[7], A. R. Sanchez Pineda[2]

E-mail: elisabetta.vilucchi@lnf.infn.it

CORRIGENDUM TO: E Vilucchi *et al* 2014 *J. Phys.: Conf. Ser.* **513** 032102

The acknowledgements section was inadvertently omitted, it should be as it follows.

Acknowledgements
This work was developed in the framework of the PRIN Project "*STOA-LHC 20108T4XTM*", *CUP: I11J12000080001*, and partially supported by it.

[1] INFN sezione di Roma1, piazzale Aldo Moro, 00146 Roma (RM), Italy
[2] INFN sezione di Napoli, Complesso universitario di Monte Sant'Angelo, via Cinthia, 80126 Napoli (NA), Italy
[3] INFN Laboratori Nazionali di Frascati, via E. Fermi 40, 00044 Frascati (RM), Italy
[4] CERN, CH-1211 Geneva 23, Switzerland
[5] GSI Helmholtzzentrum für Schwerionenforschung GmbH, Planckstr. 1, 64291 Darmstadt, Germany
[6] INFN sezione di Milano, via Celoria 16, Milano (MI)
[7] Radboud University Nijmegen and Nikhef, Netherlands