The argo YBJ daq system and the GRID based data transfer

Article *in* IEEE Transactions on Nuclear Science · March 2008

DOI: 10.1109/TNS.2007.909988 · Source: IEEE Xplore

Paolo Branchini
INFN - Istituto Nazionale di Fisica Nucleare
1,607 PUBLICATIONS **25,104** CITATIONS

Antonio Budano
INFN - Istituto Nazionale di Fisica Nucleare
165 PUBLICATIONS **2,258** CITATIONS

S. Catalanotti
University of Naples Federico II
145 PUBLICATIONS **2,675** CITATIONS

Paola Celio
INFN - Istituto Nazionale di Fisica Nucleare
42 PUBLICATIONS **948** CITATIONS

The Argo YBJ Daq System and the GRID Based Data Transfer

Alberto Aloisio, Paolo Branchini, Antonio Budano, Sergio Catalanotti, Paola Celio, Pietro Creti, Fulvio Galeazzi, Riccardo Gargana, Giovanni Marsella, Stefano Mastroianni, Federico Ruggieri, and Cristian Stanescu

***Abstract*—The Argo-YBJ experiment has now reached its final design configuration. The detector system consists of a full coverage array (about 5800 square meters) of Resistive Plate Chambers (RPCs). The throughput depends on the trigger rate and threshold. The DAQ system must be able to sustain a maximum transfer rate of the order of 15 MB/s and a high peak data flow. Data are read out using a typical front-end acquisition chain built around a custom bus. Specialized electronics have been designed and dedicated software has been written to perform this task. Data are sent to the online farm through a switch exploiting a gigabit ethernet protocol. A solution to transfer data from the YBJ laboratory to the laboratories belonging to the Argo-YBJ collaboration exploiting the GRID middleware has also been implemented. In this paper we describe the daq and the data mover main characteristics and performance.**

***Index Terms*—Data acquisition, data mover, GRID, RPC.**

I. Introduction

The Argo-YBJ experiment is a collaboration of Chinese and Italian groups. The experiment site is located at the Yangbajing International Cosmic Rays Observatory at Yangbajing (Tibet, P.R. China) at 4300 meters above the sea level, about 90 km from Lhasa.

The Argo-YBJ ground-based detector allows for the investigation of many aspects in gamma-astronomy and cosmic ray physics, spanning a large energy range thanks to its ability to operate down to a few hundreds of GeV up to the PeV [1], [19].

The telescope is optimized for the detection of small size air showers to study gamma-rays from galactic and extra-galactic sources. It monitors the northern hemisphere in the declination band $-10^\circ < \delta < 70^\circ$.

The apparatus consists of an array of dimension $74 \times 78\ \mathrm{m}^2$, with an active area of about 92%, made of a single layer of Resistive Plate Chambers (RPCs) operating in streamer mode, surrounded by a guard ring which extends the coveraged area to $99 \times 111\ \mathrm{m}^2$.

The discriminated signal (digital read-out) of the RPC saturates when the energies reach a few hundreds of TeV. In order to extend the dynamic range, the read-out of the integrated signal (analog read-out) has been implemented.

The detector is now complete and the final data acquisition system has been recently installed.

A trigger rate of about 4 kHz is sustained during normal operation of the telescope, producing 7 MB/s of throughput from the Front-End Electronics (FEE) to the online farm. The data acquisition system was designed to efficiently handle such event rates, and to combine the tasks of data logging with those of data quality control.

II. Daq System

A. Daq Architecture

The full carpet has been equipped with the final front-end electronics. We trigger on the 130 Clusters belonging to the central region of the carpet (Central Carpet).

The DAQ and Trigger basic elements are structured in modules made up of 12 RPCs, called Clusters. Each Cluster has its own modular read-out and local trigger electronics housed in a Local Station (LS) [2]. The 120 pad signals from each Cluster are stretched to 150 ns in order to guarantee that signals from particles in the same shower can be put in coincidence. The LS outputs a 6-bit Low Multiplicity (LM) weighted bus (when $\geq 1, \geq 2, \geq 3, \geq 4, \geq 5, \geq 6$ pads are fired). This read-out saturates when more than 6 pads fire in coincidence on the same Cluster.

The detection of small size showers is one of the main tasks of the Argo-YBJ experiment. An inclusive trigger able to record a minimum number of hits has been implemented, based on a four-level coincidence scheme which correlates only signals pertaining to adjacent areas. This logic reduces the spurious signal by 60% [3]. This trigger allows to sample photon-induced showers down to energies of a few hundreds GeV. Based on MonteCarlo simulations [4] the expected trigger rate due to the cosmic ray background with hit threshold of 20 counts is about 4 kHz.

Shower events are selected using a simple but powerful algorithm, by just summing the multiplicities of all Clusters across the entire carpet in a time window of $\sim$400 ns. When the total number of hits exceeds a programmed threshold the event is selected for acquisition. The spurious signals from the detector ($\sim$400 Hz/pad) represent the noise for the shower events [5].

Besides this trigger channel there are other complementary triggers designed to select showers with a dense core region with a much bigger cluster data size.

The DAQ set-up must be organized in order to have a good read-out efficiency able to sustain a data transfer rate coming

Manuscript received April 30, 2007; revised August 21, 2007, October 3, 2007, and October 4, 2007.

A. Aloisio and S. Catalanotti are with Dipartimento di Fisica, Universitá "Federico II" and INFN, Napoli 80126, Italy.

P. Branchini, A. Budano, P. Celio, F. Galeazzi, R. Gargana, S. Mastroianni, F. Ruggieri, and C. Stanescu are with Dipartimento di Fisica, Universitá Roma Tre and INFN, Roma 00146, Italy (e-mail: antonio.budano@roma3.infn.it).

P. Creti and G. Marsella are with Dipartimento di Ingegneria dell'Innovazione, Universitá del Salento and INFN, Lecce 73100, Italy.

Digital Object Identifier 10.1109/TNS.2007.909988

from the low density showers and high peak data flows from the high density showers. So the DAQ system must be able to acquire this extremely large dynamic range from a few hundreds of bytes up to few Mbyte for each trigger asserted.

The Argo-YBJ DAQ is built on a two-layer read-out architecture implementing an event-driven data collection by using two custom bus protocols, based on VME-bus.

When a trigger occurs, each LS assembles a local data frame containing both analog and digital read-out information, an incremental event number, the addresses of the fired strips and all the timing information stored in the time to digital converters.

The local data frame is electronically transferred to the Central Station at a rate of 160 Mbit/s (16-bit word in 100 ns) and pushed into a FIFO memory placed in the Argo Memory Board (AMB) [6] entering the Level-1 read-out system. The Level-1 environment is based on crates equipped with a VME and a custom bus (L1bus) [7] that uses the lines undefined by the VME standard [8]. Each Level-1 crate contains up to 10 AMB boards managing the read-out from up to 40 LS channels. A Level-1 read-out controller in each VME crate collects the front-end data via the L1bus. It implements hardware block transfer capability and its peak throughput is 50 MB/s [9]. The typical sustained throughput on the Level-1 crate is 2 MB/s. Up to 8 Level-1 controllers can be daisy-chained and acquired by one Level-2 controller through a fast one-directional custom-bus (CBUS) sustaining up to 40 MB/s [9], [10]. The typical sustained throughput on the CBUS is 7 MB/s.

The Level-1 controller builds data frames consisting of an event number, data frames from the AMB boards and a parity word: in a similar fashion, the Level-2 controller collects the data frames relevant to a given event number from all the Level-1 boards. A decoupling fifo is hosted on both the Level-1 and the Level-2 controllers to buffer the data. The Level-2 controller fifo is read by a CPU hosted on the same VME bus and the data are sent to the online farm system through gigabit ethernet connection.

The VME CPU currently in use is a Motorola model MVME6100 [11] running at 1.3 GHz, with 1 GB of RAM memory and two ethernet 100/1000 interfaces.

All the DAQ VME crates provide a slow commercial bidirectional bus (VIC-bus [12]) whose bandwidth is 4 MB/s, that is used to initialize the DAQ chains and to check the run conditions.

To improve scalability, the system can be split in several chains, each one having its own VME processor board [13].

B. Farm Online System

The online farm consists of a Blade Center by IBM [14]. The blade center chassis can host up to 14 blade server boards allowing fast and easy maintenance and redundancy. Each board hosts two Intel Xeon CPUs running at 3.06 GHz with Hyper-Threading and 512 kB cache each. A total of 3 GB RAM is installed on each board. The bandwidth to the internal SCSI disk is about 28 MB/s.

The blade center chassis is equipped with two switches to which all blade servers are internally connected: one for the ethernet protocol (10/100/1000 Mbs) and one for a fiber channel protocol.

The present configuration provides full redundancy since six boards are installed out of which only three are used:

- one for the farm system that receives the data from the MVME6100 CPU with a 1000 Mbs ethernet connection;
- one for the control of the DAQ system: a software named “Argo Run Control” written in Java language offers a GUI (Graphical User Interface) allowing a user to manage the DAQ system for configuring/controlling the apparatus and the data taking;
- one for archiving and trasfering data files after storing their information (like checksum, size) in a Data Base.

The blade center has also fiber optics connections to a disk server and to a tape library.

The disk server is an IBM model DS4100 [15] equipped with 14 disks of 450 GB each. The total space is about 3.2 TB split into 3 Raid5 arrays each made up by 3 plus 1 (parity) disks, and 2 hot spare disks. This configuration allows for a threefold disk failure before data loss, provided all rebuilds can be completed.

All the blade servers share this disk space using the General Parallel File System [16] (GPFS), the high-performance shared-disk filesystem from IBM. GPFS allows parallel applications the simultaneous access to a set of files (or even a single file) from any node that has the GPFS file system mounted while providing a high level of control over all file system operations. In the current configuration of the farm system we measured a disk write throughput of about 45 MB/s, when at the same time another machine was reading data files at about 35 MB/s.

C. DAQ System Performance

This section deals with the DAQ system performance discussing, in particular, data acquired by 130 clusters at the YBJ experimental site (the central carpet zone).

In this set-up the data transfer rate and dead time were measured by changing the trigger threshold, the DAQ performance being interesting mainly in the case of low trigger thresholds. Fig. 1 shows the trigger rate and data transfer dependences on the trigger threshold. The measurements have been performed by decreasing the trigger threshold down to 10 pads on the whole carpet. The measurement at a threshold value of 10 pads was done to check the system performance under stressed conditions. During the data taking we imposed a threshold value of 20 pads which allows us to study Gamma-Ray astronomy events while maintaining a good rejection of noise triggered events.

As shown in Fig. 1, the trigger rate and the data transfer display similar dependence on the trigger threshold. When the threshold value is 20, the trigger rate is as high as 4 kHz and the data throughput is about 7 MB/s. To fully characterize the DAQ performance it is important to investigate the dead time of the system under these conditions. The different sources of the dead time introduced by the data acquisition chain are:

- Dead Time on Transfer (DTT) is asserted by each Local Station during the data transfer from FEE to the AMB;
- Level-1 Dead Time (L1DT) is generated inside each Level-1 controller which puts in a logical-OR the almost-full FIFO flags present on each board equipped in the crate;
- Level-2 Dead Time (L2DT) is active when the Level-2 controller FIFO is almost full.

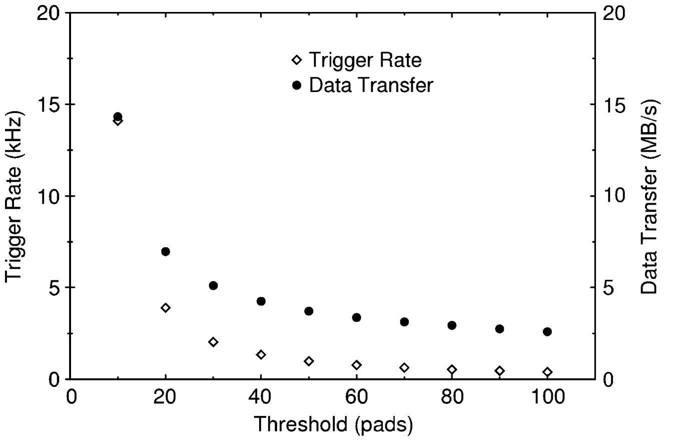

Fig. 1. Total transfer rate and trigger rate of the DAQ system versus the trigger threshold.

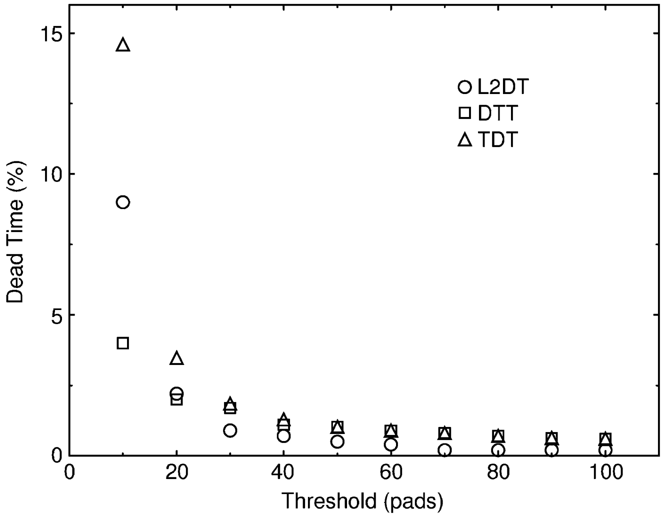

Fig. 2. The L2DT (Level-2 Dead Time) circle, DTT (Dead Time on Transfer) square and the TDT (Total Dead Time) triangle, versus the trigger threshold.

The Total Dead Time (TDT) is the logical-OR of all dead time components. While the DTT is not reducible because it depends on the local data frame from the front-end, the L1DT and L2DT are dominated by the CPU read-out speed on the VME bus and by the software running on the CPU. In order to keep the TDT as low as possible it is essential to reduce the L1DT and L2DT, and this in turn means an effective decoupling between the 3-FIFO levels (AMB, Level-1 controller, Level-2 controller).

Fig. 2 shows the TDT, DTT and L2DT versus the trigger threshold. All the dead time contributions are of the order of a few percent down to ~20 threshold pads, the dominant component being the DTT. When the threshold is lower than ~20 it is evident from Fig. 2 that the dominant contribution is the L2DT; in this case splitting across multiple Level-2 chains can be a good solution to decrease the dead time. These measurements thus show that the DAQ read-out doesn't introduce a further dead time in the threshold region of interest.

A software component running in the farm online system reformats the data coming from the Level-2 CPU before writing them to disk. This component allows to reduce the data size by a factor of 3 by removing redundant information. Redundancy is used before data reformatting to check data consistency and provides eventual error recovery. After data reformatting the Argo-YBJ experiment, in its present configuration, produces a data rate about 2–3 MB/s.

III. Data Mover

The computing resources available at the site allow only for some limited data processing and data storage. Hence, the data collected by the experiment need to be moved and analyzed elsewhere. The collaboration relies on two computing centers, one in IHEP-Beijng in China, and one in CNAF-Bologna in Italy: according to Fig. 1 and accounting for data reformatting, the minimum data bandwidth from the laboratory to the computing centers should be of the order of 50 Mbs.

Network connectivity between China and Europe is currently 155 Mbs. The fraction of such bandwidth which could be used by Argo is insufficient to support the experiment's data rate: anyway during the next months the connectivity should be increased to gigabit bandwith using the TEIN2 and ORIENT network [17].

The experimental site used to have a very limited network connection (8 Mbs) to the main Chinese network infrastructures. For this reason, the data collected by the experiment were written to tape and sent to the main computing centers, where the tapes were read and the data copied to disk. This method of operation provides an enormous instantaneous throughput (during transfers), but has the consequence that the data is available to the collaboration after some delay, of the order of several weeks. This not only affects physics analysis but also has an impact on the control of the experiment, since some subtle effects may be revealed only after a full data analysis.

Since one year, the upgrade of the network connection to 155 Mbs from the laboratory in YBJ to Beijing makes it possible to transfer the data via network. We started to develop a Data Mover Application to transfer data from the Argo-YBJ laboratory to the collaboration computing centers.

The collaboration is evolving most computing activities to Grid, since this approach provides benefits in terms of efficient resource (CPU, storage) usage, enforcement of common policies, redundancy, etc. In this framework, it has been a natural consequence to develop the "Data Mover" software using Grid services [18].

A. Requirements

The application meets the following requirements:

- automatic: need of manual intervention by an expert is minimal;
- does not interfere with data taking: this means it is able to copy the data quickly, and is designed in such a way as to avoid filling up the buffer disk, which would stop the data acquisition;
- data-safe: at every step of the procedure at least two copies of the data exist.
- based on redundant services: all Grid services which are used are duplicated

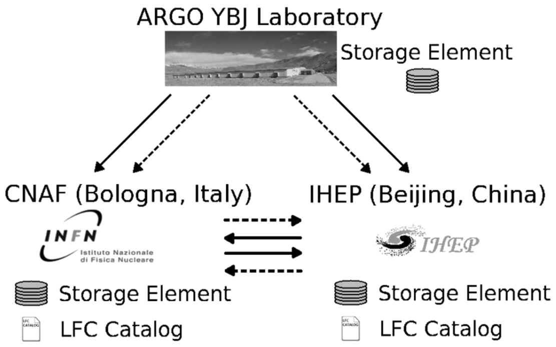

Fig. 3. Schematic diagram of the data mover application. The arrows symbolize the FTS channel, the outlined arrows are the channels of the IHEP FTS server, the other are managed by CNAF FTS server.

B. Overview

The architecture of the "Data Mover" application is shown in Fig. 3. The "Data Mover" application is based on four Grid services: the Storage Element, the File Transfer Service, the Logical File Catalog and the User Interface.

The Storage Element (SE) is the component where the data are stored in the Grid architecture.

The File Transfer Service (FTS) is the component that permits to move the data from an SE to an other. The FTS service works with "channels" that connect the SEs. A channel is a named uni-directional logical connection from one SE to an other: it is configurable in terms of bandwidth, number of streams, access policies, and more. Transfer of one file or of a group of files is called a "job". FTS jobs are processed asynchronously: upon submission a job identifier is returned, which can be used at any time to query the status of the transfer.

The Logical File Catalog (LFC) permits to the user in the Grid to assign a logical name to a physical file present on a SE. The association is one-to-many, a logical name can point to several physical copies of the same file: the copies can be on different SEs and are called "replicas".

The User Interface is the gateway to the Grid, where users are authenticated and authorized to use Grid services [18].

C. Operations

As shown in the previous section, the data are collected from the DAQ system and sent to the farm machine which is one of the blades of the server.

The data is routinely migrated to the SE by a procedure scheduled to be executed periodically. Since the SE and the farm machine share the same disk the migration involves no data copying or moving, rather only the metadata information stored in the SE database is updated. As soon as a run has been successfully migrated, a flag is set in the DAQ database.

At this point the "Data Mover" (DM) application starts. The architecture of this application can be split in three sub-programs:

1) transfer of the data from YBJ to one of the computing centers;
2) synchronization between the computing centers;
3) garbage collection at YBJ.

The data transfer procedure selects runs for which the migrated flag is set, prepares the list of relevant files, picks the FTS server and channel to be used based on their availability, and submits the files for transfer. As far as the FTS server and channel choice are concerned, the first available FTS server is contacted, and for such server the first working channel from YBJ to one of the computing centers is used: in case of problems the other FTS server and/or channel are tried.

After a run has been queued for transfer, the FTS server used and the identifier of the transfer are stored in the DM database. The application periodically checks the state for every transfer and when all data files of a run have been copied they are registered in the LFC catalog. The DM database is updated accordingly.

The synchronization process runs asynchronously at each site. It queries the LFC catalog of the other site and tries to find entries which are not in the local catalog. Missing entries are selected for transfer using the first available FTS server/channel pair, as above. When the transfer is complete the LFC local catalog is updated: furthermore, the copy of the file at the other site is registered locally as a replica and the local copy is registered as a replica at the other site.

The garbage collector is responsible for cleaning up the buffer disk at YBJ. At first all files that have been recently transfered are checked: files for which two distinct physical copies are found in the LFC catalogues are removed from the buffer disk. If, at this stage, there is still need for disk space, the garbage collector starts to copy the files to tape. Upon successful copy to tape, the files are deleted. Such tapes will then be sent to Italy where they will be read using a procedure which will take care of storing the files in the SE and registering them in the LFC catalogue.

D. Test and Performance of Data Mover

At this time the "Data Mover" application is under test. For this test a SE in Roma Tre is used to simulate the YBJ site.

The test is done starting from several runs that have some files about 100 MB and stored in the Roma Tre SE. Using the first avaiable FTS server/channel (referring to the scheme in Fig. 3, for this test CNAF site is tried first) these files are sent to the SE and registered in the LFC Catalogue at CNAF. At this point the application starts to check the differences between two LFC catalogue (IHEP and CNAF) and begins to transfer the missing data using the appropriate FTS channel.

The test was done with success. The application showed a very high level of reliability, effectively coping with network glitches or temporary server unavailability, and always granted that at least two copies of the data files existed at any time.

The Table I shows the measured transfer rates. The table shows that the transfer rate is very low between Italy and China: this is a known problem due to the present network configuration, which will be overcome in a few months when the routing via TEIN2 and ORIENT networks will be in operation. The

TABLE I
FTS CHANNEL AND THE RELEVANT TRANSFER RATE

FTS Channel	**Data Transfer (MB/s)**
ROMA TRE - CNAF	~ 15
CNAF - IHEP	~ 1
IHEP - CNAF	~ 1

next step is to do a test with larger files to check application bugs.

IV. CONCLUSION

In this work an efficient setup for the DAQ system is described. The measurements of the DAQ performance demonstrate that the system is capable to sustain a good trigger rate of 4 kHz with a low dead time (below 4%). A powerful data transfer system was built, exploiting Grid services. The test for the Data Mover application shows that the data can be transfered to the collaboration computing centers with high performance (after the upgrade of the international links). A small fraction of failed transfer, mostly due to extended periods of unavailability of GRID services during site maintenance, required operator intervention.

In a few months the final version of the Data Mover application will be installed at the YBJ site, and this will greatly improve the timely distribution and accessibility to the experiment's data.

ACKNOWLEDGMENT

The authors are particularly indebted to many colleagues who helped them during the design and the setting up of the DAQ system and the Data Mover Application, in particular F. Bitelli of the Department of Physics of University of Roma Tre and S. Cellini of INFN-Roma Tre.

REFERENCES

[1] M. Abbrescia *et al.*, Astroparticle Physics With Argo Proposal 1996.

[2] R. Assiro *et al.*, “Local station: The data read-out basic unit for the Argo-YBJ experiment,” *Nucl. Instr. Meth. A*, vol. 518, pp. 549–553, 2004.

[3] S. Mastroianni, “Studio del Sistema di Trigger per l’Esperimento ARGO-YBJ” Ph.D. dissertation, Dipt. Fisica, Università "Federico II", Napoli, Naples [Online]. Available: http://www.argo.na.infn.it/argo_pub_theses.html

[4] S. Mastroianni, “The ARGO-YBJ inclusive trigger,” in *Proc. 29th ICRC*, 2005, vol. 5, pp. 311–314.

[5] A. Aloisio *et al.*, “The trigger system of the Argo-YBJ experiment,” *IEEE Trans. Nucl. Sci.*, vol. 51, no. 4, pp. 1835–1839, Aug. 2004.

[6] A. Aloisio *et al.*, “FPGAs widen the ARGO-YBJ experiment’s eyes,” *IEEE Trans. Nucl. Sci.*, vol. 49, no. 1, pp. 401–404, Feb. 2002.

[7] A. Aloisio *et al.*, “Level-1 DAQ for the KLOE experiment,” in *Proc. Int. Conf. Computing High Energy Physics*, Singapore, 1995, p. 371.

[8] *The VME Bus Specification*, ANSI/IEEE 1014-1987, IEC 821.

[9] A. Aloisio *et al.*, “Real-time diagnostic and performance monitoring in a DAQ environment,” *IEEE Trans. Nucl. Sci.*, vol. 47, no. 2, pp. 162–165, Apr. 2000.

[10] A. Aloisio *et al.*, “Custom busses for large scale data acquisition systems,” in *Proc. IEEE Int. Conf. Electronics, Circuits Systems (ICECS ’96),*, Rodos, Greece, 1996, pp. 1155–1161.

[11] Information About Motorola VME Processor, Model MVME6100 [Online]. Available: http://www.motorola.com/content.jsp?globalObjectId=5656

[12] C. F. Parkman, “VICbus: VME inter-crate bus-a versatile cable bus,” *IEEE Trans. Nucl. Sci.*, vol. 39, no. 2, pp. 77–84, Apr. 1992.

[13] A. Aloisio *et al.*, “Argo-YBJ data acquisition system,” *Nucl. Instr. Meth. A*, vol. 568, pp. 847–853, 2006.

[14] Information About IBM Blade Center [Online]. Available: http://www.ibm.com/systems/bladecenter/

[15] Information About IBM DS4100 [Online]. Available: http://www.ibm.com/storage/disk/ds4000/ds4100/index.html

[16] Information About General Parallel File System (GPFS) [Online]. Available: http://www.ibm.com/systems/clusters/software/gpfs.html

[17] TEIN2 and ORIENT are DANTE Projects [Online]. Available: http://www.dante.net/

[18] Information About GLite Grid Services [Online]. Available: http://www.glite.web.cern.ch/glite/documentation/ [Online]. Available: http://www.grid-it.cnaf.infn.it/

[19] C. Bacci *et al.*, The Argo-YBJ Project, Addendum to the Proposal, 1998 [Online]. Available: http://www.argo.na.infn.it/