The TIER-2 site for the ARGO-YBJ experiment

F. BITELLI[1,2], A. BUDANO[2], S.M. MARI[1,2]
[1]*Dipartimento di Fisica Università Roma 3*
[2]*INFN Sezione Roma 3*
bitelli@fis.uniroma3.it, antonio.budano@roma3.infn.it
DOI: 10.7529/ICRC2011/V03/0221

Abstract: The ARGO-YBJ experiment is taking data at the Yangbajing International Cosmic Ray Observatory located at an altitude of about 4300 m a.s.l. (Tibet, P.R. China). The detector consists of an EAS array made of a full coverage RPCs carpet. The experiment is taking data since 2007 in the final configuration, the trigger rate is about 3.5 kHz producing a throughput of data of about 2.5 MB/s. Taking into account the high duty cycle of the experiment, the amount of raw data collected is about 200 TB/year and the amount of reconstructed data is about 60 TB/year. The management of this volume of data and the production of a huge amount of Monte Carlo data samples require a suitable computing power. A TIER-2 site is arranged to face out these tasks. In this paper the structure of the data transfer method and data reconstruction/analysis based on the TIERs are presented. The hardware infrastructure and software architecture of the Roma 3 TIER-2 site of the ARGO-YBJ experiment are also discussed.

Keywords: data analysis, computing, EAS.

1 Introduction

The ARGO-YBJ experiment is taking data since 2007, it has been built by a Chinese and Italian collaboration. The experiment site is located at the Yangbajing International Cosmic Ray Observatory at Yangbajing (Tibet, P.R. China) at 4300 meters above the sea level, about 90 km from Lhasa.

The detector is made of a single layer of Resistive Plate Chambers (RPC), with a central region (about 5800 m^2) with 93% active area and an almost equivalent region sampled. The detector is read-out through 18360 TDC channels (strips) which provide a detailed spatial-temporal image of the shower front, with a time resolution of about 1.8 ns. The detector was designed to operate whit high duty cycle (> 86%) [1].

The ARGO-YBJ ground-based detector allows for the investigation of many aspects in the gamma-astronomy field and in the cosmic ray physics. The apparatus can detect showers in a large energy range, from few hundreds GeV up to the PeV region, because of the digital and analog read-out of the RPC chambers [2].

An *event* (see figure 1) collected with the ARGO-YBJ detector is a collection of hits (*i.e.* the digitized information generated by charged particle belonging to the shower front and hitting the detector) recorded within the trigger time window of 420 ns. The detector is presently taking data with a trigger configuration generating a rate of about 3.5 kHz events written to disk, corresponding to a data throughput of 2.5 MB/s.

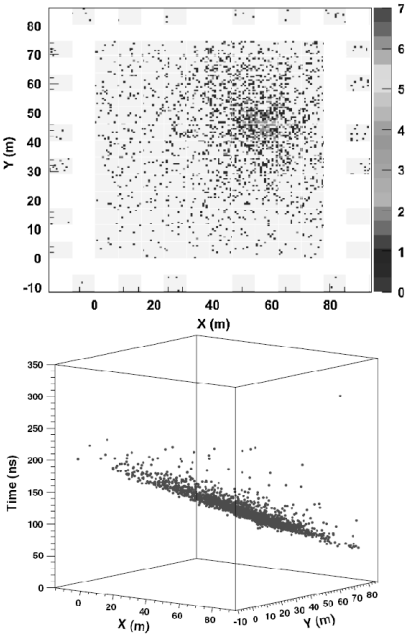

Figure 1. Two different views of a detected shower. The hit map at ground (top), the color code representing the strip multiplicity of each fired pad. A space-time view of the detected shower (bottom).

The DAQ electronics is designed to cope with a throughput three times larger which, due to the continuous mode of operation of the detector and the high uptime, extrapolates to about 17 TB/month or 200 TB/year [3]. The events collected need to be reconstructed (i.e. determining the shower direction, the shower size) before they can be used for physics analysis. The event reconstruction is a very CPU-intensive task, requiring about 500 SI2006 to keep up with data taking. Moreover, Monte Carlo simulations have similar needs as the reconstruction, in terms of CPU power and storage space. Such a computing power as well as storage space are not available at the experimental site, so the data are moved and analyzed elsewhere. The Collaboration has been arranged the computing resources needed in two main computing centers (named TIER-1), one at IHEP (Beijing, China) and one at CNAF (Bologna, Italy). The computing power needed to generate Monte Carlo events and to support the end-user analysis has been arranged in smaller sites. As of today ARGO-YBJ has accumulated about 200 TB of raw data which are stored to tape at each of the two computing centers (named TIER-1 site).

2 Data Management

Data taking is organized in RUNs, a period of data taking during which conditions are kept reasonably constant. Each RUN is made of several files (about 1 Gbyte each). The experiment aims at a very high duty cycle, and the expected amount of data collected in one day is of the order of 300 GB. The computing resources available at the experimental site allow only limited data processing and data storage so the data are distributed to the TIER-1 sites and copied on a tape library. In the TIER-1 sites a software is responsible to process the raw data and to reconstruct the events. The data reconstructed are finally copied to the TIER-2 sites for the analysis activity.

3 The ARGO-YBJ TIER-2 Site

The TIER-2 Computing Center hosted in the INFN Roma Tre was founded in 2008 thanks to a project that involved the ARGO-YBJ experiment and the sharing of resources between INFN and Department of Physics. After the first year of use of resources by the founding members of the experiments, the interest from other groups within our structure has been growing and has therefore decided to extend the sharing of resources to all groups belonging to INFN and the University that require more computing power.
The Data Center consists of five racks (42 U) cooled by an air conditioning system with direct expansion of gas-based air flow inside the rack. This architecture is therefore "heat-isolated" from the room.
Inside the room is also a UPS that can prevent the sudden shutdown of equipment in the event of a blackout. All the machines are connected within the rack through a Gigabit Ethernet network. The resources within the structure can be divided in 3 types: CPUs dedicated to the calculation, CPUs dedicated to services and equipment storage disk.
Concerning the computing power, the TIER-2 site is equipped with about 40 logical CPU, in particular: 5 Blade servers SuperMicro (2 Intel Quad Core E5520 2.27 GHz, 48 GB RAM equipped with Infiniband [4]): 80 slots (the year of purchase 2010); 3 Blade SuperMicro (2 Intel Six Core E5660 2.8 GHz 48 GB RAM equipped with Infiniband); 16 HP Blade servers (2 Intel Quad Core 2 GHz, 16 GB RAM): 128 slots and 3 Graphical Processor Unit (GPU) [5] system NVIDIA TESLA 2070.
All the services needed to manage the computing site are supported by: one Computing Element (CE), two User Interface (UI) units, four machines for storage management, one Grid Storage Element (StoRM) and 2 machines to monitor all the TIER-2 system [6,7].
The hardware described provides a CPU power of about 500 SI2006 needed for the Monte Carlo events production and to support the end-user analysis.

3.1 The Computing Element (CE)

The CE is a queue manager with batch queuing system based on PBS/MAUI [8].
Different queues are arranged in the cluster and each user is allowed to submit to one or some of them according to permission policy decided by the administrator .
The CE is responsible to process jobs requirement and assign the job to the best resource available in the cluster.
A queue is configured in the cluster is for the multiprocessor jobs (such as Message Passing Interface (MPI) job [9]). These jobs uses the network for interchange message and in case of MPI jobs the bandwidth between the processor is a important factor. Thus in this case CE routes this MPI jobs to resources interconnected with Infiniband that is a type of communication link for data flow between processors that offers throughput of up to 5.0 GB/s. The bandwidth performance has been measured. The values obtained are the following:

(size)	(transfer time)	(bandwidth)
32 bytes	0.000001 s	35.135531 MB/s
2048 bytes	0.000003 s	588.351684 MB/s
131072 bytes	0.000035 s	3725.642545 MB/s
8388608 bytes	0.002566 s	3269.482466 MB/s

These data show that the bandwidth is greater than 3 GB/s as expected. In a similar way the CE may address jobs requiring the GPU, that is a new architecture that permits to have a lot of processor in a single cards, to the correct subsystem environment. The resources are shared between different experiments so the CE is configured to assign different priorities to user jobs and furthermore the scheduler (MAUI) is configured to use Fairshare. The Fairshare allows to incorporate the historical resource utilization information into job feasibility and priority decisions. This feature allows site administrators to set system utilization targets for users, groups, accounts,

classes, and QOS levels. Administrators can also specify the time frame over which resource utilization is evaluated in determining whether or not the goal is being reached. The scheduler configuration parameters allow to specify the utilization metric, how historical information is aggregated, and the effect of Fairshare state on scheduling behavior. Fairshare targets can be specified for any credentials (i.e., user, group, class, etc) which administrators wish to have affected by this information.
In the CE is also installed the Grid Middleware (Glite3.2) [6] and so the CE is also able to receive jobs from Grid users. The Grid CE type installed is the new CREAM CE that in addition to the default Glite services offers a web service interface to manage jobs requests. The jobs processed during one year (2010) of operation by the TIER-2 site are shown in figure 2.

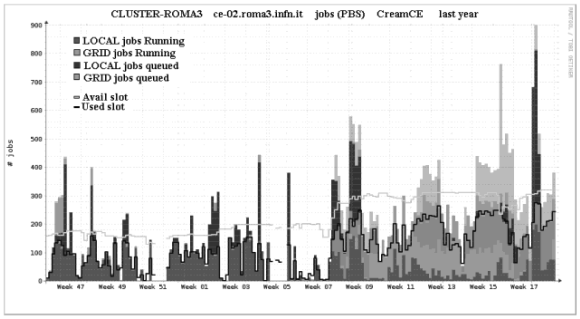

Figure 2. Jobs processed by the Computing Elements during last year

3.2 The User Interface (UI)

This machine is a high performance CPU machine that may also allow users to compile and execute simple test jobs before submitting them to the cluster.
The Roma Tre TIER-2 site has two User Interface (UI) elements that constitute the interface between the computing resources of the system and the users. From this machine the users can submit their jobs to local cluster and to the Grid. In the user interfaces are also installed the typical software packages that allow users to analyze and create custom layout for their results (eg: ROOT suite [10], PAW [11], GNUPLOT [12],...).
One of the user Interface hosts an LDAP server [13] that store information about users and provides these information to each machine of the cluster. A second LDAP replica machine is also present in the cluster.

3.3 The Storage Architecture

All the users data and experimental data is stored in one of the three storage system present in the cluster for a total amount of about 170 TB.
These disks are hosted in three different systems, in particular, a SUN 6140 System that has 48 HDD 500 GB SATA with a total amount of data about 20 TB; an E4 E6500 system with a set of fast SAS disks for the user home and software area and a set of SATA disks with a total of storage space of about 100 TB and finally an HP MSA2000 system with 24 HDD 2 TB SATA disks.
All these systems are interconnected with 4 machines (named Data Server (DS)) using 4 Gbps Fiber Channel connections [14]. In order to share the data between all the cluster, in each DS is installed GPFS (General Parallel File System server provided from IBM [15]. GPFS allows parallel applications to access simultaneously 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. The storage area for the ARGO-YBJ experiment is organized in one dedicated filesystem, a second one file system is also present for the user directories and one more for the experimental software area. The data distribution system is configured to grant load balancing between the four DS machine and furthermore each of these machine has four Ethernet Gigabit Interface aggregated in order to have a bandwidth of about 4 Gbps to the cluster network. All the network switches present in the cluster are also interconnected each other using LACP Link Aggregation Control Protocol [16]. Part of the storage system is also served to grid users trough the Grid storage Element StoRM (Storage Resource Manager) [7] that is a lightweight solution that provide an SRM Interface to simply manage disk operations.

3.4 The ARGO-YBJ software

The software area of the ARGO-YBJ experiment, as described above, is shared in the computing cluster using GPFS and it contains all the software for reconstructing and analyzing data and for the Monte Carlo production.
This area is also shared between local and Grid users, this permits to all jobs submitted in the TIER-2 site to access to the same software release. Furthermore the installation of the software in the site can be done using Grid jobs (using a special user named Software Grid Manager (SGM)), so that few people in the collaboration, with the role of SGM, can control in a easy way the release installed in each TIER sites. In the ARGO-YBJ collaboration the official software release are installed in the TIER-1 site and then distributed to the other TIER sites.

3.5 The Monitoring System

All the machines and services of the cluster are monitored in real time and statistics of used resource are collected and published in web pages. A Ganglia client [17] is installed in each node and it is responsible to send to a server relevant information about resources (such as CPU consuming, memory, network and disk usage). Furthermore we have installed a Nagios server [18] that monitors hosts and services and in case of some problems it is able to send notification via email and SMS to the admin staff.

4 Conclusion

The ARGO-YBJ experiment is taking data at the International Cosmic Rays Laboratory, located in Tibet (P.R. China). The trigger rate is about 3.5 kHz corresponding to a throughput of about 2.5 MB/s. This trigger rate produces a high data flow of about 200 TB/year. The analysis of this amount of data is based on a massive Monte Carlo events production which requires storage capability of about 200 TB/year and a computing power of about 500 SI2006. To face out this task a computing center based on the TIER-2 architecture has been successfully realized at the Roma Tre University. The performances of this computing site fulfill the requirements of the ARGO-YBJ experiment and support the end-user analysis also.

5 References

[1] Aielli G. et al. (ARGO-YBJ Collaboration), Nucl. Phys. B Proc. Suppl., 2007, **166** , 96.
[2] M. Abbrescia *et al.*, Astroparticle Physics With Argo Proposal 1996.
[3] Aloisio A. et al., IEEE Transaction Nuclear Science, 2008, **55** (1), 241-245.
[4] More information about Infiniband can be found at: http://www.infinibandta.org.
[5] More information about GPU can be found http://www.nvidia.com/object/GPU_Computing.html
[6] More information about Glite Middleware can be found at: http://glite.cern.ch/
[7] More information about Storm can be found at: http://storm.forge.cnaf.infn.it/
[8] MAUI and PBS are at: http://www.clusterresources.com/
[9] More information about MPI can be found at: http://www.mcs.anl.gov/research/projects/mpi/
[10] http://root.cern.ch/
[11] http://wwwasd.web.cern.ch/wwwasd/paw/
[12] http://www.gnuplot.info/
[13] http://www.openldap.org/
[14] More information about Fiber Channel can be found at: http://www.fibrechannel.org/
[15] More information about GPFS can be found at: http://www-03.ibm.com/systems/software/gpfs/
[16] More information about LACP can be found at: http://standards.ieee.org/index.html
[17] More information about Ganglia can be found at: http://ganglia.sourceforge.net.
[18] More information about Nagios can be found at: http://www.nagios.org.