ARGO-YBJ Computing Model. Data Analysis and Hardware/Software Architecture of the Processing Farm

Paola Celio,[1] Severino Bussino,[1] Alfredo La Rosa, [1] Stefano Mari,[1] Daniele Martello,[2] Cristian Stanescu, [1] and Antonio Surdo[2] for the ARGO-YBJ Collaboration
(1) INFN and Dipartimento di Fisica dell'Universita' di Roma Tre, Rome, Italy
(2) INFN and Dipartimento di Fisica dell'Universita' di Lecce, Lecce, Italy

Abstract

The Chinese-Italian experiment ARGO-YBJ in Tibet is going to start next spring the first period of data-taking with 36 RPC clusters installed. The paper describes the computing model and the hardware resources required for data reconstruction and analysis. The present configuration of the processing farm, composed by 24 biprocessors computing elements and 4 diskserves (7.5 TB total) is describeds. The software designed and developed for data organization, reconstruction job submission, data analysis and error management is briefly reported. The management software of the farm, including the GRIDWARE queuing and submission package, the DB design and the monitoring software are also described.

1. Introduction

ARGO-YBJ is a cosmic-ray telescope placed at 4300 m of altitude in Tibet based on single layer of RPC covering 74x78 m^2 (1608 chambers full coverage carpet plus a ring of 240 chambers). The experiment will start the data-taking this year with a partial configuration, ie with 36 clusters (1 cluster = 12 RPC). In the meantime the construction of the apparatus will continue with another partial data-taking before using the complete configuration. The high trigger rates will produce important raw data flows and we foresee up to 200 TB of data/year with the full apparatus [1]. The processing of the data will be done by a dedicated farm, thoroughly designed.

2. Design Hardware Configuration

In our computing model [2] we suppose to take data for 50-70% of the year time during the first period, and for 80-90% later. We foresee to keep the reconstructed data on disks for long periods of time (one year). The first period of data taking introduces however some "special" request: the reconstructed data will contain not only the reconstructed particle data but also the raw information, the events will undergo to many re-constructions, different versions of the

Table 1. Table of Data Taking and Processing Numbers.

Clusters	Tr rate (KHz)	Events/y	Raw TB/y	Reco TB/y	SPECint2000
36	4.8	$1.02\ 10^{11}$	24	5	30000
54	7.1	$1.76\ 10^{11}$	41	10	46600
154	25.	$6.24\ 10^{11}$	200	36	166000

programs will be used, etc.

The data flows and numbers are reported in Table 1, for the three phases of data-taking. The numbers are based on average raw event size (250 byte), reconstructed event size (50 byte) and processing time per event (max 20 ms using a CPU PIII 1GHz clock, equivalent to 420 SPECint2000).

The farm for data processing and analysis is organized in a traditional way, with a number of "computing elements" accessing data via a local network (a mix of GE and FE switches) from few disk servers. We are using two kinds of RAID disk servers: one SCSI (3TB of disks, more efficient) and one SerialATA (2.5TB). Both work at RAID 5 level. There are other two small disk servers (SerialATA RAID 0, 1TB of disks each) for raw data manipulation and for users data. The current farm layout, dimensioned for the first period of data-taking, is described in Figure 1.

The system is protected against job and data loss by an UPS and a dedicated card for anomalies signaling. Two tape systems are installed: a DLT loader (initial data-taking and backup) and a more powerfull LTO robot (20 slots) for the other phases. In one year of full data-taking we need 1000-750 second generation LTO tapes/year (n).

The architecture and the components that will be added to the farm to cope with the requests of the next periods of data-taking will follow the "Moore law". The WAN connectivity was planned to be raised to 16 and later to 32 MBPS bandwith in the next three years.

3. Software Environment and Experimental Data Processing

All the farm components are running now 7.3.1 RedHat Linux installed via a kickstart mechanism. To unify the computing resources we are using the free version of Sun One GridEngine software, a complex job queuing and submission system, distributed by SUN (initially developed by Gridware Inc.) [3].

The Sun GridEngine has a distributed architecture, with one or more masters that manages all the declared resources, the requests, the priorities and the permissions and decides how and where to schedule the jobs. It checks the queues and the jobs via tables containing the status of all the components. The queues

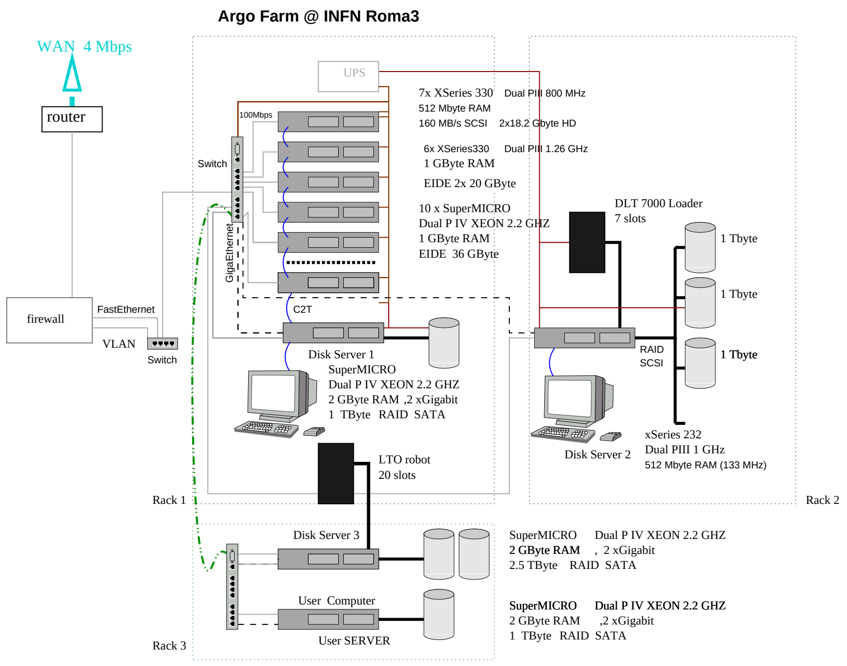

Fig. 1. Argo Farm configuration.

are managed locally by an “execution daemons”. Submission and administration functions can be done via separate hosts, or can be included in the master.

We choose this software for its reliability, the availability of a checkpoint mechanism and the possibility to have “user defined parameters” (called complexes) for the queues. In this way the scheduler can submit jobs only to those queues belonging to a pre-defined category (for example: MC, production, analysis, etc.).

The experimental data are processed by the farm software, in three distinct phases. In the first phase a certain number of Perl scripts verify automatically the correctness of the raw data and the correlated RUN information (RUN log, geometry, calibration and slow-control). After, the information about the run context, the raw data and their location on the disk server are permanentely saved into the production DB. The second phase, also guided by a series of scripts and API's, starts querying the DB for the RUNs to be processed, checks the farm resources and submit jobs to the GRIDEngine. The processing steps of each job are traced into a dedicated work DB, used to handle possible errors. The third phase, the physics analysis, is done again by API and scripts, querying the DB and submitting jobs to the farm.

The sensitive point in this process is the DB organization, performance and robustness. We choose POSTRESQL DB and the design of the tables and their relations reflects the logical data flow processing (see Figure 2). The tables containing the RUN information (RUN number, type, organization, related in-

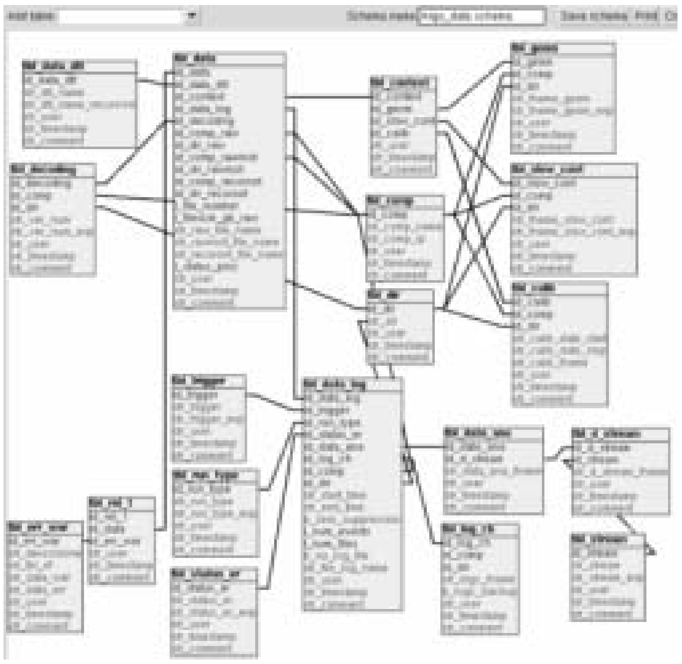

Fig. 2. DB production schemaArgo.

formation, missing files, etc) are filled during the first phase. The reconstructed data info (files, their location, version of the program, errors, etc) are stored in the production DB at the end of the second phase. The "standard" analysis performed on the data are also registered in the DB. Many other tools, to refine the data analysis, as well as to monitor and to face emergencies during the data processing, are under development.

The complex software developed for the farm installation (kickstart, NIS, web server, Gridware sw, nfs, etc) and data processing were tested during intensive MC data production.

4. Conclusion

The actual hw/sw configuration proved to be up to now efficient and reliable, and we are confident about the future enhancement, when the computing power of the farm will pass in few years from actual 30000 SPECint2000 to final 170000 SPECint2000 configuration and a disk space of the order of 30 TBytes.

1. A.Surdo et al. (ARGO-YBJ Coll) 2003, in this proceedings
2. S.Bussino et al.,"Computing Model for the ARGO-YBJ", ARGO note 005/01
3. http://gridengine.sunsource.net/project/gridengine/documentation.htm