Data simulation and analysis for ARGO-YBJ experiment using the GRID tools.

P. Celio*, S. Mari* S. Mastroianni†, P. Montini* and C. Stanescu*
on behalf of the ARGO-YBJ collaboration

**Dipartimento di Fisica, Università Roma Tre and INFN Roma Tre, Roma, Italy*
†Dipartimento di Fisica Università Federico II and INFN Napoli, Italy

***Abstract*. The ARGO-YBJ software was integrated into the GRID environment and the full set of data manipulation software instruments (simulation, data transfer and data analysis) are now GRID compliant. In this paper the software environment setup for a massive Monte Carlo data production is presented as well as the complex database structure needed for an efficient data files retrieval and manipulation.**

***Keywords*: Monte Carlo, GRID, database.**

I. Introduction

The Chinese-Italian cosmic-ray telescope ARGO-YBJ in Tibet is taking data with the full RPC carpet installed. The carpet made by a single layer of 1848 RPC chambers is organized in clusters of 12 for data acquisition plus an external ring made by 240 chambers. The detector is optimized for small air showers detection for studies in gamma astronomy, gamma-ray bursts, the Anti-p/p ratio, the primary proton spectrum, solar physics, and other possible fields [1]. The observation of low energy phenomena is possible due to the location of the experimental apparatus at high altitude (4300 m above the sea level) and to the high active surface ($> 92\%$). The trigger is based on pad multiplicity, that corresponds to a data acquisition rate of about 4 KHz of events, producing after data compression around 2.5 Mbyte/s of experimental raw data.

II. The evaluation of Monte Carlo data request.

The ARGO-YBJ experiment is gathering around 1.3×10^{11} events a year at the nominal trigger rate of 4KHz, producing around 100 TB/year of data. The off-line reconstruction produce in its turn 15 TB of data/year. The need for Monte Carlo production was initially estimated in 5-10% of the experimental events, foreseeing the reuse a certain number of times the simulated showers data. The computing power requested by the MonteCarlo production was evaluated at 200 KSPECInt2000. This is explained by the complexity of the shower simulation process and by the quite complex chain of application software implied: the air-shower simulation (CORSIKA [2] and FLUKA [3]) and the simulator of the ARGO-YBJ apparatus response (ARGOG) based on GEANT-3 [4]. The data are afterwards analyzed by Medea++, the ARGO-YBJ reconstruction and analysis software. The ARGO-YBJ collaboration prepared a plan of Monte-Carlo production, subdivided by the different kind of primaries, different ranges of energies and angles and taking into account the needs of the specific groups of physics analysis (gamma astronomy, cosmic ray spectrum, the study of Anti-p/p ratio, moon shadow, etc.). In this first phase of data analysis we produced the simulation of 0.5×10^9 showers induced by 5 different kind of primaries and their interaction with the experimental apparatus. The production was done on local farms as well as on GRID farms, providing an access to the data files based on the inquiry of a dedicated database and on the use of a GRID LFC catalog. The computing power used for this production was estimated in 80 KSPECInt 2000 and the disk space in 70 TB.

III. Production procedures and data base access

The leading idea of this job was to implement a systematic simulation process, developed for the computing farm environment but also for the Grid environment, an approach already developed in the past for experimental data transfer and processing [5]. We prepared the full process of simulation running the Cosmic Air Shower Simulation (CORSIKA), the software that simulate the detector response (ARGOG) and the program of reconstruction (Medea++) in different contexts. The development phase was done for the most part on INFN Roma Tre farm.

The software design and implementation covered different aspects:

- Implementation of a Monte Carlo production database
- Development of jobs submission procedures
- Storage and retrieval of the simulated data files
- Development of a User Interface.

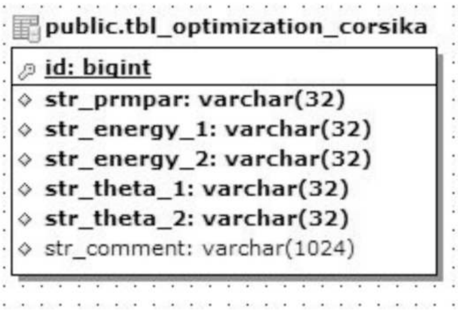
The MonteCarlo production was split by the job submission procedures in smaller parts to take into account the management aspects of the data files and the possible failure situations and recovery. The job submission procedures can operate either directly on the local farm, based on PBS queue system, or via GRID, just modifying a flag that the submitter has to pass like parameter to the procedure.

The submission procedure is done in the following steps:

1) The users are authorized to operate through an authentication process based on a specific table in the database, and through ARGO-YBJ Virtual Organization, in the case of GRID resources usage.
2) The Monte Carlo simulation parameters for each job are specified in a short file
    - NSHOW 9000000
    - PRMPAR 402
    - ESLOPE -1
    - ERANGE 1.E3 3.16E3
    - THETAP 0. 15.
    - PHIP 0. 360.
    - XRDM 62500. 62500.
    - MTRG 20
    - TRIG 4
    - CORSIKA 1
    - INP_COR 1
    - ARGOG 1
    - INP_ARG 1

   In this file are defined the number of showers to be generated, the kind of primary, the energy slope, the energetic range, the theta and phi ranges, the area to be used for the detection, the kind of trigger and the number of time that events will be reused. The version of CORSIKA and ARGOG are also specified. Other sets of parameters used for the simulation are usually fixed and are referenced by a flag in the DB. The values of these parameters are also deposited in a DB table.
3) The procedure reads the previous file and create the scripts for job submission and the input files for CORSIKA and ARGOG, inserting in the corresponding tables their content.
4) Run of the job submission scripts in the local farm or in GRID environment.
5) Each job produces a log file.
6) A script was developed to examine these log files, identify the status of the job and update the database with their status. A dynamical parameter is also used to report waiting to start jobs older than a fixed time (usually 3 days), allowing to clean them.
7) A user-friendly web interface helps to check the current status of the jobs (under development).
8) The Monte Carlo data files correctly produced are introduced in a GRID LFC (Logical File Catalog), doesn't matter if the production was done locally or in GRID.

We have used a Prostgres Database and we have prepared a procedure that is able to perform a DB backup once per week. We are also studying the possibility to replicate the database for a faster interrogation. The scripts were developed in Perl, a language that has been proved to be really suitable for performance in database interrogation and I/O files manipulation. We developed a modular procedure with a deeply detailed log file that enable us to perform also useful analysis on the behavior of the simulation process. The production was carefully followed in terms of performance creating a table (see fig. 1) where to save the space and CPU-time used by each job. The table was used to optimize the available resources and the modularization of the production files.



public.tbl_optimization_corsika

id: bigint
str_prmpar: varchar(32)
str_energy_1: varchar(32)
str_energy_2: varchar(32)
str_theta_1: varchar(32)
str_theta_2: varchar(32)
str_comment: varchar(1024)

Fig. 1. Table for production optimization

IV. Database content

In Fig. 2 is described the database structure, the tables content and their relations. The tables defined for ARGO-YBJ Monte Carlo production are the followings:

- public.tbl_user; list of authorized users for general purpose MC production and analysis.
- public.tbl_corsika_ver; specifies the used CORSIKA version.
- public.tbl_corsika_link_option; the parameters used to compile and link the CORSIKA program (cross sections, atmospheric model, use of Fluka, etc.).
- public.tbl_fluka_ver; used Fluka version.
- public.tbl_corsika_work; table used for specific production tasks, linking other tables for input files (how the production is subdivided in small tasks, the random numbers used) and for output files (file names, LFC catalog name).
- public.tbl_input_main_corsika; complete input card for CORSIKA production, containing variable parameters for specific production.
- public.tbl_input_sec_corsika; more general parameters for CORSIKA (shower altitude, interaction models, etc.) - generally fixed for a certain production.
- public.tbl_optimization_corsika; used to optimize the organization of the production.
- public.tbl_argog_ver; the version used for ARGOG, apparatus simulation program.
- public.tbl_argog_work; table used for specific production tasks, linking other tables for input and output files
- public.tbl_input_main_argog; complete input card for ARGOG, containing variable parameters for specific production.
- public.tbl_input_sec_argog; more general parameters for ARGOG (ARGO-YBJ experiment configuration, etc.) - generally fixed for a certain production.

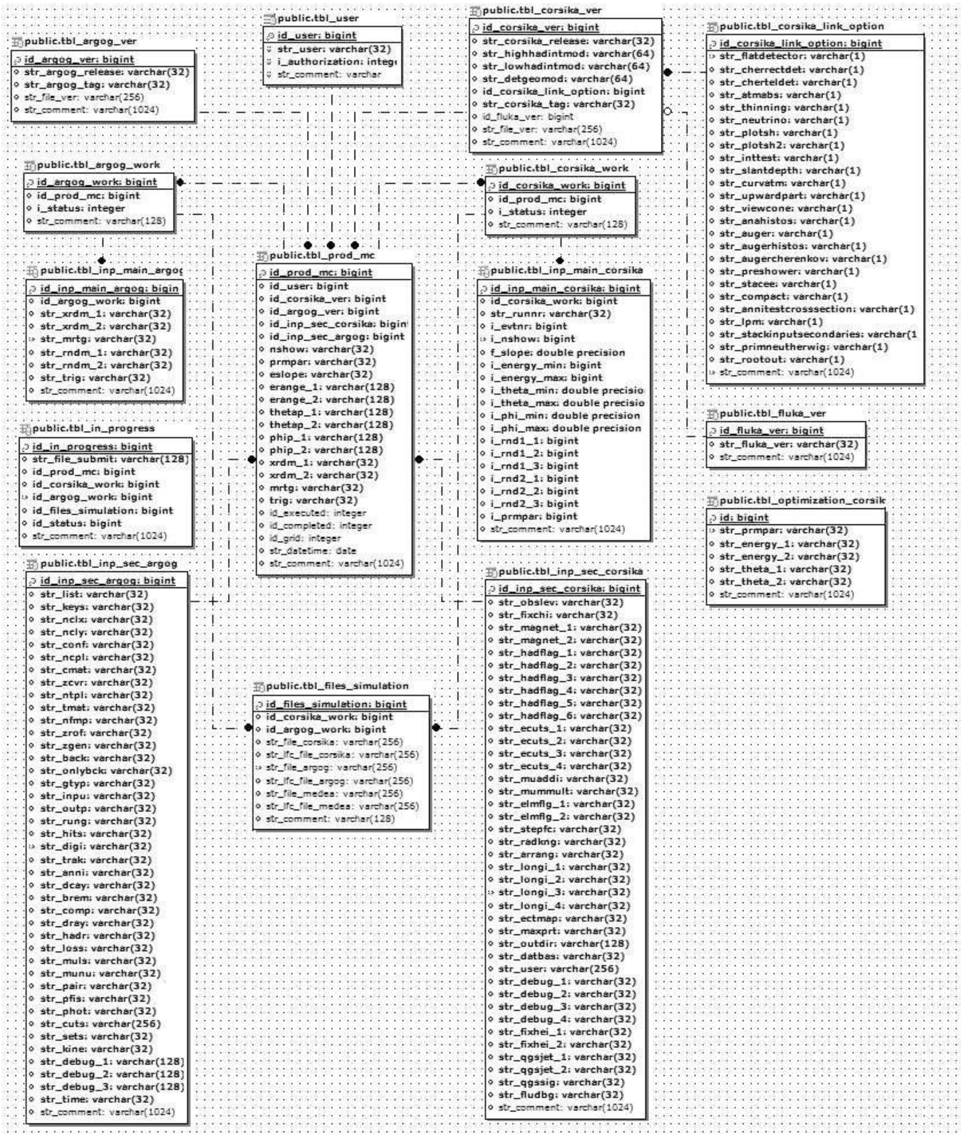


Fig. 2. Database schema

- public.tbl_files_simulation; produce files names and the relative LFC catalog names.
- public.tbl_prod_mc; parameters specifying the current production (showers number, kind of primary particle, range of energy, range of zenith and azimuth angles, trigger used, etc.).
- public.tbl_in_progress; dynamic information about the current production (status of each task)

V. Conclusions

The procedures for Monte Carlo data production were widely tested either on Roma Tre farm and on GRID environment (mainly Roma Tre and Naples GRID farms). This phase of simulated data production was completed and the database efficiency tested.

The simulation environment is now available also for more personalized Monte Carlo production of single users.

References

[1] M.Iacovacci and E. Rossi, *Result overview from the ARGO-YBJ experiment*, CRIS2008 Melfa, Salina Island, Italy (Sept. 15-19 2008)

[2] D. Heck et al., Report **FZKA 6019** (1998).

[3] A. Fassò, A. Ferrari, J. Ranft, and P.R. Sala, *FLUKA: a multi-particle transport code*, CERN-2005-10 (2005), INFN/TC_05/11, SLAC-R-773

[4] GEANT- Detector Description and Simulation Tool, CERN Program Library, Long Writeup W5013 (1993)

[5] C.Stanescu, A.Budano, P.Celio, S.Cellini, F.Galeazzi, Y.Q.Guo, S.Mari, L.Wang and X.M.Zhang, *A GRID approach to ARGO-YBJ experiment data transfer and processing*. ICRC2007, Merida, Mexico