Storage Strategy of AMS Science Data at Science Operation Centre at CERN

Storage Strategy of AMS Science Data at Science Operation Centre at CERN

V Choutko[1], O Demakov[1], A Egorov[1], A Eline[1], B S Shan[2,4] and R Shi[3]

[1] Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA
[2] Beihang University, 37 Xueyuan Road, Haidian Qu, Beijing 100191, China
[3] Southeast University, 2 Sipailou, Xuanwu Qu, Nanjing, Jiangsu 210018, China

E-mail: baosong.shan@cern.ch

Abstract. The Alpha Magnetic Spectrometer (AMS) has collected over 95 billion cosmic ray events since it was installed on the International Space Station (ISS) on May 19, 2011. The AMS science data includes original flight data, reconstructed and simulated ones, as well as the metadata of all of them. The total data volume is more than 1000 TB per year of operation in average, and is now over 6200 TB. One of the major responsibilities of AMS Science Operation Centre (SOC) is the management of science data, including: to receive the original data collected by the detector, to store the data of different types/stages in corresponding places, to back data up, and to create/update metadata (index) for all the stored data. The data validation, the metadata design, and the ways to preserve the consistency between the data and the metadata are presented.

1. Introduction

The Alpha Magnetic Spectrometer [1] is a high energy physics experiment on board of the International Space Station (ISS). The detector has a geometrical acceptance of $0.5m^2 \cdot sr$, and is equipped with: permanent magnet, TimeOfFlight hodoscope, nine layers silicon Tracker, gaseous Transition Radiation Detector, Ring Image Cherenkov Detector and Electromagnetic CALorimeter. The maximum event collecting rate can reach 2 KHz. The physics goals of the AMS are to search for antimatter in the universe on the level of less than 10^{-9}, to search for dark matter in various physics channels and to perform high statistics measurements of cosmic rays composition as well as γ rays.

As the primary payload of the space shuttle Endeavour's last flight, the AMS-02 detector was launched on May 16 2011, and installed on the International Space Station on May 19, 2011. Ever since then, the detector has been collecting the cosmic rays' data and transferring the data to the ground, without significant interruption. Up to now, more than 95 billion events have been collected, resulting 200 TB of scientific data (in RAW format). The average size for each cosmic ray event is 2.2 KB, and the average size for each RAW file is 1.5 GB. The volume of RAW data is in average 35 TB per year, and the volume of the final reconstructed flight data for analysis is in average 130 TB per year.

[4] Corresponding author

We have two storage levels: active/live data and backup data. Active data, including RAW data and datasets active for current analysis, are stored on SOC own storage and/or EOS [2], which can be accessed at any time, and backed up on CASTOR [3]. Due to the limited quota of active storage, inactive data are only stored on CASTOR, but they can be copied to active storage in case of necessary, with a delay of a couple of hours. Tools are developed to automate the data moving and backup procedures.

2. Computing model and data types

The work flow of AMS offline computing is shown in Figure 1:

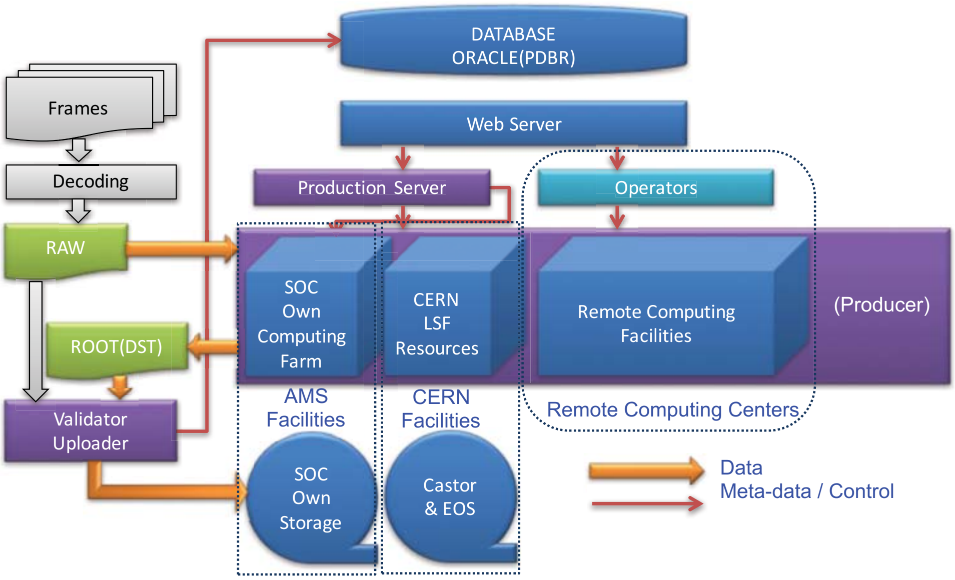

Figure 1. Work flow of AMS offline computing.

2.1. Original flight data

The data collected by the detector contain original electronic signals triggered by cosmic ray particles and recorded by sub-detectors. Before being transferred, the original data are split and packed into *frames*, and each frame contains data collected during the same minute. The one-minute frames are transferred by the Tracking and Data Relay Satellites (TDRS) to White Sands, and then transmitted to the AMS Payload Operations Control Centre (POCC) by the ground system. These data are stored in the POCC storage and shared to SOC by NFS [4]. Frame data are also backed up on CASTOR.

2.2. Preproduction – from frames to RAW

To facilitate detector calibrations and alignments, data of each quarter of the ISS orbit (between either pole and the equator), typically about 23 minutes, are labeled as one *run*. Data from the same run are decoded from frame files, and repacked into one RAW file. Cyclic redundancy check (CRC) [5] is done on each RAW file for consistency check, and then copied to EOS. The metadata of the RAW files, including the run number, the path, the event numbers and collecting time of the first and last events, the total number of events, the CRC checksum, the frame range from which the RAW file is built, etc., are inserted into the Oracle parallel database [6].

2.3. Production – from RAW to ROOT (DST) + TDV

After the metadata for RAW are recorded in the database, production jobs can be requested, and submitted to run on SOC own computing farm or CERN LSF [7] hosts. Production jobs produce ROOT [8] files and Time Dependent Variables (TDV) files for AMS conditional database. TDV files contain the environmental variables which are important for data analysis, for example, the temperature variation of the silicon Tracker sub-detector affects the alignment parameters and further data analysis.

ROOT files are validated and uploaded to the permanent storage, EOS or SOC own storage, depending on the availability, and backed up on CASTOR, and the metadata are also recorded in the database.

TDV files are stored on AFS [9] and published to CVMFS [10] weekly.

Production includes two stages: the first production and the second production. The first production runs in a fully automated manner and produces the data summary files for the quick detector performance evaluation. Usually the reconstructed data are available in two or three hours after the raw data are validated and registered in the database. The second production uses all the available calibrations, alignments and ancillary data from the ISS as well as monitoring values (temperatures, pressures, and voltages) to produce the physics analysis ready dataset. The second production usually runs every 6 months incrementally. However, in case there are major updates in the reconstruction software, a full reproduction may be needed. To identify the "generations" of the second production data, pass-N is used, where N is increased when a reproduction is done by a new version of reconstruction software. The latest second production data is pass-6.

Besides flight data, Monte-Carlo [11] (MC) production generates the simulated data, which include ROOT and RAW files, and are stored on EOS and backed up on CASTOR.

Table 1 is a summary of AMS science data types, the storage places, and the data volumes.

Table 1. Summary of AMS data types, storage destinations, and data volumes.

Data type	Storage	Comment	Per year volume
Frames	POCC storage, CASTOR	One-minute science data	37 TB
RAW	SOC storage, EOS, CASTOR	Data of one run (1/4 ISS orbit)	35 TB
ROOT	SOC storage, EOS, CASTOR	Data of one run (1/4 ISS orbit)	130 TB
Metadata	Oracle DB		-
TDV	AFS, CVMFS		100 GB
Source code	AFS, CVMFS		-

3. Evolution of storage strategy

3.1. EOS as the primary storage

Up to the beginning of 2013, AMS science data were stored on SOC own storage and backed up on CASTOR. In 2013, the storage strategy was changed, and EOS started to be used as the primary storage for science data. SOC own storage is used as a redundant system in case EOS service is degraded or unavailable.

After 4 years operation, EOS proved to be a reliable storage service with good performance, especially for big files. To transfer files to EOS after validation, we use the tool xrdcp [12], which can reach a stable rate of over 100 MB/s, and with the recently enabled Third-Party-Copy option, which means the data will be copied between EOS and CASTOR disk servers, the copying rate can reach over 300 MB/s.

3.2. Deferred CASTOR backup

The volume of AMS science data grows rapidly in the past few years, as shown in Figure 2. Backing up of the Monte-Carlo simulated data and the second production data has been moved out from the validating step to ensure the validation can be done without significant delays. Now the CASTOR backing up is done by CERN FTS3 [13] pilot service, which schedules the data copying directly between the EOS and CASTOR disk servers, resulting a performance boost.

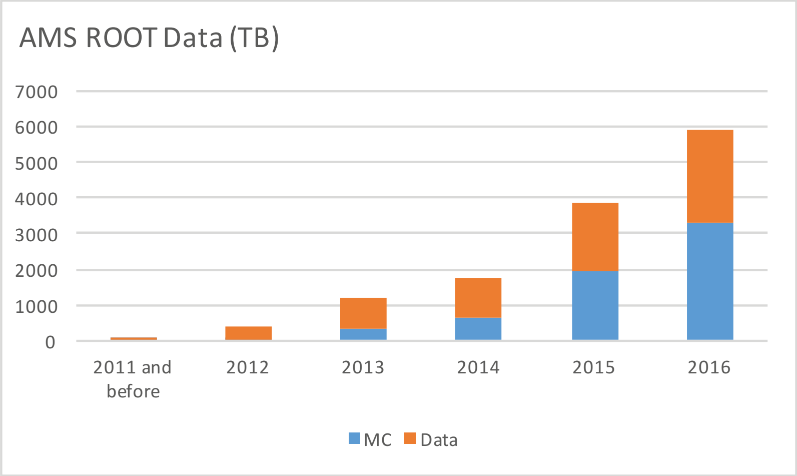

Figure 2. Total amount of AMS ROOT data (accumulative by year).

FTS3 works as a delegation service, which accepts a list of original/destination path pairs, and processes the transferring requests by copying between storage servers. To automate the backup procedures, a backup daemon is developed to generate backup list, submit and monitor FTS3 jobs, and update metadata of the files which have been backed up.

The daemon runs on a dedicated virtual host managed by Puppet [14], and uses a robot certificate for authorisation. A sqlite3 [15] database is used to record the information of each backup task.

3.3. Generating backup list

The daemon queries the Oracle database to find out which files on EOS are not backed up, and for each file, it writes the SRM [16] URL [17] of the EOS path and the destination CASTOR path into a text file, for example:

```
srm://srm-eospublic.cern.ch:8443/srm/v2/server?SFN=/eos/ams/MC/AMS02/2014/O.B1050/o16.
    pl1.l19.1664000.112/872813981.00000001.root srm://srm-public.cern.ch:8443/srm/
    managerv2?SFN=/castor/cern.ch/ams/MC/AMS02/2014/O.B1050/o16.pl1.l19
    .1664000.112/872813981.00000001.root
```

3.4. Managing FTS3 jobs

After the input text file is ready, the daemon initializes its GRID [18] certificate, and submits an FTS3 job using the text file as the transferring list. Following a successful submission, the daemon registers the information of this backup task to its sqlite3 database, and starts to monitor the status of the FTS3 job.

3.5. Updating metadata

Once the job terminates, the daemon scans the file list of the completed transfers, compares the checksum values on EOS and CASTOR, and if they agree, updates the Oracle database of the CASTOR time of this file.

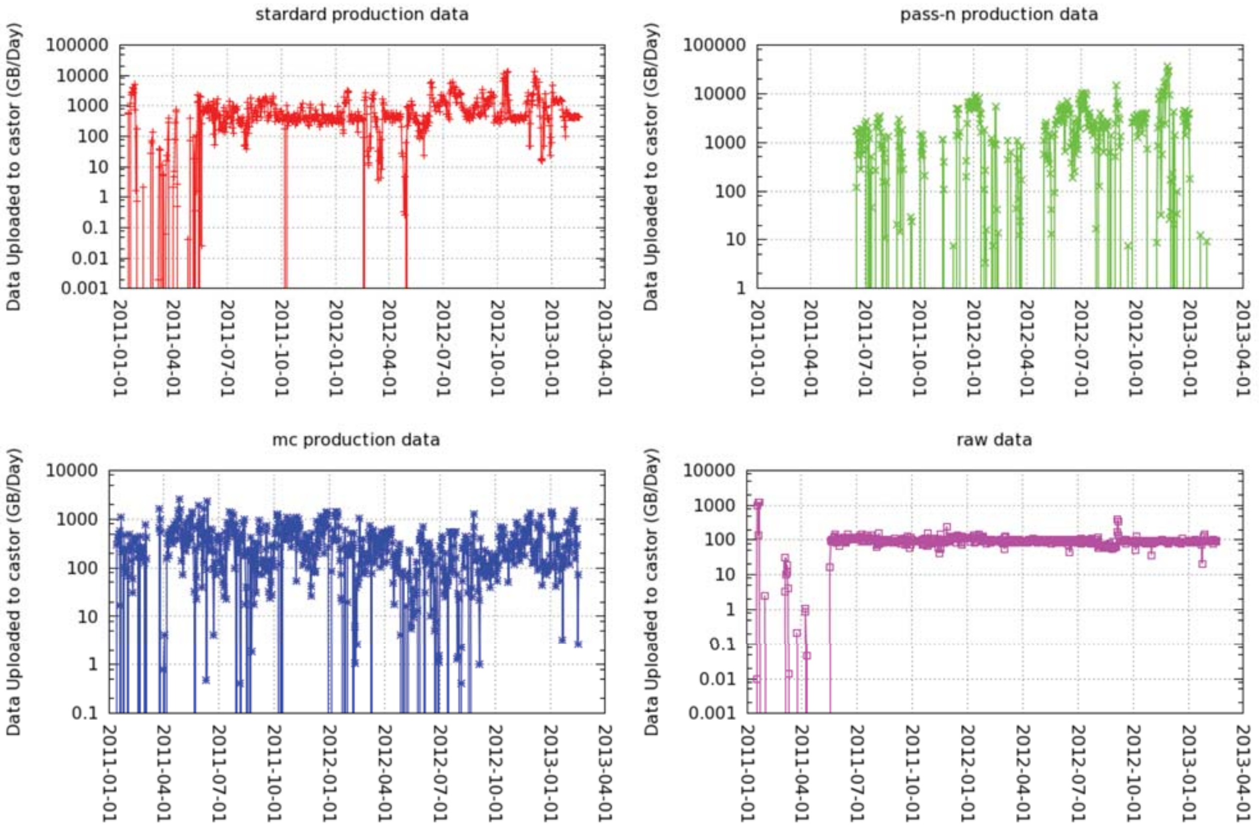

Figure 3. CASTOR backing up rate for different streams.

3.6. Performance of FTS3 backup

Figure 4 shows the performance of backing up pass-6 production data to CASTOR. For this backup task, totally 54 TB data from 31528 files was copied from EOS to CASTOR in 9 hours, and the average transferring rate is 1.7 GB/s, equivalent to 149 TB per day.

Figure 5 shows the performance of backing up electron simulation data (el.B1064) to CASTOR. For this backup task, totally 85 TB data from 103972 files was copied from EOS to CASTOR in 11 hours, and the average transferring rate is 2.1 GB/s, equivalent to 182 TB per day.

4. Storage and curation policies of AMS science data

As several storages are used for keeping AMS science data, and the volume is growing fast, it is necessary to develop new tools and to improve existing tools to automate the data moving procedures.

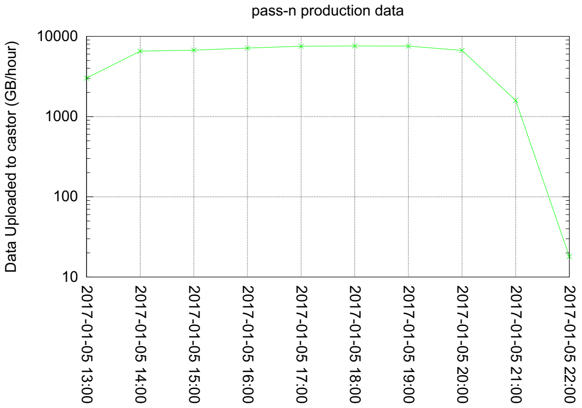

Figure 4. FTS3 backup rate of pass-6 data.

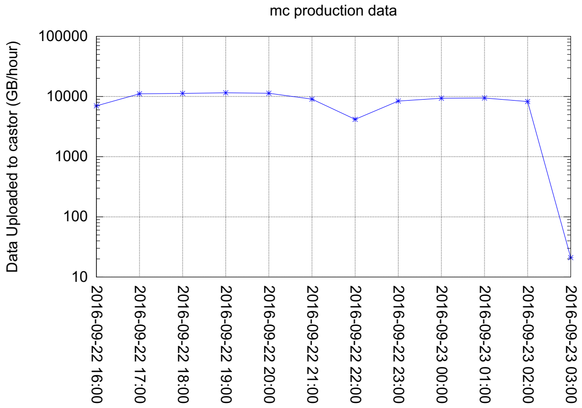

Figure 5. FTS3 backup rate of MC data.

4.1. Dataset/template based storage policies

All the data production and Monte-Carlo simulation tasks are organised by datasets and templates. Within the same *dataset* the jobs have the same version of software and even the same particle type (only for MC simulation). A dataset consists of one or more job *templates*, and the jobs from the same template have the same reconstruction/simulation parameters.

We store a set of policies for each dataset/template pair, to specify which storages the data should go, and which data moving tools should be used. Table 2 is an example of several datasets and templates and their storage policies. The first production (std.job) is running constantly, and the backing up is done by the validator; pass-6 and simulated data are backed up by FTS3; pass-4 data is only stored on CASTOR as they are not actively in use.

Table 2. Example of data placement policies

Dataset/Template	Active storage	Backup storage	Backing up tool
ISS.B1070/std.job	EOS/SOC own storage	CASTOR	By validator
ISS.B950/pass6.job	EOS/SOC own storage	CASTOR	By FTS3
He.B1081/*	EOS/SOC own storage	CASTOR	By FTS3
ISS.B620/pass4.job	-	CASTOR	-

4.2. Data flow

Figure 6 shows the data flow between different storages.

EOS works as the primary storage and files store on SOC own storage only if EOS is not available. In such case, those files will be transferred to EOS by a crontab task running on SOC cluster when EOS is available again. And backup daemon always copies files from EOS to CASTOR.

In case for some reason an inactive dataset/template which has been removed from EOS becomes active again, a similar daemon uses FTS3 service to transfer data from CASTOR to EOS and updates the metadata in the Oracle database.

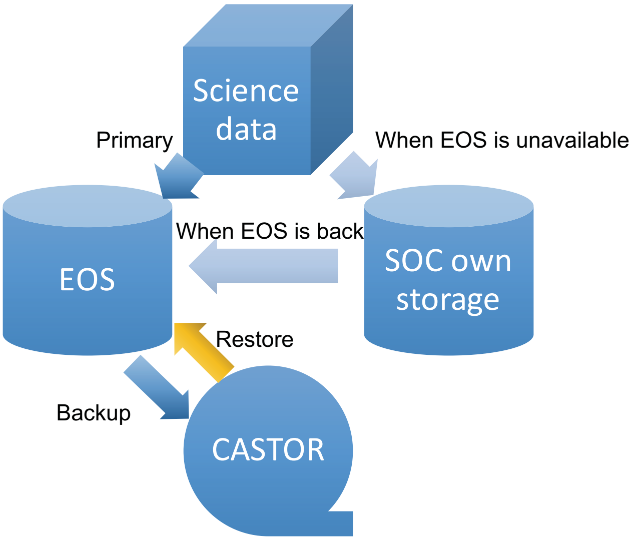

Figure 6. Data flow between storages.

5. Conclusion

The storage strategy for AMS science data is designed to work with the variety of different file systems, and to cope with the changing of the data volume and storage systems. The data management tools ensure appropriate data placement and efficient data moving among storages.

References

[1] Ting S 2013 *Nuclear Physics B-Proceedings Supplements* **243** 12–24

[2] Peters A J and Janyst L 2011 Exabyte scale storage at CERN *Journal of Physics: Conference Series* vol 331 (IOP Publishing) p 052015

[3] Presti G L, Barring O, Earl A, Rioja R M G, Ponce S, Taurelli G, Waldron D and Dos Santos M C 2007 CASTOR: A distributed storage resource facility for high performance data processing at CERN. *MSST* vol 7 (Citeseer) pp 275–280

[4] Nowicki B 1989 NFS: Network file system protocol specification Tech. rep.

[5] Sobolewski J S 2003

[6] DeWitt D and Gray J 1992 *Communications of the ACM* **35** 85–98

[7] Zhou S 1992 LSF: Load sharing in large heterogeneous distributed systems *I Workshop on Cluster Computing* vol 136

[8] Antcheva I, Ballintijn M, Bellenot B, Biskup M, Brun R, Buncic N, Canal P, Casadei D, Couet O, Fine V *et al.* 2011 *Computer Physics Communications* **182** 1384–1385

[9] Satyanarayanan M 1993 *Distributed Systems. Addison-Wesley and ACM Press,* **821** 145–154

[10] Aguado Sanchez C, Bloomer J, Buncic P, Franco L, Klemer S and Mato P 2008 CVMFS - a file system for the CernVM virtual appliance *Proceedings of XII Advanced Computing and Analysis Techniques in Physics Research* vol 1 p 52

[11] Binder K 1987 *Quantum Monte Carlo Methods* 241

[12] Dorigo A, Elmer P, Furano F and Hanushevsky A 2005 *WSEAS Transactions on Computers* **1**

[13] Ayllon A, Salichos M, Simon M and Keeble O 2014 FTS3: New data movement service for WLCG *Journal of Physics: Conference Series* vol 513 (IOP Publishing) p 032081

[14] Loope J 2011 *Managing Infrastructure with Puppet: Configuration Management at Scale* (“ O’Reilly Media, Inc.”)

[15] Owens M and Allen G 2010 *SQLite* (Springer)

[16] Perelmutov T, Bakken J and Petravick D 2004 Storage resource manager *Proceedings of the 2004 Conference on Computing in High Energy and Nuclear Physics (CHEP?04)*

[17] Berners-Lee T, Masinter L and McCahill M 1994 Uniform resource locators (url) Tech. rep.

[18] Foster I and Kesselman C 2003 *The Grid 2: Blueprint for a new computing infrastructure* (Elsevier)