



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

V. Choutko¹, O. Demakov¹, A. Egorov¹, A. Eline¹, B. Shan², R. Shi³

¹Massachusetts Institute of Technology ²Beihang University ³Southeast University

Abstract

This paper introduces the storage strategy and tools of the science data of the Alpha Magnetic Spectrometer (AMS) in the Science Operation Center (SOC) at CERN.

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 currently reached over 5200 TB. We have two storage levels: active/live data which is ready for analysis, and backups on CASTOR. Active/live data are stored on SOC own storage, and on CERN EOS. Tools are designed to automate the data moving and data backup.

The data validation, the metadata design, and the ways to preserve the consistency between the data and the metadata are presented.

Computing model and data types

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

- Original flight data arrives at AMS SOC
The data collected by the detector are packed into one-minute *frames* and transferred to SOC.
- Preproduction – frames → RAW
Frames are decoded and repacked into *RAW* files, and each RAW file contains one run, which consists of data of 1/4 of the International Space Station (ISS) orbit.
- Production – RAW → ROOT (DST) + TDV
After the *metadata* for RAW are recorded in the database, production jobs will be requested, and submitted to run on SOC own computing farm and CERN LSF resources. Production jobs produce *ROOT* files and *Time Dependent Variables (TDV)* files for AMS conditional database. ROOT files are validated and uploaded to the permanent storage, SOC own storage or CERN EOS, and backed up on CERN CASTOR, and the metadata are also recorded in the database. TDV files are stored in CERN AFS and published to CERN CVMFS periodically. Production includes two stages:
 - Standard: for performance checking and calibration
 - Pass-n: for physics analysis
- Monte-Carlo production
 - Simulated data, transferred back to CERN

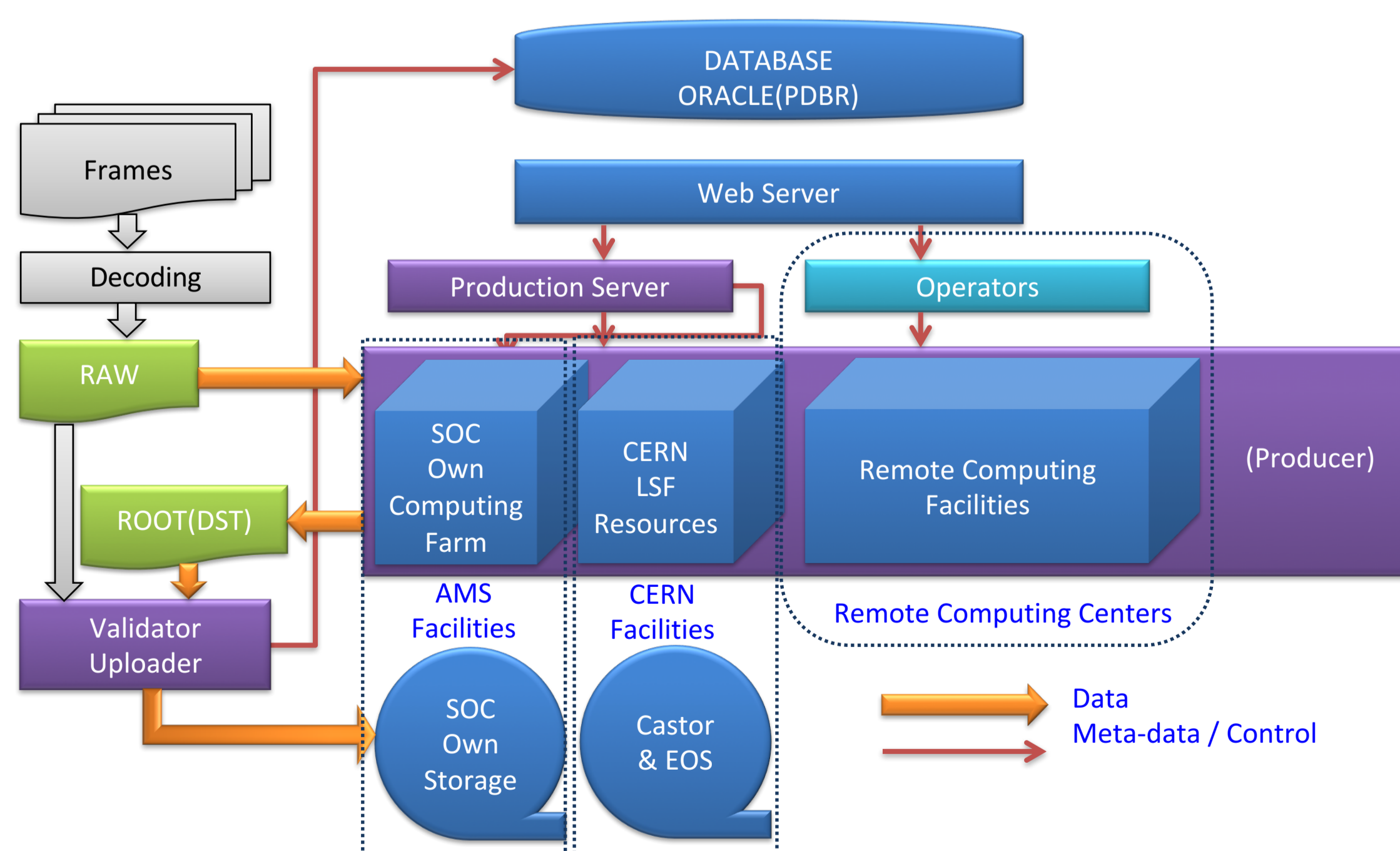


Figure 1: Work flow of AMS offline computing

Data type	Storage	Comment
Frames	SOC own storage, CASTOR	One-minute science data
RAW	SOC own storage, EOS, CASTOR	Data of one run (1/4 ISS orbit)
ROOT	SOC own storage, EOS, CASTOR	Data of one run (1/4 ISS orbit)
Metadata	Oracle DB	
TDV	AFS, CVMFS	
Source code	AFS, CVMFS	

Table 1: Summary of AMS data types and storage destinations

Evolution of storage strategy

Up to 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.

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

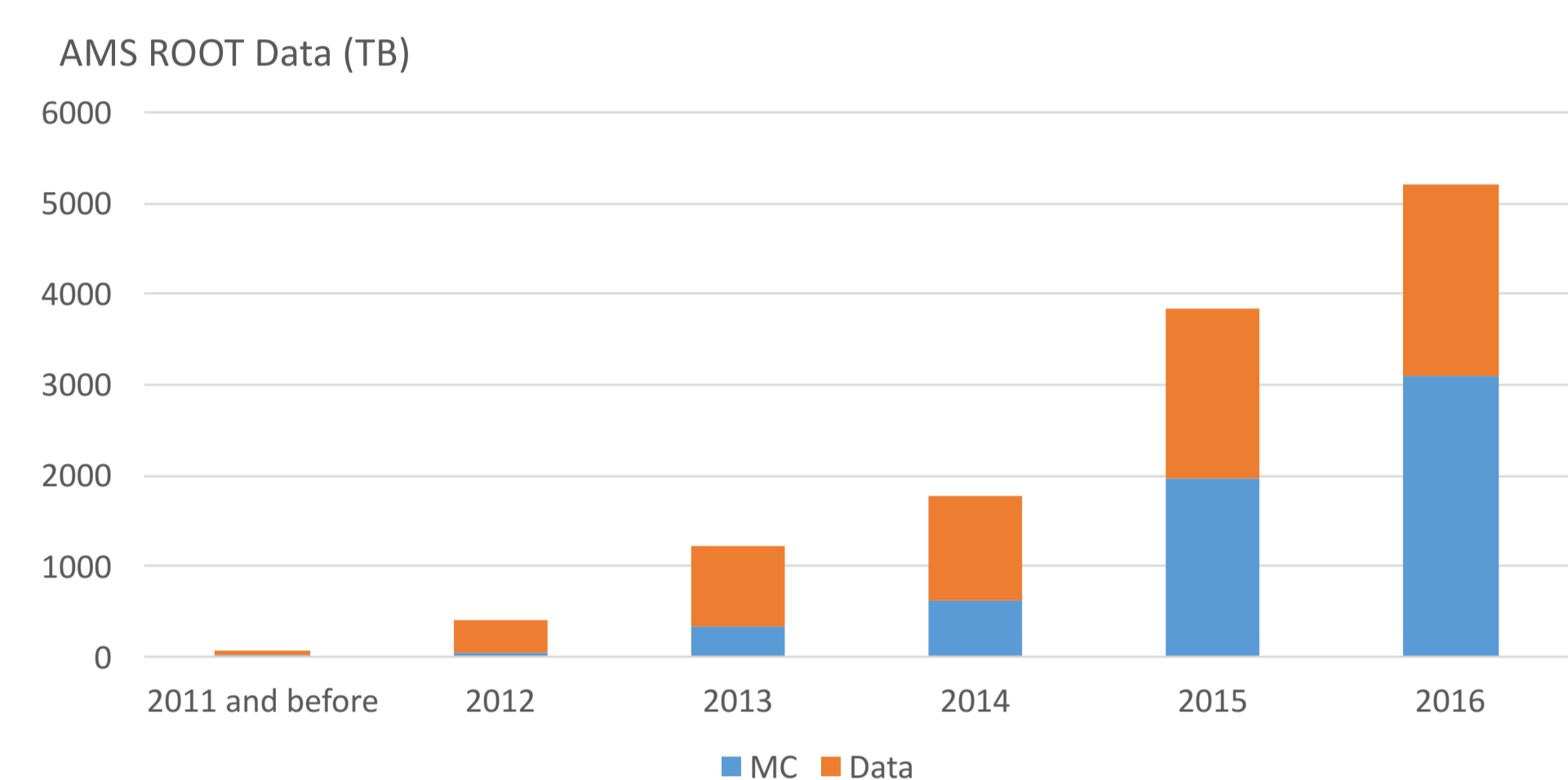


Figure 2: Total amount of AMS ROOT data.

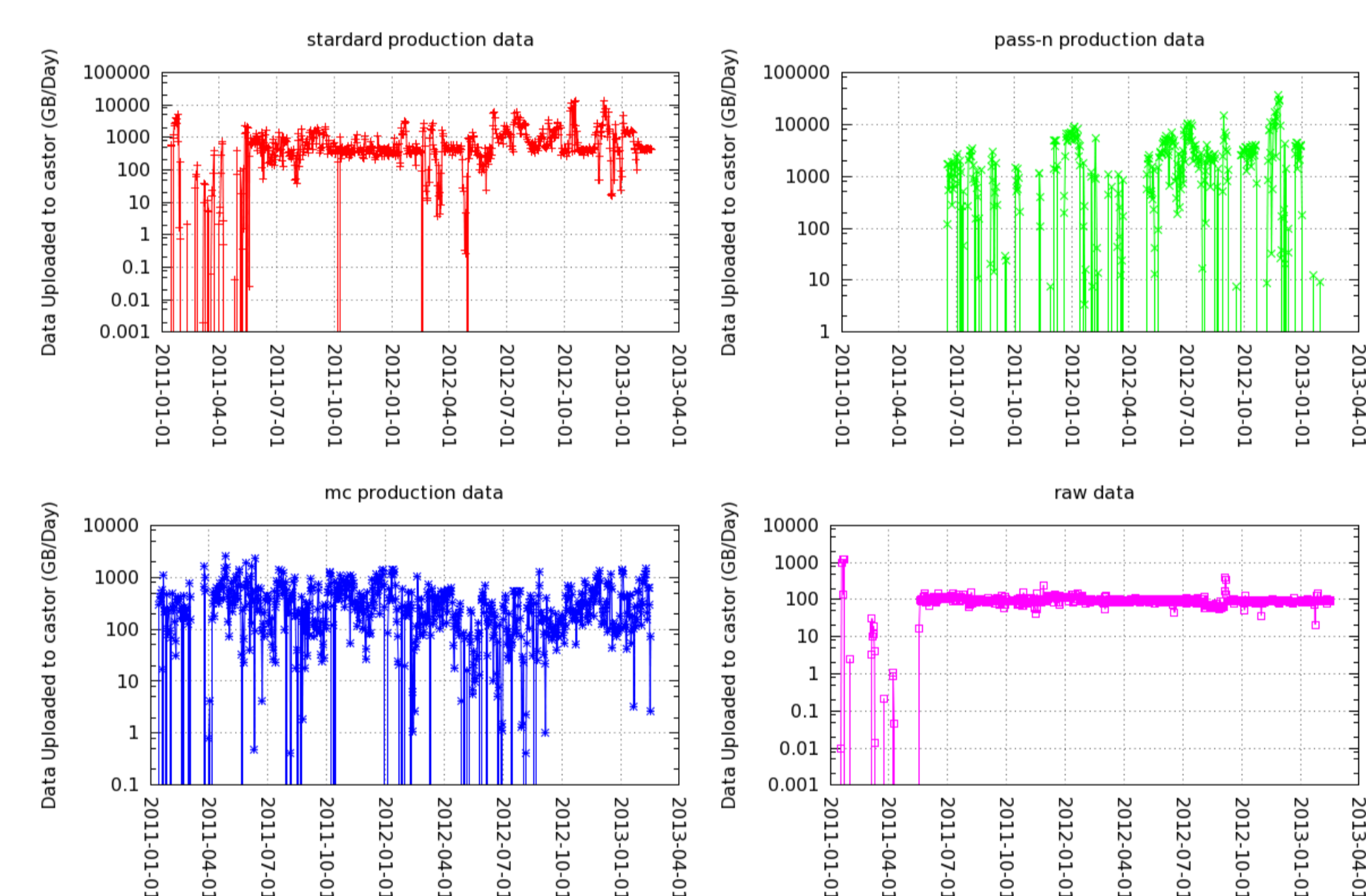


Figure 3: CASTOR backing up rate for different streams.

Data management tools

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.

Dataset/template based storage policies:

We store a set of policies for each of our 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. Standard production 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.

Dataset/Template	Storage	Backing up tool
ISS.B1070/std.job	EOS, CASTOR	By validator
ISS.B950/pass6.job	EOS, CASTOR	By FTS3
He.B1081/* .job	EOS, CASTOR	By FTS3
ISS.B620/pass4.job	CASTOR	-

Table 2: Example of storage policies

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 data volume and storage systems. The data management tools ensure appropriate data placement and efficient data moving among storages.