AMS-02 Monte Carlo Production in Science Operation Centre at Southeast University

IOP Conf. Series: Journal of Physics: Conf. Series **898** (2017) 082022 doi:10.1088/1742-6596/898/8/082022

AMS-02 Monte Carlo Production in Science Operation Centre at Southeast University

Junzhou Luo[1], Jinghui Zhang[1], Fang Dong[1], Aibo Song[1], Runqun Xiong[1], Jiyuan Shi[1], Feiqiao Huang[3], Renli Shi[1], Zijian Liu[1], Vitaly Choutko[2], Alexander Egorov[2] and Alexandre Eline[2]

[1]School of Computer Science and Engineering, Southeast University, Nanjing, China
[2]Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA
[3]Baidu, Inc.

E-mail: jhzhang@seu.edu.cn

Abstract. Southeast University (SEU) Science Operation Centre (SOC) is one of the computing centres of the Alpha Magnetic Spectrometer (AMS-02) experiment. It provides 2016 CPU cores for AMS Monte Carlo production and a dedicated ~1Gbps Long Fat Network (LFN) for AMS data transmission between SEU and CERN. In this paper, the development and deployment of SEU SOC's automated Monte Carlo production management system is discussed in detail. Data transmission optimizations are further introduced in order to speed up the data transfer in LFN between SEU SOC and CERN. In addition, monitoring tool for SEU SOC's Monte Carlo production is also presented.

1. Introduction

The Alpha Magnetic Spectrometer (AMS-02) is a general purpose high-energy particle physics detector designed to operate on the International Space Station (ISS) [1]. AMS-02 detector has a weight of 6 metric tons and its key components include: the permanent magnet, time of flight, silicon tracker, transition radiation detector, ring image Cerenkov detector and electromagnetic calorimeter. It uses the unique environment of space to study the universe and its origin by searching for antimatter and dark matter in the universe while performing precision measurements of cosmic rays composition and flux in the energy range from GeV to TeV [2].

AMS-02 was brought to the ISS with Space Shuttle Endeavour On May 16th 2011. Data taking began immediately after AMS-02 was mounted on the ISS on May 19, 2011. Since then, more than 90 billion cosmic ray events have been collected and transmitted to the ground. As shown in Figure 1, collected AMS data are transmitted to ground via the TDRS satellites and then relayed to Marshall Space Flight Center (MSFC). At MSFC data are firstly transferred to the AMS relay computers and then to the AMS Science Operation Center (SOC) at CERN [3]. Subsequently, data are distributed from CERN SOC to AMS remote SOCs in China, Germany, Italy, France and etc.

Southeast University (SEU)'s SOC is one of AMS major remote SOCs in China and provides 2016 CPU cores for AMS Monte Carlo (MC) production [4]. In MC production, hundreds of billions of simulated AMS events are produced using dedicated software developed based on the GEANT4 [5] by the AMS collaboration. This software simulates electromagnetic and hadronic interactions of particles in the material of AMS and generates detector responses. The simulated events then undergo the same

reconstruction procedure as used for those events collected by AMS detector in ISS. As shown in Table 1, AMS MC production uses the computing power of remote computing centers. The AMS MC data processing flow, including each center's requesting MC job from CERN, executing those job on local computing facility and transferring generated MC data back to CERN, have much in common for all the remote computing centers participating in AMS MC Production as shown in Figure 1.

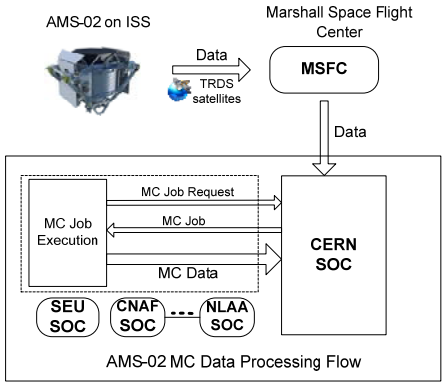

Figure 1. AMS-02 data processing flow.

Table 1. List of AMS MC Production centers.

Remote Center	CPU Cores for MC Production
CNAF, Bologna	500
JUROPA, Juelich	2000
IN2P3, Lyon	700
Acad. SINICA, Taipei	1000
SEU, Nanjing	2016
NLAA, Beijing	1024

However, each computing center's existing hardware, networking and software varies much from each other, e.g., NLAA uses 3-node shared memory HPC system while SEU is a typical 168-node HPC cluster; CNAF, JUROPA and IN2P3's networking condition is much better than SEU's network connection across Asia and Europe to CERN; NLAA simply uses function tightly coupled with operating system to schedule computing resources, and instead SEU adopts IBM LSF to allocate resources to MC production job. Therefore, a customized MC production management system has been developed for improving the efficiency of MC production at SEU SOC. More specifically, SEU's MC production management system is specifically developed for its deployment on SEU's clustered servers with particular optimizations on automated MC production execution and data transfer on the 1Gbps long fat network between SEU and CERN.

In this paper, we introduce the customized Monte Carlo production management system at SEU SOC in Section 2. The details of data transmission optimization for long fat network between SEU and CERN are described in Section 3. In addition, Monte Carlo production monitoring tool is introduced in Section 4. Conclusions are given in Section 5.

2. Monte Carlo Production Management System at SEU SOC

Monte Carlo production is massively conducted at SEU SOC by active request for Monte Carlo production job from CERN, submission/execution of job at SEU SOC and final transmission of the output data to CERN SOC. The output of each MC job is organized as two data files. One is "root" file that represents the simulated events based on ROOT [6] package and can be used for data analysis. The other "raw" file is with format equivalent to AMS raw data, and it can be used subsequently to rerun of the simulated data with different AMS offline software versions if needed.

In order to automate the MC production, the following hardware are specially configured, organized and deployed at SEU SOC, which is shown in Figure 2. SEU SOC's existing hardware is divided and organized as a management server, a set of computing servers, several data transfer servers and a storage pool. The management server coordinates and manages the overall MC production, while the computing servers are used as aggregated computing resources for carrying out massive MC job executions. The output data of MC production job are then written to the storage pool, and the data transfer servers are responsible for transferring the output MC data from local storage pool to CERN SOC.

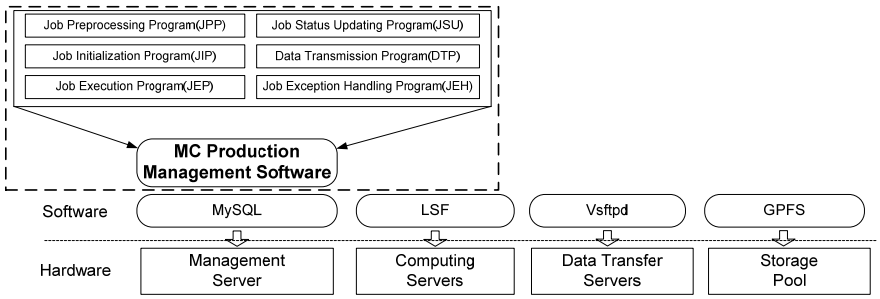

Figure 2. Software and hardware deployment for MC Production at SEU SOC.

In addition, the necessary software is also deployed on SEU SOC's hardware. For instance, the General Parallel File System (GPFS) is deployed on all servers from the storage pool and Load Sharing Facility (LSF) [7] is installed on all computing servers for management of job scheduling and execution. Vsftpd, using as FTP server, is deployed on data transfer servers accordingly. A MySQL-based database is installed on the management server to dynamically aggregate the information of MC job status and the growth of output data during the MC production.

Most importantly, MC production management software, as the core part of the SEU's automated Monte Carlo production, is specifically developed for its deployment on SEU's clustered servers with particular optimizations on automated MC production execution and data transfer on the 1Gbps long fat network between SEU and CERN. The software is composed of the following sub-programs:

- Job Preprocessing Program (JPP): this program is responsible for preprocessing procedures after the MC job have been requested.
- Job Initialization Program (JIP): it is used for initialization before starting MC job execution.
- Job Execution Program (JEP): this is used to start job execution on computing servers.
- Job Status Updating Program (JSU): it is for monitoring the job status and managing job execution.
- Data Transmission Program (DTP): its function is to manage the data transmission of MC production.
- Job Exception Handling Program (JEH): it is used to handle the exception during the MC production.

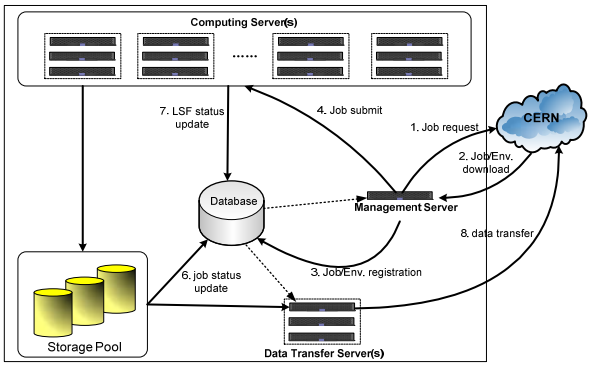

Figure 3. Automated Monte Carlo production at SEU SOC.

The complete procedure of automated Monte Carlo production is customized for SEU SOC, which is shown in Figure 3 and can be described as follows. Each Monte Carlo job includes a few job description/script files and several related software packages. Job need to be downloaded remotely from CERN and extracted on local computing servers automatically, and then are registered into database. After that, the MC production management software will check the database regularly and submit job to SEU computing cluster by LSF for execution. Along with the job execution, it will produce two types of files: the output file that contains both "root" file and "raw" file; the "journal"

file that has logs for all activities of job execution. During the job execution, the status of a job will be regularly refreshed by checking its related "journal" file. Once there is a "root" file being validated, this file will be transferred to CERN immediately. When a job is completely finished, its related "raw" file and "journal" file will also be transferred to CERN automatically.

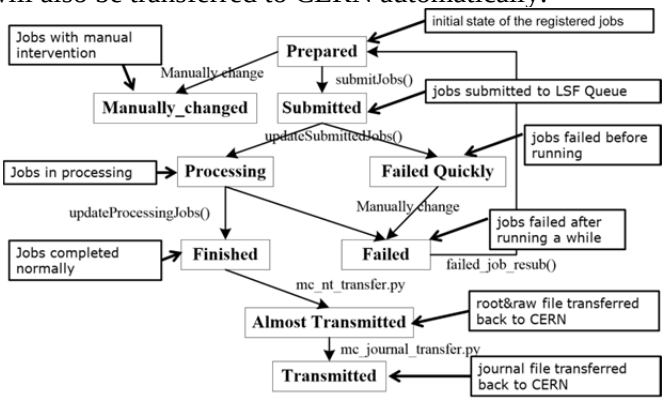

Figure 4. Monte Carlo job state transition diagram.

More specifically, as introduced by the MC job state transition diagram in Figure 4, the complete MC production can be organized into steps as follows.

2.1. Job Request

Each Monte Carlo job consists of two parts: job files (the "mc-scripts" files) and the execution environment files (the "mcdb" and the "mcdb-addon" packages). In order to execute a Monte Carlo job, all these files/packages need to requested and downloaded to local computing servers. For each Monte Carlo job, the "mcdb" file, and the "mcdb-addon" package can be fetched via the CERN VM File System. However, the "mc-scripts" file can only be requested by manually filling out the AMS Production Remote Client Form. After filling out the form as requested, the AMS production web site will send an email with attached "mc-scripts" files that correspond to those requested job to our specified email address.

2.2. Job Retrieving and Preparation

When the JPP program probes a new coming email attached with a "mc-scripts" file, it will download this file and extract it in a folder on the local directory. Subsequently, the program will also start to download two related running environment packages (mcdb.tar.gz and mcdb.addon.tar.gz) through the CERN VM File System. These two packages will be saved into the same directory with job-scripts "mc-scripts". If all the three packages are obtained without errors, an empty flag file that has the same name as the job folder and has a suffix ".new" will be created by the JPP program.

2.3. Job Registration and Submission

Once a new Monte Carlo job represented by ".new" suffix is detected, the JIP program will check the directory to verify the running environment packages, mcdb.tar.gz and mcdb.addon.tar.gz and these packages will be extracted after verification. After that, the JIP program will also parse the filename of each job file and write the related information into the database, e.g., job_name, job_runid, data_set. Up to now, the status of this batch of job in database is marked as "prepared". The JEP program will check the database and invoke a LSF "bsub" command to submit a "prepared" job and update its status as "submitted" in database. If everything above runs successfully, the job now enters the stage of execution.

2.4. Job Execution Management

Job execution management is mainly responsible for achieving two goals. One goal is to monitor the status of the job execution and update the job status in database correspondingly. The other is to

perform various management operations according to the status of the job as demonstrated in Figure 4. More specifically, the database is updated by the JSU program every half hour and corresponding operations will be taken depending on the status changes:

- For the "submitted" job, if they are processed in normal way (according to the information of journal files and LSF status obtained by the JSU program), their status will be updated as 'processing' and the information of those corresponding "root" files will be recorded into related tables in database.
- For the "processing" job, the JSU program will check their journal files, and if their status changes to finished, the job status will be updated as 'finished' in database. The related "raw" files and "journal files" will be marked as ready for transmission, and some other information (such as finished time, CPU time, events processed, etc.) will also be recorded into database. If their status still is processing, the JSU program will check the LSF status to make sure there is no exceptions and repeat the process until they are finished.
- If the JSU program detects any exceptions during job running progress, it will mark this job status as 'Failed Quickly' or 'Failed' in the database and terminate this job. We have a special exception handler program JEH to handle these 'Failed Quickly' or 'Failed' job. These job will be resubmitted for execution by this JEH program.

2.5. Data Transfer Management

During the job execution, the DTP program will regularly refresh MC job output "root" files' status by checking job related "journal" file. Once there is a "root" file being validated, this file will be transferred to CERN immediately. After all "root" files have been transferred to CERN, job related "raw" file will be transferred to CERN and job status will be updated as "Almost Transmitted" in the database. The DTP program will also regularly check all the "Almost Transmitted" job "journal" file. If the job status is completely finished, its related "journal" file will also be transferred to CERN automatically and this job status is updated as "Transmitted" in the database.

3. Monte Carlo Production Data Transfer Optimization for Long Fat Network

Massive Monte Carlo production output data, that includes all "root", "raw" and "journal" files generated by successful job execution at SEU SOC, need to be transferred instantly back to CERN for physicist's active demand. Therefore, data transmission optimizations are made to speed up the data transfer in LFN between SEU and CERN.

3.1. Introduction to the Data Transfer

EOS is an open source distributed disk storage system widely used at CERN, and AMS also uses EOS to store massive Monte Carlo production data. Though EOS cannot be accessed from outside CERN because of strict security restrictions, the lxplus [8] cluster at CERN is open to outside CERN and EOS can be accessed from there by mounting EOS to lxplus as a local filesystem with EOSFUSE. However, there are several known issues with EOSFUSE, including high latency on file creation, limited number of files and transport endpoint disconnection caused by crashed EOSFUSE daemon [12].

Therefore, we currently adopt the other approach by using lxplus cluster as the relay servers for transferring the MC production data from SEU to EOS storage at CERN: MC Production data generated in SEU are firstly transferred to the lxplus cluster, and then transferred from the lxplus cluster to EOS. Since the size of "root" files may reach a few hundred megabytes and the networking distance between SEU and CERN is quite long, we empirically choose lftp [9] as the software to transfer MC production data from SEU to the lxplus cluster for the data transmission reliability achieved. Besides that, because the lxplus cluster and the AMS EOS storage are all inside CERN, we use xrdcp [10] to copy MC Production data between them. The "journal" files transferred to CERN are used to validate whether the "root" files and "raw" files are results of successful job execution, so the "journal" files are transferred independently, and accordingly two separate scripts for transmission are

designed: mc_nt_transfer.py for transferring the “root” and “raw” files, and mc_journal_transfer.py for transferring the “journal” files.

3.2. Optimization of the Data Transfer

The data transfer network between SEU and CERN is a typical long fat network for that the bandwidth is around 1Gbit/s and the round-trip delay time (RTT) is approximately 300 ms. Due to its high latency and large capacity, data transfer via such a network may incur relatively slow transmission speed. In order to address this, two optimized transmission strategies in application layer and TCP layer respectively, are designed to improve the transmission performance respectively.

(a) *Dual-host transmission in application layer*

The characteristics of segmented data transmission between SEU and CERN are taken into consideration: when files are being transferred from the lxplus cluster to the AMS EOS storage, the network from SEU to CERN would be idle. Therefore, the transmission can be speed up by parallelizing the transmission between these two networks. As shown in Figure 5, two data transfer servers in SEU cluster and three servers in the lxplus cluster are simultaneously used for transferring MC production data from SEU to CERN. More specifically, we run mc_nt_transfer.py simultaneously on two servers in the lxplus cluster to receive MC production output data: once one server in the lxplus cluster has received MC production output data and is transferring them to the AMS EOS storage, the other server in the lxplus cluster can simultaneously receive other data files from SEU and thus increase the network utilization. In addition, multithread is adopted in these transmission scripts with a varying number of thread n according to network conditions: when the network has sufficient bandwidth, we increase n to maximize network utilization; otherwise, we simply set n to 1 as to transfer files sequentially with the earliest possible time.

(b) *Congestion control in TCP layer*

In order to speed up the data transmission in LFN between SEU and CERN, we need to make the TCP window size reach the optimal value as soon as possible and adjust the size dynamically according to the network condition. Therefore, we use TCP sliding window mechanism and TCP congestion control mechanism to tune the performance of the long RTTs network between SEU and CERN. Hybla [11], which is suitable for increase the transmission rate of large latency network, is adopted as the TCP congestion control strategy. For the increasing of congestion window size (*cwnd*), Hybla uses the following formulas for slow start and congestion avoidance:

$$cwnd = \begin{cases} cwnd + 2^{p} - 1 & \text{slow start} \\ cwnd + \dfrac{p^{2}}{cwnd} & \text{congestion avoidance} \end{cases}$$

And $p=RTT/RTT_0$, RTT is set to 300ms and RTT_0 is set to the Linux default value of 25ms. By adopting the optimized transmission strategy in TCP layer and application layer, the data transmission rate in LFN between SEU and CERN is ~2 times faster (in blue) than before (in black) and the amount of data transferred per day can reach a maximum of 2.5TB, as shown in Figure 6.

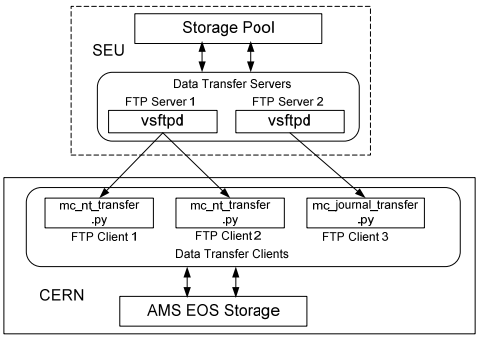

Figure 5. Data transfer facility at SEU and CERN.

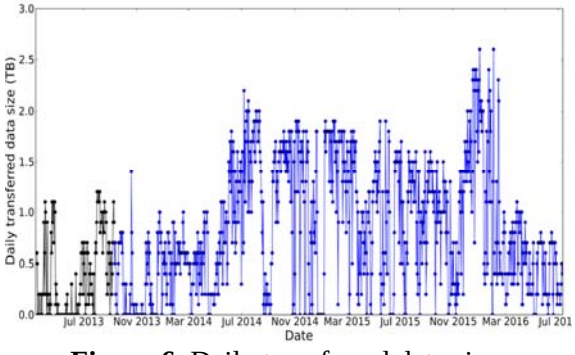

Figure 6. Daily transferred data size.

4. Monte Carlo Production Monitoring Tool

Production monitoring tool provides much user friendliness in the visualization for AMS Monte Carlo production. Database design for MC production and production monitoring website are introduced in this section.

4.1. Database for Monte Carlo Production

For recording the processing progress of each job, we created a database named "seumcjobs", in which three main tables named "jobs", "root-files" and "transmission-status" are designed. The table "jobs" is the basic information table, and all the Monte Carlo job requested from CERN will be registered to this table at the first time. It stores production job information (such as job-runid, job name, dataset, log file etc.) and its running environment configuration (such as mcdb name, mcdb version, mcdb update time, cpu time, job threads etc.). The table "root-files" stores the generated root/raw files' information including the runid (consistent with the job-runid in table "jobs"), name, size, location, status etc. The "transmission-status" table records the transmission information of each file, such as run-id (consistent with the job-runid in table "jobs"), file name, file size, type, transmission duration, host names etc.). Table 2 shows tables' information in the database seumcjobs.

Table 2. Main tables in seumcjobs.

#	Tables	Comments
1	amsuser	the information of database users
2	jobs	the basic information of MC job
3	root-files	the information of the MC root/raw files
4	Transmission-status	the information of each file transmission

4.2. Production Monitoring Tool

To visualize the MC production information, a production monitoring website is designed as the monitoring tool. We use php to get the production information from the MySQL database and display the information by a Javascript chart library ECharts.

After login the website displays four kinds of information: MC Production, MC Transmission, RAW Backup and disk space. For each job there are 8 possible states (prepared, submitted, processing, finished, almost transmitted, transmitted, manually changed and failed). The MC Production page shows the amount of job at each state (see Figure 7). The failure reasons for the failed job are also posted on this page. The MC Transmission page reveals the daily statistical MC transmission status within one year. The RAW Backup page displays the raw data backup status of the past 9 days. The Disk Space page shows the SEU cluster disk usage status.

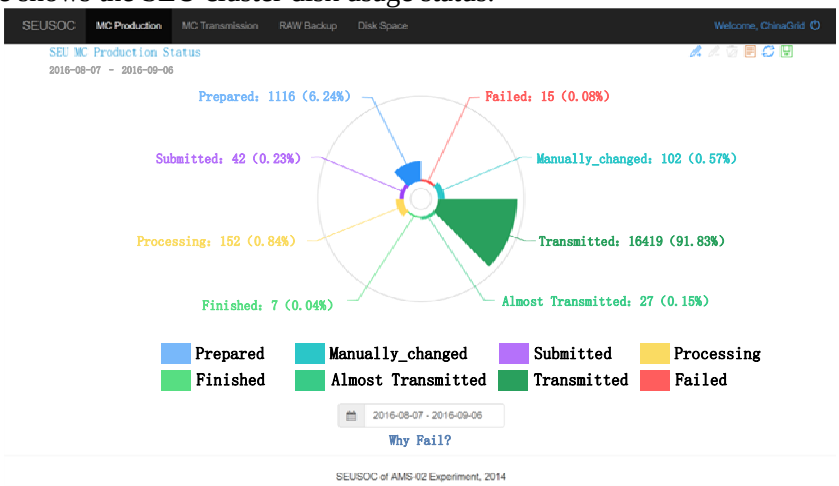

Figure 7. SEU SOC's MC production monitoring website.

5. Conclusions

An automated Monte Carlo production system with optimized data transmission strategy is developed and deployed at SEU SOC. In particular, by adopting the optimized transmission strategy proposed in this paper, the data transmission speed in LFN between SEU and CERN is ~2 times faster than before. This automated Monte Carlo production system allowed reliable running of the AMS Monte Carlo production at SEU SOC during over 4 years of operations of AMS on ISS, resulting in more than 1 PB of simulated events available for the AMS Collaboration. As a result, SEU SOC contributed a total of 30% of the CPU time in AMS Monte Carlo simulations and ranked 1st among all remote AMS Monte Carlo centers around the world.

6. Acknowledgements

This work is supported by National Natural Science Foundation of China under Grants No.61572129, No.61502097, No. 61632008, No.61320106007, No.61370207, National High-tech R&D Program of China (863 Program) under Grants No.2013AA013503, International S&T Cooperation Program of China No. 2015DFA10490, Jiangsu research prospective joint research project under Grants No.BY2013073-01 , Jiangsu Provincial Key Laboratory of Network and Information Security under Grants No.BM2003201, Key Laboratory of Computer Network and Information Integration of Ministry of Education of China under Grants No.93K-9, and partially supported by Collaborative Innovation Center of Novel Software Technology and Industrialization and Collaborative Innovation Center of Wireless Communications Technology.

References

[1] Kounine A 2012 The Alpha Magnetic Spectrometer on the International Space Station, *Int. J. Mod. Phys. E* **21** 1230005
[2] Aguilar M et al. 2013 First Result from the Alpha Magnetic Spectrometer on the International Space Station : Precision Measurement of the Positron Fraction in Primary Cosmic Rays of 0.5-350 GeV *Phys. Rev. Lett* **110** 141102
[3] Choutko V, Egorov A, Eline A and Shan B 2015 Computing Strategy of the AMS Experiment *Journal of Physics: Conference Series* **664** 032029
[4] Roesler S, Engel R and Ranft J 2001 Advanced Monte Carlo for Radiation Physics *Particle Transport Simulation and Applications* 1033-1038
[5] Allison J, Amako K, Apostolakis J, Araujo H, Dubois P A, Asai M, Barrand G, Capra R, Chauvie S, Chytracek R et al. 2006 Geant4 Developments and Applications *IEEE Transactions on Nuclear Science* **53**(1) 270-278
[6] Antcheva I, Ballintijn M, Bellenot B, Biskup M, Brun R, Buncic N, Canal P, Casadei D, Couet O, Fine V et al. 2011 ROOT-A C++ framework for petabyte data storage, statistical analysis and visualization *Computer Physics Communications* **182**(6) 1384-1385
[7] Schwickerath, Ulrich, and Lefebure V 2008 Usage of LSF for batch farms at CERN *Journal of Physics: Conference Series* **119**(4)
[8] Lxplus: http://information-technology.web.cern.ch/services/lxplus-service
[9] Lameter C. "lftp—Sophisticated ftp program." Internet Document: LFTP Manpage, http://www. dca. fee. unicamp. br/cgi-bin/man2ntml/n/net/man1/1ftp, 2001
[10] XRDCP: http://xrootd.org/doc/man/xrdcp.1.html
[11] Cainin C and Firrincieli R 2004 TCP Hybla: a TCP enhancement for heterogeneous networks *Int. J. Satell. Commun. Network* **22** 547–566
[12] EOSFUSE: https://cern.service-now.com/service-portal/article.do?n=KB0004616