PhEDEx high-throughput data transfer management system

J. Rehn, CERN, Switzerland; T. Barrass, Bristol University, UK;
D. Bonacorsi, INFN–CNAF, Italy; J. Hernandez, CIEMAT, Spain;
I. Semeniouk, IN2P3, France; L. Tuura, Northeastern University, USA
Y. Wu, FNAL, USA

Abstract

Distributed data management at LHC scales is a staggering task, accompanied by equally challenging practical management issues with storage systems and wide-area networks. CMS data transfer management system, PhEDEx, is designed to handle this task with minimum operator effort, automating the workflows from large scale distribution of HEP experiment datasets down to reliable and scalable transfers of individual files over frequently unreliable infrastructure. Over the last year PhEDEx has matured to the point of handling virtually all CMS production data transfers. CMS pushes equally its own components to perform and the heavy investment into peer projects at all levels, from technical details to grid standards to worldwide projects, to ensure the end-to-end service is of sufficient quality.

We present the throughput and service quality we have reached in the current daily 24/7 production work, the steps taken in LCG service challenges for the next generation transfer service, and the resulting changes in performance. We also report results from our scalability stress tests on PhEDEx alone. We offer an analysis of transfer-related problems we have encountered and how they have been affecting CMS data management.

INTRODUCTION

PhEDEx[1] is a data transfer management system designed to handle large scale data transfers for the **C**ompact **M**uon **S**olenoid (CMS)[2] **H**igh **E**nergy **P**hysics (HEP) experiment.

PhEDEx meets CMS' requirements for large scale distribution of data by managing a blend of traditional HEP distribution infrastructure with new grid[3] and peer–to–peer[4][5] replication tools.

Historically HEP experiments have relied on manpower-intensive techniques for managing such tasks as:

- Ensuring data safety
- Large–scale data replication
- Tape migration/stage of data

PhEDEx provides a scalable infrastructure for managing these operations by automating many low level activities. This allows a system manager to focus on handling data at high level by managing customizable logical sets of files rather than individual files.

PhEDEx does not place any constraints on the choice of grid or other distribution tools. It harmonizes perfectly with the different existing grid flavours and currently couples resources from **L**HC **C**omputing **G**rid LCG[6], **O**pen **S**ience **G**rid OSG[7] and NorduGrid[8].

PHEDEX WORK FLOW

Traditionally data management and data placement are very manpower intensive operations. Each participating site in a distribution network has to perform and supervise typical transfer processes, like those in Fig. 1. Such operations are unscalable when dealing with $\mathcal{O}$(PB) of data, millions of files and hundreds of sites.

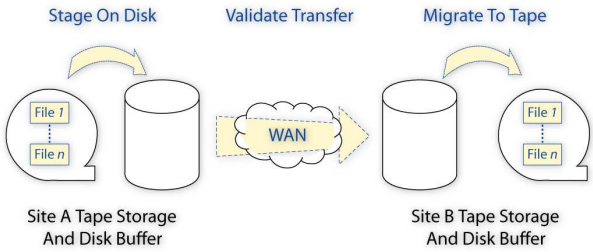

Figure 1: Traditional data replication workflow

PhEDEx automates such workflows, leading to a significant reduction in service manpower requirements. It comprises a set of agents, each undertaking a unique task in a reliable way. Every participating site runs a suite of agents dealing with tasks like file replication, routing decisions, tape migrations and file pre–staging. This partitioning of functionality in subsets of simple tasks is one of the key elements that makes PhEDEx robust and xreliable. Information exchange between the individual agents is performed via a central *blackboard* system, which is realized as a **T**ransfer **M**anagement **D**ata **B**ase (TMDB) running on a multi–server Oracle platform.

PhEDEx was designed to use any available grid infrastructure and to provide data mangement funtionality suiting CMS' needs, although much is now suitably generic that it could be taken up by a variety of scientific and other projects. This goal was achieved by using a layered design[9] (see Fig. 2) where each layer adds functionality leading to a reliable service with low operational effort. Typically one person per participating site is needed to run operations at the level of approximately 0.2 **F**ull **T**ime **E**quivalent (FTE). However, our current experi-

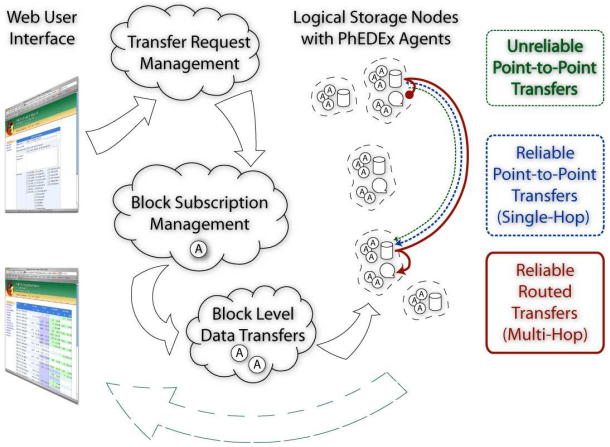

Figure 2: Layered design of PhEDEx. Each site runs a set of specialized agents (denoted by A) running "close" to the site's storage systems. A web front end provides easy administration and monitoring.

ence shows, that problems with the underlying fabric and **S**torage **R**esource **M**anagement (SRM)[10] interoperations can take up to $\mathcal{O}(1)$ FTE.

PhEDEx provides high–level replication using the concept of data blocks instead of dealing with individual files, following the CMS data management concept of datasets and owners.

The grid infrastructure offers several point–to–point file replication tools. Experience shows, that such tools are not typically reliable enough to manage data placement at PetaByte scale. PhEDEx provides a layer of components which ensure that point–to–point transfers are reliable. This is achieved by storing filesize and checksum informations in the TMDB, which can be used to independently verify the internal checks made by the grid replication tools.

SYSTEM PERFORMANCE IN PRODUCTION

PhEDEx has had to cope with an ever growing demand as new sites have joined the distribution network. This demand was met using an appropriate design, aiming on scalability, high throughput, and low operation effort, which has proved capable of handling the load imposed by existing production requests.

CMS has been using PhEDEx since mid-2004 to ship simulated data from the production centres to sites suporting user analyses. Since then a total of 230 TB of data has been successfully replicated among the participating sites. Currently the PhEDEx disitribution network houses a total of 35 sites, of which 8 are regional centres (T1), which in term serve a set of 27 local centres (T2) or smaller sites.

Fig. 3 illustrates the cumulative amount of data moved using the PhEDEx production instance, between Tier 1 and Tier 2 centres, where the Tier 2s typically downloaded files

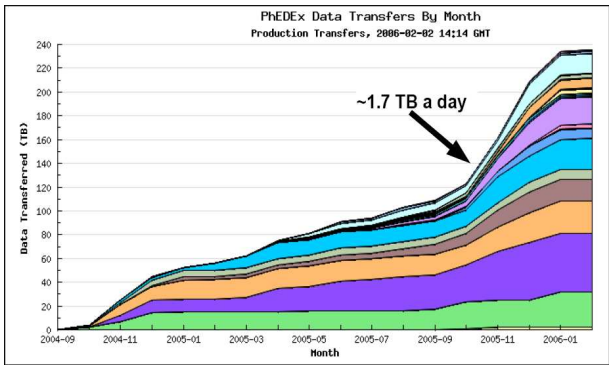

Figure 3: Total amount of data transfered to all participating sites using the production instance of PhEDEx. This plot includes regional T1 centres as well as T2 sites.

from the closest of the Tier 1s. From the slope a maximum performance of 1.7 TB per day can be deduced. Note that this is a typical value sustainable between currently deployed production systems, not testbeds, incorporating disk and tape resources and real local file catalogue servers, all of which are contended by other users.

The main limitations experienced in the production instance of PhEDEx were related to mass storage access caused by heavy load on tape stagein systems, network related issues like firewall misconfigurations, or unresponsiveness of SRM services. It also turned out to be difficult to trace the different error conditions to their source and to correctly identify problems, since parts of the grid middleware does not properly propagate low level error messages.

SYSTEM PERFORMANCE IN SC3

In order to test the system in an environment similar to what we expect to be dealing with during LHC start–up, CMS also participated in the LCG **S**ervice **C**hallenge 3 (SC3)[11] using PhEDEx as its data transfer and data placement tool. Here new storage systems and dedicated high bandwidth network links were available for stress–tests. The first phase of SC3 (also known as *the throughput phase*) intended mainly to determine how to optimize use of available bandwidth and to exercise data import and export at high rate at the individual centres.

During the first phase of SC3, PhEDEx replicated an integrated total of 290 TB of data between the participating computing centres (see Fig. 4). The transfer performance peaked at about 17 TB a day across all sites and the system proved to be capable of handling the increased data volume without problems.

The target replication performance for this phase of SC3 was a sustained 12 TB a day per T1 centre. This goal was clearly not achieved; and the main reasons for missing that goal are discussed below. Note, that this is an aggregated rate of successfully completed transfers, not an average of low level through-put, which might take into account some unsuccessful transfers as well.

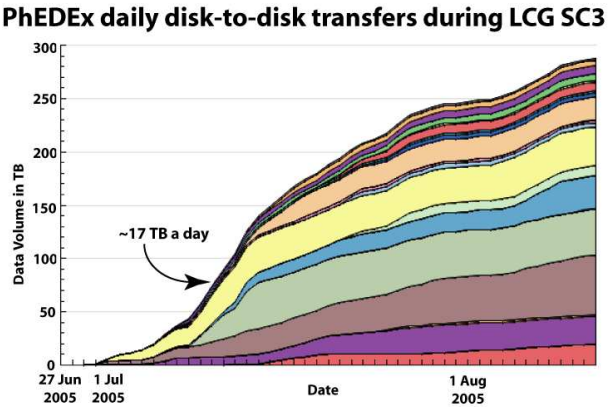

Figure 4: Total amount of data transfered to all participating sites during SC3 throughput phase using PhEDEx. This plot includes regional T1 centres as well as local T2 sites.

The recent service challenge provided a nice opportunity to not only study performance and bottlenecks, but also to study a variety of error conditions and their effect on the PhEDEx system. Data consistency remains of paramount importance for scientific experiments and so the problems associated with making transfers robust and reliable must be addressed.

During the CMS participation in SC3, only about 50 % of the transfer attemps succeeded using the underlying grid replication tools. This gave reason to analyze the source of those problems (see Fig. 5).

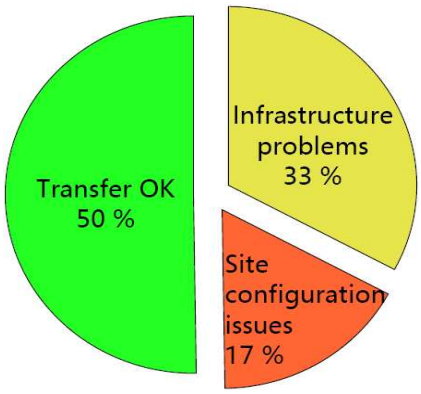

Figure 5: Classification of transfer level problems observed during SC3 data replication. PhEDEx prevented data corruption successfully in all cases.

Most of the issues (33 % of the cases) were related to storage or network infrastructure, caused by:

- instabilities of the mass storage systems
- interoperability problems between SRM instances
- network outages and site network infrastructural problems

The second set of problems (17 % of the cases) were related to site local issues or misconfigurations, of which the most prominent were:

- firewall issues or reconfigurations
- local interventions at sites
- hardware failures
- difficulty of coordination of distributed problem solving

The issues experienced during the CMS participation in SC3 were similar to the problems already observed during production operation. In addition startup difficulties with some of the newly deployed systems showed up, which typically don't appear in a production environment using reliable and well tested hardware and software.

It should be emphasised at this point that PhEDEx only provides a layer that makes the management of large scale transfers manageable. It doesn't aim to provide the underlying fabric or tools, and relies on other projects for these. As the pie chart shows, most problems with PhEDEx transfers are to do with the underlying fabric, either due to direct problems, or due to complexity of site configuration. Many of these problems are only solved by manual intervention at the fabric level. Such intervention increases the manpower load required to run large-scale transfer operations, and has led to suggestions that PhEDEx is unscalable. However, it can be shown that PhEDEx itself is highly scalable; the results of such a study follow.

SCALABILITY TESTS

To ensure that PhEDEx scales elegantly to meet the demands of LHC startup we undertook a series of scalability tests aiming at identifying processes that limit PhEDEx' performance.

A profiling of the PhEDEx workflow at file replication level was performed by analyzing the timing informations in the log files at one particular site. At this level of *reliable single hop point to point transfers* (see Fig. 2), the PhEDEx workflow comprises 4 individual steps:

- cleaning up remnants of previous (unsuccessful) transfer trials
- performing the WAN data replication using the underlying grid fabric
- validating data replication by comparing file sizes and eventually checksums
- publish the new file to a local file catalogue, on completition of a successful transfer

The processing time for each of those steps was measured exploiting the logs of transfers performed during the SC3 throughput phase at one particular Tier 1 site. It reveals that most of the time is spent in actually transfering the files over the WAN (see Tab. 1). Since only one site was

step in workflow	average time spent per file
remove old replica	9.22 ± 0.78 s
transfer over WAN	86.93 ± 10.08 s
validate transfered file	1.72 ± 0.40 s
publish to file catalogue	16.15 ± 2.91 s

Table 1: Profile of PhEDEx workflow using an average file size of 120 MB. Listed are the average procesing time per task and the corresponding statistical fluctuation. Clearly the transfers over the WAN dominate the total replication time per file, although the administrative operations required to verify and publish the data on successful transfer represent a significant portion of the total time – about 32%

taken into account, the result might systematically depend on the individual choice of service and deployment.

In order to identify bottle necks in the workflow of the system, idealized transfers requiring zero time were simulated by replacing the transfer command with `/bin/true`.

Since a realistic setup comprises several centres exchanging data with each other, the test environment simulated 5 Tier centres, each centre connected with the four others. Data subscriptions were organized such, that every centre started with a unique set of files, which were distributed to the other simulated sites leading to a maximum of parallel transfers and hence to a maximum of load on the workflow and the TMDB. These stress tests were repeated several times in order to accumulate sufficient statistics and to monitor the file replication performance over a period of 12 days (see Fig. 6).

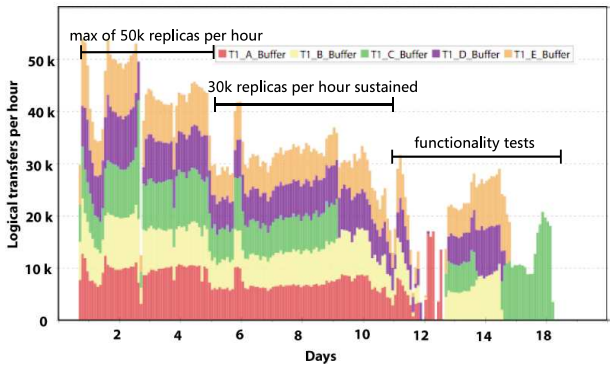

Figure 6: Scalability test using 5 simulated centres connected to each other and logical WAN replications.

PhEDEx achieves a peak replication performance of about 50k logical replications per hour. A sustained level of 30k file transfers per hour was achieved over a period of several days. Taking into account the current average file size of about 2 GB and the CMS target transfer rate of $\mathcal{O}(1$ TB) per day total, this translates into an expected replication rate of $\mathcal{O}(1000)$ successful file replications per day. Hence, the PhEDEx workflow seems to be adequatly prepared to meet the CMS requirements.

CONCLUSIONS AND OUTLOOK

PhEDEx provides scalable and reliable data replication using the underlying grid replication tools as well as data management capabilities to steer and monitor the data distribution at the level of logical file blocks.

PhEDEx has been used to distribute simulated LHC data among collaborating CMS analysis sites for one and a half years. During that period, PhEDEx was constantly improved and new features were added to suit the requirements imposed by High Energy Physics (HEP). It became apparent that the underlying grid transfer infrastructure and the interfaced storage systems currently cannot provide the needed reliability and responsiveness, in particular with respect to the increased requirements during LHC startup.

This impression was confirmed by the results obtained during the CMS participation in SC3, where PhEDEx was able deliver a reliable service. The main goal of achieving a high transfer speed of 12 TB per Tier 1 centre at a sustained level could not be achieved for various reasons related to the reliability of the underlying fabric and lack of interoperability.

PhEDEx scales well, logically, and yet day-to-day production operations at large scale are still not manageable. This unscalability is generated in the underlying fabric, by the tools and the storage systems that are used. The complexity of configuration at each site, which can vary greatly even between sites deploying the same system, and a raft of interoperability problems between implementations of supposedly similar systems mean that manpower is typically required at each site to monitor frequent, often recurring problems.

This weakness in the underlying fabric threatens the ability of all the LHC experiments to undertake large-scale transfers and needs to be addressed with all speed.

REFERENCES

[1] Barrass et al, “Software agents in data and workflow management”, Computing in High Energy Physics (CHEP04), Interlaken, 2004

[2] CMS Collaboration, “The Compact Muon Solenoid Computing Technical Proposal”, CERN/LHCC 1996-045 (1996)

[3] Foster and Kesselman, “The Grid”, Morgan Kaufmann, 1999

[4] “Bittorrent”, http://www.bittorrent.com

[5] “GnuTella”, http://www.gnutella.com

[6] “The LHC Computing Grid”, http://lcg.web.cern.ch

[7] “The Open Science Grid”, http://osg.grid.iu.edu

[8] “NorduGrid”, http://www.nordugrid.org

[9] Barrass et al, “Techniques in high-throughput, reliable transfer systems: break–down of PhEDEx design”, Computing in High Energy Physics (CHEP06), Mumbai, India, 2006

[10] Shoshani et al, “Storage Resource Managers: Middleware Components for Grid Storage”, MSS, 2002

[11] Tuura et al, “CMS experience in LCG SC3”, Computing in High Energy Physics (CHEP06), Mumbai, India, 2006