Journal of Physics: Conference Series **664** (2015) 042029 doi:10.1088/1742-6596/664/4/042029

Data preservation for the HERA experiments at DESY using dCache technology

Dirk Krücker, Karsten Schwank, Patrick Fuhrmann, Birgit Lewendel, David M. South

DESY, Notkestraße 85, 22607 Hamburg, Germany

E-mail: dirk.kruecker@desy.de

Abstract. We report on the status of the data preservation project at DESY for the HERA experiments and present the latest design of the storage which is a central element for bit-preservation. The HEP experiments based at the HERA accelerator at DESY collected large and unique datasets during the period from 1992 to 2007. As part of the ongoing DPHEP data preservation efforts at DESY, these datasets must be transferred into storage systems that keep the data available for ongoing studies and guarantee safe long term access. To achieve a high level of reliability, we use the *dCache* distributed storage solution and make use of its replication capabilities and tape interfaces. We also investigate a recently introduced *Small File Service* that allows for the fully automatic creation of tape friendly container files.

1. Introduction

The HERA accelerator has a circumference of 6.3 km and performed hadron-electron collisions from 1991 to 2007 with a centre-of-mass energy of up to 318 GeV. It was the largest particle accelerator at DESY and served the four experiments H1, ZEUS, HERMES and HERA-B. Up to now, and also for the foreseeable future, there is no other accelerator built to explore electron-proton interactions at comparable high energies. The datasets produced by the experiments are undoubtedly of unique value and must be considered for long-term data preservation.

An important institutional change took place in 2015. After many years of successful work, the HERA experiments are no longer independent projects at DESY. The collaborations still exist but without dedicated manpower for the old experiments. As a consequence, the responsibility for keeping the HERA legacy safe and alive is transferred to DESY. While access to the data is still controlled by the collaborations, the responsibility for preserving data and documentation is or will be shifted mainly to the IT department and the DESY library. This implies a transition from experiment specific solutions for storage, computing, web services and other forms of documentation to institutional solutions.

The challenges of long-term data preservation had been extensively studied at DESY as part of the *ICFA study group on Data Preservation and Long Term Analysis in High Energy Physics* [1]. Data preservation in our context has three major aspects: documentation, software- and bit-preservation. During the past years great care has been taken to collect all available digital and non-digital documentation accumulated by the experiments over many years of running. Non-digital documentation has been collected and catalogued, partly digitised, and safely stored in the DESY library archive. Many forms of digital documentation, such as electronic logbooks, online


Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

monitoring and all kinds of web-based communications must be transferred into self-contained forms. Web sites has been translated into static, script-free versions without references to the different storage systems that had been used over the years, such that they can run on virtual machines with self-contained storage.

The data alone is of no use without the capability to analyse the preserved data and to perform or replicate physical analyses. Here the collaborations follow different strategies. While ZEUS decided to transfer all data into a pure ROOT [2] format, H1 considers the availability of the complete custom-made H1 software as indispensable.

Finally, the aspect of bit-preservation is concerned with the data itself. Data in this context includes the raw-data taken by the experiments, as well as the results of extensive Monte Carlo simulations performed over many years. In this paper, we focus on the aspect of bit-preservation and describe the structure of the new *dCache*-based [3] store for the HERA data that is currently built up at DESY.

2. Data and expected use case

The HERA legacy dataset consist of the raw data of the four experiments and large samples of corresponding Monte Carlo simulations, which had been produced until early 2015 and which are typically ten times larger by volume than the raw data. In a first step, the preservation dataset has to be defined by the experiments, a process which is almost finalised. In a second step, the data has to be copied to a new storage setup, due to the special situation that the old *dCache* instance hosting the HERA data will be decommissioned soon.

For the new setup, we had to take into account that the HERA experiments are still active. There are ongoing analyses, not yet finished PhD theses and new ideas how to analyse the data that may still come up. We therefore make provisions to meet these requirements by a two-folded strategy: the long-term availability of the data must be assured and at the same time easy access to the data must be guaranteed now and for the coming years.

The new storage system includes an archive part where the data exists in two tape copies and an online part where data is available on read-only disk pools. The online part consists of a large subset of the data that has been selected by the experiments. Consequently, the HERA legacy data exists at least on two physically separate media and, depending on the technological developments, the archive may be copied onto whatever storage media will become fashionable in the future.

3. Storage structure and status of bit-preservation

The demanding aspect of the legacy data is not only its size of about 1.1 PiB but also the large number of about 4 million individual files with a broad range of file sizes from a few KB to a few hundreds of GB. To achieve a high level of reliability, we use the dCache distributed storage solution and make use of its replication capabilities and tape interfaces. The data is manually catalogued and packaged into plain tar archives. For each tar file two redundant copies are saved on two different generations of tape cartridges. Presently, we use LTO-4 and LTO-6 cartridges [4] for this purpose. Data from the tape archive is not generally accessible for the user. If access is required, the data has to be manually restored by the *dCache*-operation team at DESY.

For data that is still in active use we provide a second, so-called online part with a size of 700 TiB provided by 47 disk pools. It allows read-only access through *dCache* doors and is mounted on our batch system and the different workgroup servers used by the experiments.

Our main effort during the last months was related to bit-preservation. The legacy datasets had been defined up to a level of about 98% and the new online store had been filled up to approximately 92%. Figure 1 shows the distribution of data between the different HERA experiments and the status of the long-term tape archive. At the time of writing, 73% of the

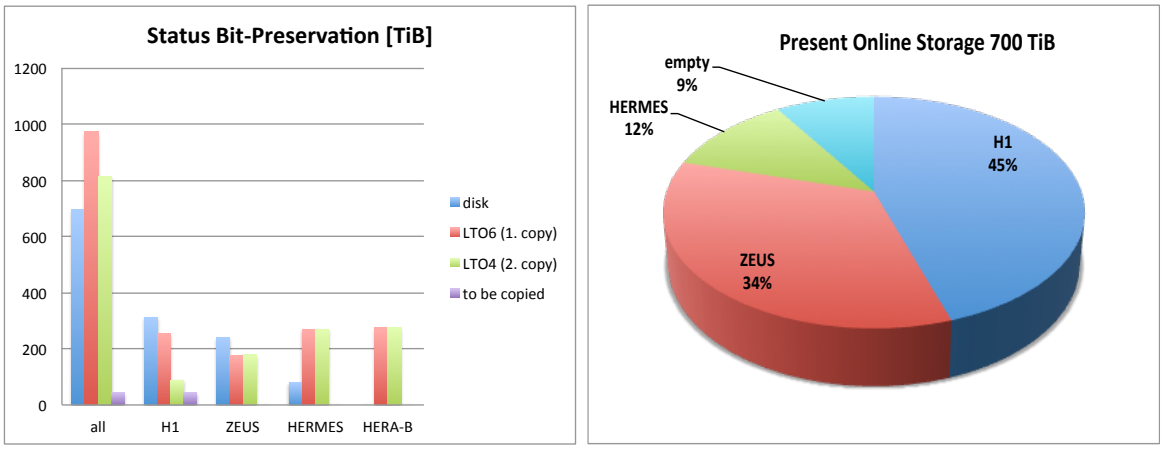

Figure 1. Status of the HERA data preservation as of early 2015 separated for the different storage media and HERA experiments (left). Distribution of disk space between the HERA experiments for the online store alone (right).

expected data for the tape archive has been written to tape in two redundant tape copies and we expect to complete the archive within the next 6 months.

4. Design study for an extended HERA data storage

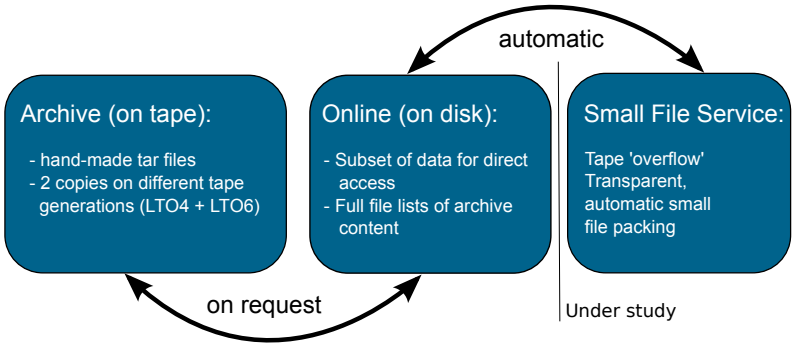

Figure 2. Structure of the HERA data storage.

We have more data in the tape archive than in the online storage and, as mentioned, the user cannot access the tape archive directly. This is by design and follows the economical reasoning not to waste a high performance and high availability storage system for data that may not be accessed for years. On the other hand, the decision which data belongs to which category is not always obvious and we are faced with two conflicting views: the users would like to have access to as much data as possible without having to contact the *dCache* operations team, while from the data management point of view keeping all data all the time on disk is expensive. In addition, requests for locating and unpacking large numbers of small files puts a high additional workload on the operators. It would be convenient if we could provide automatic access for the users to all data without manual intervention.

To meet these requirements we have included a recently developed service, called the *Small File Service* [5], which allows the automatic creation of tape friendly container files. From the

user's point of view, this is done in a fully transparent way in terms of container file creation and access to the data. The data would be accessed through the same mount point and would be visible to the user within the directory structure of the online space. The service is implemented in such a way that it can be attached to any recent *dCache* version and provides easy direct access to all data without manual intervention. The service bundles files into container files in the background and unpacks them again on data requests. The bundling allows improved average data rates for reading and writing to tape without introducing an additional point of failure. All data necessary to restore the individual files is kept in the *dCache* instance, while the service is operated largely independent of the *dCache* instance.

4.1. Performance

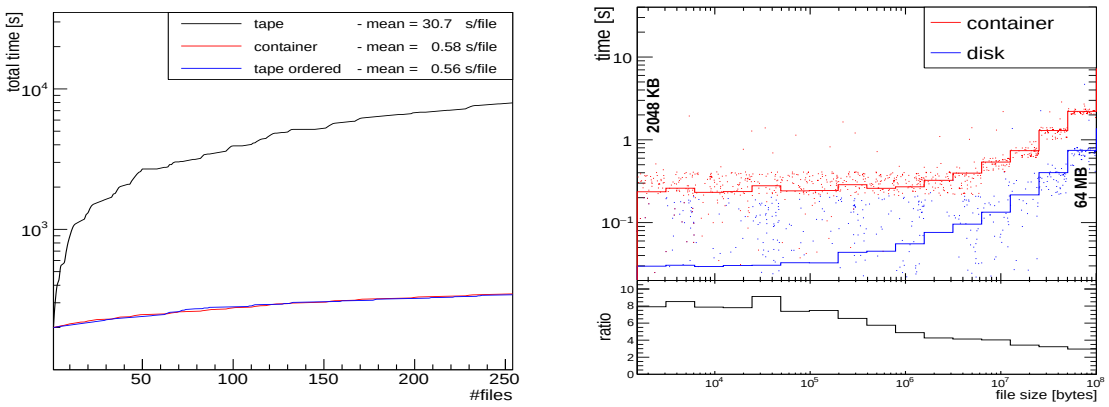

Figure 3. Performance of the *Small File Service*. Accumulated access time for random/ordered file access from tape compared to random access from container files (left) and dependence on file size (right).

The *Small File Service* is developed with the assumption that typically many files from one subset of data are requested by the user at the same time. A subset may be defined for example by a common directory or file name pattern and could be bundled into a common container file. Reading many files back from tape at once upon a single file request will be more efficient in such cases. The first file request triggers the reading of the container file back from tape. Subsequent requests for files belonging to the same container can then be handled by the *Small File Service* without tape access. All unpacked files are read directly from disk and the *Small File Service* does not introduce any additional time penalty. Reading the complete container file from tape adds a certain time overhead but restoring data from tape is in any case limited by the speed of the tape robot. The packaging of files into the container is handled by the service asynchronously, thus it does not influence the write performance into *dCache*.

We have carried out several studies to check the performance of the *Small File Service*. The results are shown in Figure 3. For the first test we use a real example with several hundred log files from the ZEUS experiment. The file size varies between a few hundred KB and several dozen MB. The files were accessed in random order and the accumulated time over 250 files is shown in Figure 3 (left). The time measurement does not include the time the tape robot needs to load the tape cartridge. The advantage of the container files is clearly seen by the difference of the total time of about 2.5 h compared to 10 min with the *Small File Service*. The slow reading from tape is related to the random access of files on the tape. The slow tape positioning can be avoided if the files are read back in the same sequence as they have been written to tape.

This best case is shown in Figure 3 as 'tape ordered'. On average the *Small File Service* allows a read performance similar to this best case.

For a second test we used pseudo data files of logarithmically increasing size. Figure 3 (right) shows the time for reading single files of different size directly from *dCache* where the files reside on disk (blue line), compared to reading files through the containers (red line) created by the service and which for this test are already on disk. The small plot below shows the ratio of the two measurements. The lines show the average over 100 read operations, the individual read operations are shown as dots. The spread is related to the different load of *dCache* and network. We can see that reading a file from a container introduces a time overhead of about 200 ms and, as the files grow bigger, the overhead gets more and more negligible. More technical details on the *Small File Service* can be found elsewhere in these proceedings [5].

5. Summary

The uniqueness of the HERA dataset is clearly acknowledged by all involved parties and extensive preparations have been taken in recent years for the long-term preservation of the HERA legacy. As dedicated manpower for the old experiments comes to an end, the responsibility for data preservation has been shifted to DESY as an institution. Bit-preservation is a central element to these efforts. The HERA data archive is defined and built up to a large extent and we expect this work to be finalised within the next 6 months. Our efforts in building-up the archive are not only of preservation purposes but we additionally want to keep the data easily accessible for current and future users.

The HERA data store consists of a tape archive of about two times 1.1 PiB where the data is kept in two redundant tape copies and an online store where a selected part of the data is kept on disk pools with a total size of 700 TiB. In the present setup less frequently used datasets can only be restored by the *dCache* operators. To allow direct access to all data the *Small File Service* is under discussion and the first performance tests for our use case are promising. The *Small File Service* would allow transparent swapping of files between disk and tape, while providing efficient data rates even for random access to small and medium sized files.

References

[1] Akopov Z *et al.* 2012 Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics (*preprint* hep-ex/1205.4667)
[2] `http://root.cern.ch`
[3] `http://www.dcache.org`
[4] `http://www.lto.org/`
[5] Schwank K *et al.* 2015 Transparent handling of small files with *dCache*, proceedings CHEP-2015, Okinawa