IJDC | General Article

Design and Implementation of the First Generic Archive Storage Service for Research Data in Germany

Felix Bach
Steinbuch Centre for Computing (SCC)
Karlsruhe Institute of Technology (KIT)

Björn Schembera
High-Performance Computing Center (HLRS)
Universtity of Stuttgart

Jos van Wezel
Steinbuch Centre for Computing (SCC)
Karlsruhe Institute of Technology (KIT)

Abstract

Research data as the true valuable good in science must be saved and subsequently kept findable, accessible and reusable for reasons of proper scientific conduct for a time span of several years. However, managing long-term storage of research data is a burden for institutes and researchers. Because of the sheer size and the required retention time apt storage providers are hard to find.

Aiming to solve this puzzle, the bwDataArchive project started development of a long-term research data archive that is reliable, cost effective and able store multiple petabytes of data. The hardware consists of data storage on magnetic tape, interfaced with disk caches and nodes for data movement and access. On the software side, the High Performance Storage System (HPSS) was chosen for its proven ability to reliably store huge amounts of data. However, the implementation of bwDataArchive is not dependant on HPSS. For authentication the bwDataArchive is integrated into the federated identity management for educational institutions in the State of Baden-Württemberg in Germany.

The archive features data protection by means of a dual copy at two distinct locations on different tape technologies, data accessibility by common storage protocols, data retention assurance for more than ten years, data preservation with checksums, and data management capabilities supported by a flexible directory structure allowing sharing and publication. As of September 2019, the bwDataArchive holds over 9 PB and 90 million files and sees a constant increase in usage and users from many communities.

All authors contributed equally to all sections of the paper.

Received 12 January 2018 ~ *Revision received* 27 September 2019 ~ *Accepted* 22 October 2019

Correspondence should be addressed to Dr. Felix Bach, Steinbuch Centre for Computing (SCC), Karlsruhe Institute of Technology (KIT). Email: felix.bach@kit.edu

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

As predicted in (Hey, Tansley and Tolle, 2009), today's science is largely data-driven. Results, for example from computer-simulations, experiment recordings, surveys or digital reproductions of cultural goods, play a crucial role in scientific progress. The need to preserve data over longer time is inevitable, and prominently backed by national funding agencies, such as the German Research Foundation (DFG) (DFG, 2013). However, in practice researchers are forced to store their data either on facilities with limited data storage capacity in their home institution or at external providers. Storing large amounts of data at the home institution is often just impossible, e.g. because the storage must be cleared to allow other users, whereas external providers cannot be used because of either security and privacy considerations or prohibitively high costs (Tristram et al., 2016).

In High Performance Computing (HPC), huge amounts of data are produced during simulation runs (Schembera et al., 2017). For example, at the High-Performance Computing Center Stuttgart (HLRS), whose main users stem from engineering, computational fluid dynamics, molecular dynamics or climate research, a typical large project generates thousands of files, each up to several gigabytes in size. After successful simulation runs and the evaluation and publication of the results, the data becomes inactive and is either deleted or stored on e.g. a local storage system. Since the data is inactive, but should be available for at least a ten-year period there is a need for a long-term archiving service that can handle a rapidly increasing amount of data and ensures bit-stream preservation. To develop a solution for this requirement the bwDataArchive research project was established to build an inter-operable, state-wide data archive. After successful implementation of the services, selected users of HLRS and of the Karlsruhe Institute of Technology (KIT) were given early access for evaluation and to help improve the service. The service of bwDataArchive has moved to production in November 2016 and has seen a continuous increase in users and usage since.

Project bwDataArchive

The three-year project with the goal to develop a reliable and cost effective infrastructure for long term-data storage started in January 2014 and was funded by the Ministry of Science, Research and the Arts of the State of Baden-Württemberg. The project and early production phase, delivered a long term storage service that is easy to use, widely accessible and in principle can support any scientific community (van Wezel et al., 2015).

Scope and Timeline

At the beginning of the project, hardware and software was procured and installed at the Steinbuch Centre for Computing (SCC), the information technology center of KIT. In a tender, the High Performance Storage System[1] (HPSS) was selected on which the bwDataArchive service was to be implemented. The complete hardware setup comprises several computers, storage units and tape systems. The setup is described in

1 HPSS: http://www.hpss-collaboration.org/

detail in a next chapter. For operation, the bwDataArchive makes extensive use of the service infrastructure of SCC that provides networking, installation, monitoring and user support services.

Archive Storage Software

The project selected HPSS data and tape management system (Watson and Coyne, 1995) as the core of the archive solution. HPSS manages file-based data storage and presents a POSIX compatible file system layer that is addressable via an API. It manages disk as well as tape resources and orchestrates the migration of data between these resources through managed policies. The HPSS system is used by many large supercomputer centers[2] is actively developed by a collaboration, and has a set of features that are beneficial for archival storage. The HPSS license allows access to the data for reading after contract expiration. These qualities are very important for long-term storage of data, since it adds to the reliability, sustainability and therefore trust in the system. The HLRS also employs the HPSS system where it is operated as HSM solution in conjunction with the Lustre parallel file system. Before the project, KIT exclusively used the IBM Spectrum Protect[3] software for retention and archiving of data, which include the data from the GridKa[4] data center. Valuable experience and performance evaluation baselines from both data centers and products could therefore be re-used in the project. HPSS will replace Spectrum Protect for several use cases at KIT.

The feature set of HPSS makes it a good candidate to manage the archive storage. In principle bwDataArchive is independent from HPSS and could have been build with Spectrum Protect or another software. However, the level of complexity of connecting the user interface with the archive storage system functions also depends on the functions offered by archive storage software. For example, the file system interface of HPSS access allowed us to support known easy to use services. As another example, the check sums generated in HPSS can be accessed by the user. With the Spectrum Protect, these features have to be programmed and ultimately, maintained.

Checksums and Data Integrity

A digital archive must insure that its content has not changed over time. The requirement for algorithmically generated checksums was investigated in a feasibility study in the EUDAT2020 project. The joint research activity (JRA) aimed to implement universal check sum support for long term archives. Experts from three European data centers were involved, each operating a different commercial archive software solution. The results of the project, including a product independent checksum system in python is described in (Krauß, Cadolle Bel, Kennedy and Jankowski, 2015) and (Krauß, Jankowski and Kennedy, 2018).

Related to the checksums is the end to end data integrity checking feature of HPSS that follows the standardised T10 PI data integrity specification[5]. For every data object that enters the archive system a checksum is generated that is written on disk and tape

2 Sites using the HPSS system: http://www.hpss-collaboration.org/customersM.shtml
3 Spectrum Protect was previously named TSM.
4 GridKa is the German T1 computer center of the world wide LHC computing grid, dedicated to store and analyze data for the Large Hadron Collider (LHC) at CERN in Geneva, Switzerland http://www.gridka.de
5 T10 PI specification: http://www.t10.org/ftp/t10/document.03/03-224r0.pdf

alongside the data. If data is corrupted while in rest, the storage hardware will detect a checksum error while reading so the system can automatically correct the error by replacing the damaged copy with the second copy.

Commonly practiced in e.g. archives, museums and libraries, the requirement for fixity checking (Barsness et al., 2017), which means reviewing the data by comparing stored and generated checksums, may actually increase the risk of data corruption when executed on tape systems. Because the verification involves retrieving the data from tape, the additional mechanical stress may add to risk of read errors (Cloud providers may not be immune to this effect either since many of them utilise tape storage in the background). Data from physical sciences which is the focus of the bwDataArchive project actually runs into hundreds of terabytes and petabytes, which makes it even prohibitive to read back data at shorter intervals.

Fixity checking on small data sets that live (and stay) on on-line media may be useful as an additional protection layer. For data on tape this is certainly different. From a technical viewpoint and based on the experience from our large storage operations we decided that in order to properly protect the data over the years we rely on the following to ensure data integrity:

- Writing two data copies to separate tape volumes.
- Hardware supporting the T10 PI end to end data integrity checking.
- Migrating all data every five to seven years to new tape media. This can be done as part of a regular hardware refresh to keep up with the technological progress.
- The concurrent use of different tape technologies. If one day it becomes known that one technology has a systematic error, and for example results in a severe data corruption on tapes, only one copy is affected.

Should better or cheaper systems become available it must be possible to migrate the date and continue with a new system. We estimate a migration duration of six to 12 months with the current size of the archive and depending on the available hardware. Alternatively, a new system can work alongside the existing archive. Access to existing data is read-only and new data is written to the new system. The HPSS software allows read-only data access for a limited time after the license acquired for writing expires.

Sizing and Costs

Initial volume and data rate requirements from HLRS were supplemented with the well known requirements of the GridKa T1 center and those from universities and institutions in the state of Baden-Württemberg (Potthoff et al., 2014). Together these constitute an estimated volume of 30 PB and a combined I/O rate of 10 GB/s[6]. These numbers were used to size the HPSS hardware. After deployment and stabilization of the service during which only a selected group of users will be able to store data, existing archives will be transferred to the bwDataArchive service, including the 20 PB of data from the GridKa T1 center.

Research data is kept for reference or further analysis for many years. Therefore project proposals today commonly include a chapter on how they plan to handle data accumulated during the project and after the project is completed (EU, 2016). The data management plan (DMP) is a formal document which includes, among other information on the produced data, an estimate of the costs to store data after the project

6 Estimates for the first year after start of service of the archive in 2016.

finishes (Jensen, 2011). If the costs of an archive are known before the research project starts, funding can be put aside or acquired for that purpose. However, the costs must be known well before the actual use and apart from the stored volume, the costs will change over time.

One of the goals of the bwDataArchive project was to develop a sustainable and dependable costs – and contract-model that will help researchers establish data management plans. The requirements and the payment model were established in cooperation with the Research Data Management team[7], the KIT Library and the RADAR[8] project of the Leibniz Institute for Information Infrastructure in Karlsruhe. Requirements also stem from the Helmholtz Portfolio Program LSDMA (Meyer et al., 2014), in which researchers from diverse communities, e.g. climate, energy, medicine, work jointly with computer experts from KIT and other computer centers on the development of novel data services.

The bwDataArchive service is the first known long-term data storage on a pay-per-use basis for universities and research infrastructures and projects outside the well-known commercial 'cloud' offerings.

Components of the Research Data Archive Service

Data Center

The hardware of the bwDataArchive service is able to handle the large data streams coming from the HLRS HPC center and consists of many components. The components, some of them double for redundancy, listed below are depicted in Figure 1:

- Front-end nodes for user access and transfers
- A cache disk system with servers that buffer the data
- A database node storing technical metadata
- Two physically distinct and geographically separated (ca. 13 km) tape libraries equipped with Oracle T10000d, LTO7 and IBM TS1155 tape drives. Enterprise tape technologies such as Oracle and IBM serves for the first copy due to its better performance, whereas LTO is used as the fallback for the second
- A node to schedule and manage GridFTP third party transfers (bwdahub, details in a later chapter)
- User management and integration of an identity provider (IdP) through the bwIDM[9] infrastructure.
- IP network with 10 and 40 Gbit links.

7 RDM@KIT: http://www.rdm.kit.edu/
8 RADAR project: https://www.radar-projekt.org/display/RD/Home
9 bwIDM federated identity management: https://www.bwidm.de

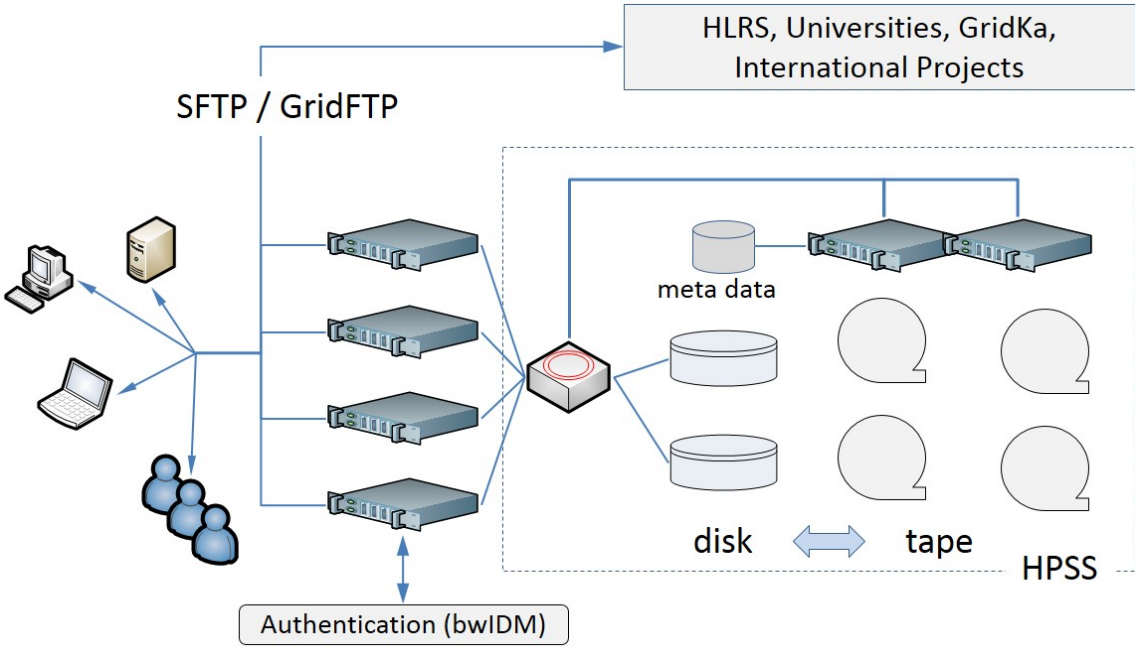

Figure 1. Hardware layout of the bwDataArchive service. The service consists of front-end nodes that can be accessed by the users and where transfers can be scheduled. HPSS acts as the back-end and consists of a metadata service, disk cache for buffering of the data and a tape back-end storage. The infrastructure is based on 10 and 40 Gbit links.

The hardware is complemented with:

- Service and support information, documentation, a help desk portal and support work flows;
- The set of service descriptions and contracting, accounting, and billing work flows.

Service Management

Data in archives will stay there for a long time, if not forever. Over the lifetime of the data, the researcher who stored the data may no longer have a relationship with the institution i.e. KIT, that allowed him to use the archive. So ultimately the data will be orphaned. It is therefore required to keep the identification credentials of the users of the service separate from those used at the institution to grant an extended access (see next chapter for implementation details).

At the same time, the permission to store data must be granted by the institution because costs will incur. Therefore, the service is connected to the federated identity management federation bwIDM, in which all universities and many other institutions in Baden-Württemberg take part (Köhler et al., 2014). bwIDM serves as a trusted source for user identities and, over time, will enable dynamic authorization and role management.

The service documentation and user support builds on the existing services and infrastructures at SCC. Written documentation is provided via the website

bwDataArchive[10] and Wiki pages[11]. The Wiki contains background information about the technology in use, pointers to other information sources, answers to common questions, instructions and best practices regarding the different access protocols, as well as pre-formatted documents, i.e. the service level agreement (SLA) and manuals for downloading. Users may request help via a web-based portal, which is also used for other state wide services, or by sending e-mail to the help desk of the SCC, or directly to the support e-mail list of the service. Because the service is offered to diverse user groups, it was deemed necessary to provide different support entries as well. Behind the scenes the support work flow directs requests to the team responsible for running the service.

The sustainability of an information technology infrastructure in a research context has always been a challenge and implementations thereof were developed with varying success. Most successful have been models where funding is secured up-front and subsequently pledged for on a regular basis. A positive example is the WLCG[12], a cooperative infrastructure that processes and stores data from the LHC at CERN (Adamova, 2013) and more recently the EUDAT CDI[13], a consortium that stems from the EUDAT2020[14] project.

Implementing the pay-per-use scenario for the use of the archive is not the primary expertise of the core team of the bwDataArchive project. In close cooperation with the legal and procurement department at KIT, the project drafted a contract document set that contains the description, pricing, service level agreement, and other components to comprise a legally binding contract for the delivery and use of long-term data storage offered through the bwDataArchive service. Initial customers are universities and public institutions in Baden-Württemberg. In time the service will be used by (customers of) international infrastructure operators such as the aforementioned EUDAT CDI and the future European data infrastructure developed in the EOSC-hub[15] project.

Implementation and Features

The list of features offered by the bwDataArchive service fulfils initial requirements and could be implemented straightforwardly. The available set of user-level features at the start of the service consists of a) data security/redundancy by two data copies at two separate locations b) data accessibility by common storage protocols like SFTP and GridFTP c) long-term preservation by data retention of ten years and more upon request, d) the generation of checksums on input and optional checksum verification on output e) authorization based on IdP/Shibboleth, and finally f) flexibility of usage by a directory layout that allows adoption of new use cases.

Some components that are not commonly found in other storage services have been specially developed or adapted for use in bwDataArchive. This is particularly true for the user management subsystem, the setup of the software stack needed to provide access to the archive via standard protocols, and the particular implementation for high-volume data transfers. Finally, this chapter discusses the directory layout that enables write protection and sharing.

10 bwDataArchive: https://www.rda.kit.edu/
11 bwData wiki page: http://wiki.scc.kit.edu/lsdf/index.php/BwDataArchiv_FAQs
12 WLCG: http://wlcg.web.cern.ch/
13 EUDAT CDI: https://www.eudat.eu/eudat-collaborative-data-infrastructure-cdi
14 EUDAT2020: https://eudat.eu/
15 EOSC Hub: https://www.eosc-hub.eu/

User Access

Access to the archive is provided via the well-known and well understood protocols SFTP and GridFTP (Allcock, Bester et al., 2002). The latter supports high speed parallel data transfer and users of the service can make use of the gtransfer[16] tools which shield users from most of the complexity of GridFTP. GridFTP is the transport protocol of choice for high speed movements of large data. On the other hand, SFTP clients are available for virtually all computing platforms and both, command-line and graphical user interface variants, exist. Its ease of use lowers the threshold for using the archive. Since users of bwDataArchive mostly come from data-intensive science, where CLI is wide-spread, a user interface like SFTP is no impediment at all. Additionally, high level services like the aforementioned generic research data repository RADAR and the central KIT repository KITopen[17] are based on the archive engine. They provide users with helpful functionality, such as an advanced rights and role concept, flexible metadata management and publication of datasets with DOIs, that are all conveniently accessible through a web interface.

Authentication, Authorization and Account Management

Authorized users may interact with the archive to store, retrieve and list data. The bwDataArchive is an infrastructure service that organizations, in particular institutes and universities in Baden-Württemberg, can choose to offer their employees. The authorization procedure and policy of the organization determines who is entitled to use the service. The account management of bwDataArchive authenticates users via the bwIDM federated identity management system, but can also register users independently. The bwIDM authentication system builds on the Shibboleth/SAML standard to forward authentication and authorization information of users from an Identity Provider (IdP), the institution using the service, to a Service Provider (SP), the bwDataArchive service. This implementation of this authentication scheme is fully GDPR compliant.

At the IdP, the archive service user's account is tagged with a dedicated service entitlement for the bwDataArchive service. The entitlement is set by the organization the user is associated with, and thereby authorizes the use of the bwDataArchive service. During the Shibboleth/SAML registration handshake, the account of the user is queried for having the proper entitlement which, when present, is reported by the IdP of the local user organization to the bwDataArchive service. Through this mechanism each site can decide which user is authorized to use the archive service. At the end of the registration process, the user is required to accept the terms of use and the user related data is recorded in a service-specific user database.

Among the usual items, e.g. name and institution, the user may optionally and by consent only register an additional email address and an ORCID[18] that permits contacting the registrant after they are no longer member of the organization. Data associated with a bwDataArchive user may be kept in the archive storage for ten years or more, whereas the person who stored the data may no longer be a member of the organization that granted the use of the storage. For that reason, bwDataArchive remembers users as long as associated data is stored. When a user has left the

16 Gransfer software repository: https://github.com/fr4nk5ch31n3r/gtransfer
17 KITopen stores the 'data' portion of the repository in the archive: https://www.bibliothek.kit.edu/cms/kitopen.php
18 ORCID: https://orcid.org

organization, an event that is tracked by the bwIDM federation, the system no longer allows new data to be stored. Access to existing data remains possible but changes and additions are prohibited.

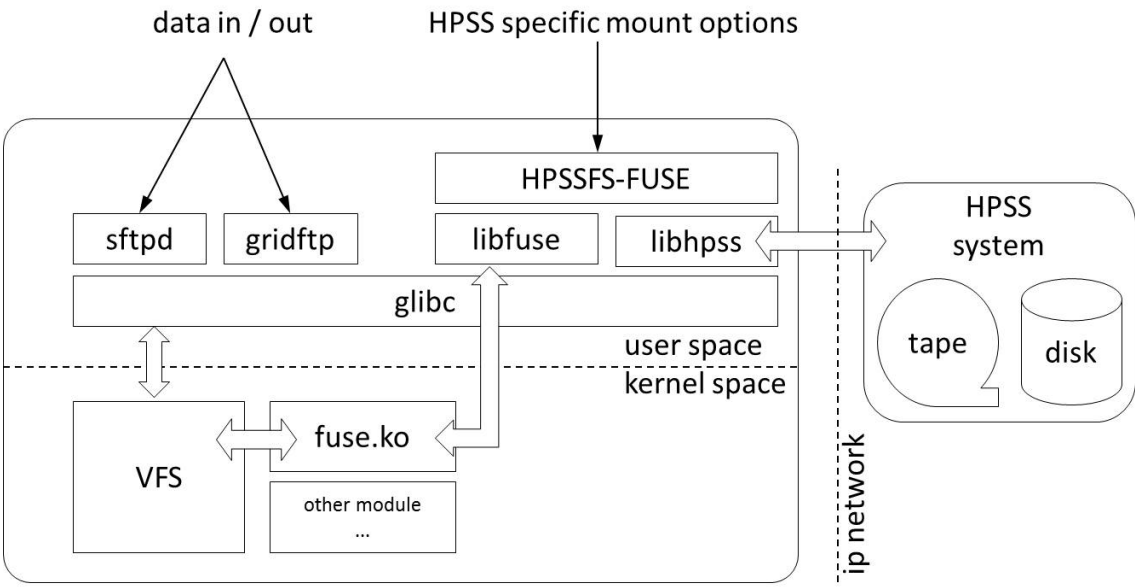

Figure 2. Interaction of storage applications, FUSE components and HPSS.

If data is deleted, the deletion is recorded in the database. This means the data, though still on tape, is no longer accessible to users or even administrators and after a while is overwritten by a tape management process. Additionally, users are able to remove their account and all of their personal data completely, fulfilling the requirements from GDPR. Once started the deletion process is irreversible.

Access to HPSS with FUSE

The HPSS-managed storage is accessed using common protocols through the FUSE[19] file system abstraction library. HPSS-FUSE is a Linux FUSE module that presents HPSS as a file system to applications. Although only a basic set of I/O and metadata operations is used, the implementation allows selecting some HPSS specifics, such as the tape family (to gather data on a particular set of tapes), the class of service (to support files with different file sizes), and the file checksum (to support user verifiable checksums). Additionally, files can be pinned to disk or recovered (undelete) from the HPSS trashcan.

Figure 2 shows a schematic overview of all components involved. Data from SFTP or GridFTP is sent via glibc, the VFS layer, the FUSE kernel module, the libfuse and libhpss to HPSS. Depending on the workload, performance of the FUSE access to storage is comparable with direct storage access. However, because of the relatively large meta-data overhead, writing and reading of large files results in better performance as compared to I/O with small files (Vangoor, Tarasov and Zadok, 2017). Data transport to and from the bwDataArchive service is exclusively done via HPSS-FUSE. Currently, the service is based on five machines (Intel(R) Xeon(R) CPU E5-2630 v2 @2.60GHz, 64 GB) dedicated for data transport to clients. Two of the nodes are reserved for GridFTP

19 FUSE: https://github.com/libfuse/libfuse

transfers, the others for SFTP. More nodes will be added if required since client tools can (and should) open many parallel sessions in order to improve throughput.

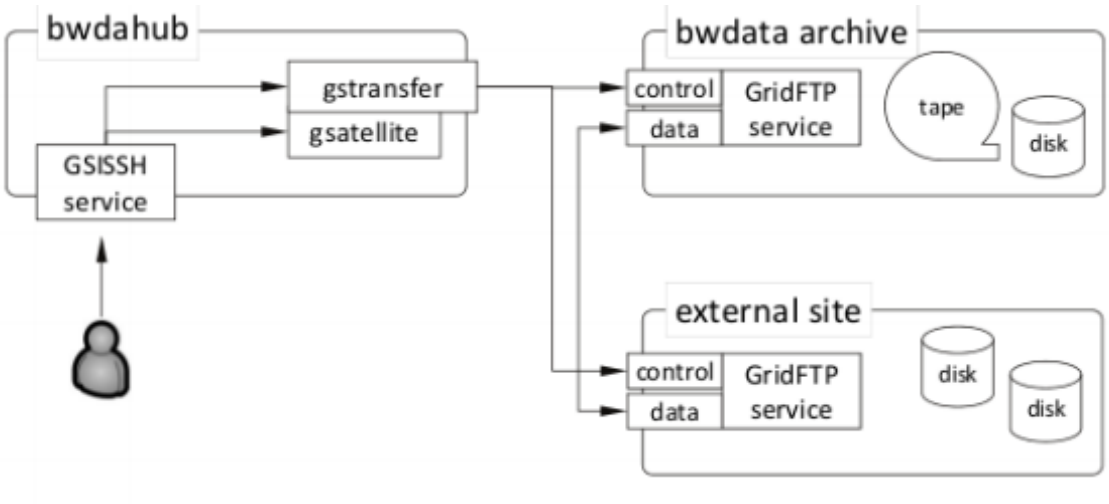

Figure 3. User interaction with the front-end node bwdahub to manage high volume data transfers via the GridFTP protocol from an external site to bwDataArchive and v.v.

Managing and Scheduling High Volume Data Transfers

The archive is typically used to store migrated data from HPC environments, like that at HLRS. A large number of files needs to be moved easily and efficiently i.e. without much user involvement. The bwdahub node acts as a user friendly front-end to schedule high-speed and high-volume data transfers. Users can log in via GSISSH and subsequently trigger data transfers between the bwDataArchive service and external sites, as depicted in Figure 3. All necessary tools are pre-installed on this node, so users don't have to install the software on a computer at their home institution. The front-end tools for transferring data are gtransfer and gsatellite and both rely on the Globus GridFTP client (globus-url-copy), uberftp, as well as tgftp. The user does not directly interact with the latter applications.

For high-throughput data transfers, the GridFTP protocol is used. Besides delivering high data rates, GridFTP can checksum and encrypt data during transfer (Allcock, Liming, Tuecke and Chervenak, 2002). Logically, GridFTP consists of two parts: a protocol interpreter (PI), responsible for managing the transfers; and a data transfer process (DTP) dedicated to transferring the data. To increase the performance of the transfers, it is recommended to use multiple (at least four) DTPs. At KIT, two GridFTP servers are installed and configured with four DTPs each on the two archive front-end machines. Each PI can make use of all eight DTPs, effectively spreading the load over both machines if concurrent or striped transfers are requested. At HLRS, a GridFTP server with a split configuration was set up: the PI (GridFTP front-end node) resides on a different machine than the six DTPs (GridFTP back-end node). The HLRS local HPSS file system is mounted on the GridFTP back-end node. Additionally, the HLRS Lustre file system can be accessed directly via another GridFTP node in order to transfer online simulation data from the supercomputer workspaces to the bwDataArchive.

Directory Structure and Directory Functionalities

The directory structure as well as the filenames in bwDataArchive are the only form of (indirect) metadata in the system since bwDataArchive is intended as a base archival storage and not as a research data repository. However, it is well suited to serve as the sustainable, secure and scalable base for repositories. These services build their (metadata) functionality on top of the directory structures of bwDataArchive.

The archive service uses an elaborate directory layout that aligns with the currently known business cases i.e. registered users and projects. It provides a chrooted user directory with shared access for selected groups, function directories in the near future. The latter cater for special functions or options that are made available for the content within the function directory. Function directories are made visible into the registered users directory as bind-mount with the assistance of an autoFS. Use of autoFS for this functionality is still experimental and promising but has difficulties handling many simultaneously active users.[20]

Function Directories

The internal directory structure of the service is shown in Table 1. Within the archive root (AR), a directory (RU) is created for each registered user, which itself is located inside a first level (PL) directory and a second level (SL) directory. The second level directory layer reduces the number of entries per directory, which improves traversal speed and caching at the client. The system automatically creates the function directory 'private/' in the RU directory where users have read and write permission based on POSIX ACLs. This is the central work space for a user to store files and create directories. It is typical that users ingest data themselves, whereas data managers might ingest data into the shared group directory. Deleting data is possible anytime as long as the user has the correct permission or the data is not yet immutable.

The function concept allows for multiple directories at SL/ that are managed by the system. Within the SL/ a directory is usually a bind-mount that shows the content of a shared directory or a directory which allows storing huge files (i.e. > 1 TB) etc.

File Exchange and Mutual Directory Access

To enable file exchange between users, a group share directory is created for each group of users that require mutual access to files. The concept can be as generic as a group share for all members of a scientific institution, or very specific for a group consisting of only two users. The group access can cover multiple institutions and may also include users that have no institution in the context of bwIDM. The data itself does not change owner and accounting is still done against the owner of the data. The group share is created for the group leader and each group member has this directory mapped into his/her RU directory. The future user management console enables the archive admins to designate a user as group leader and group leaders can add or remove users to and from the group.

20 autoFS: https://wiki.archlinux.org/index.php/Autofs

Table 1. Directory layout and permissions. Example: The user with username 'sne' will store data in: /hpss/bwda/000000/sne/private. Description of the structure:
1) Archive Root directory
2) Project Level directory: Sample projects are RADAR and bwDataDiss.
3) Second Level directory: Improves traversal speed. The name is a 6 digit number.
4) Registered User directory: The name of the registered users directory (RU) is equal to the registered user name. The directory is contained within a changed root environment and not writable for users.
5) Private Archive directory: 'private/' is a function directory created by the system. Users must change to 'private/' to store data.
6) Bind-Mount directory: The directory <group_name>/ is visible inside the RU directory.
7) Shared Group directory: The exported group share so co-workers can access the shared directory.

Type	File System Path	Owner.Group	Permissions
1	AR/	root.archiv	drwx–x–x (711)
2	AR/PL/	root.archiv	drwx–x–x (711)
3	AR/PL/SL/	root.archiv	drwx–x–x (711)
4	AR/PL/SL/RU/	root.archiv	drwxr-xr-r-x (755)
5	AR/PL/SL/RU/private/	uid.gid (gid=uid)	
6	AR/PL/SL/RU/<group_name>		
7	AR/PL/groups/<group_name>	archiv.gid	dr-xrwxr-x (575)

Immutable Data

An important feature of the archive is to prevent data being changed after the archive is locked. There are two main cases where data should be made immutable: removal of writing authorisation and referenced data. Once a registered user is no longer member of a home institution, writing to the archive is made impossible and no additional costs will result. Similarly, when archived data is referenced in a publication, changes to the data are no longer allowed. The user should still have read access with his or her registered account that is independent from his or her home institution account. This function is only available at directory level and is implemented by mounting the RU directory as read-only. Currently this is implemented using autoFS maps. Changes and deletions won't be possible even if the user has access rights thus effectively preventing accidental deletion. This solution is fail-safe because the contents of the RU directory must be made explicitly available and are not visible elsewhere.

Conclusion and Outlook

Hosting the bwDataArchive service at a large institution, such as KIT, assures sustainability and dependability of the service. Both qualities in turn, add to the trust of the data storage service. The bwDataArchive project has built such a service which is already in use by researchers of KIT and is being tested by projects at HLRS and

several universities in Baden-Württemberg. With a focus on the requirements of users that need to move data away from expensive disk storage and in general seek affordable long term storage, the service has incorporated some unique features that resulted in quick acceptance.

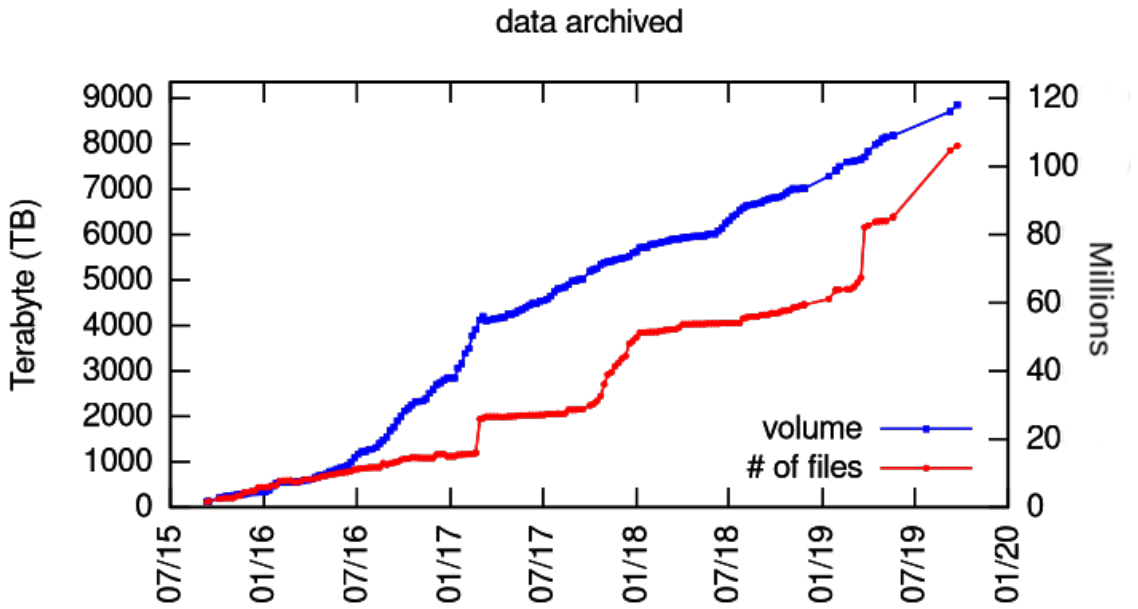

Figure 4. Volume and number of files archived since 2015. The boxed line shows the increase in volume, the dotted line the number of files in the archive.

The infrastructure and features of the installation match the requirements and can scale with growing demands. As of September 2019, the archive holds over 9 PB of data in roughly 90 million files from more than 400 users. Figure 4) shows the trend of both the volume and number of files since 2015.

The amount of data stored does not prohibit migration to a new platform or different technology in the future though moving the data may take some time.

Acknowledgments

The authors thank the Ministry of Science, Research and the Arts of the State of Baden-Württemberg for funding this project. We also wish to thank Bettina Bauer, Martin Beitzinger, Thomas Bönisch, Ahmad Hammad, Dorin-Daniel Lobontu, Iris Mayer, Frank Scheiner and Sara Ramezani for their contributions to various parts of the project and preparing the manuscript.

References

Adamova, D. (2013). Current status of the wlcg data management system, the experience from the three years of data taking and future role of grids for the lhc data processing. In *Proceedings of the 51st international winter meeting on nuclear physics* Bormio, Italy.

Allcock, B., Bester, J., Bresnahan, J., Chervenak, A. L., Foster, I., Kesselman, C., . . . Tuecke, S. (2002). Data management and transfer in high-performance computational grid environments. *Parallel Computing, 28*(5), 749–771.

Allcock, B., Liming, L., Tuecke, S. & Chervenak, A. (2002). Gridftp: A data transfer protocol for the grid. In *Grid forum data working group on gridftp*.

Barsness, S., Collie, A., Gallinger, M., Kussmann, C., Schaefer, S. & Truman, G. (2017). 2017 Fixity Survey Report. Retrieved from https://osf.io/grfpa/download

DFG. (2013). Safeguarding good scientific practice. Retrieved from http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf

EU. (2016). H2020 programme guidelines on fair data management in Horizon 2020. Retrieved from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Hey, T., Tansley, S. & Tolle, K. (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research. Retrieved from https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/

Jensen, U. (2011). Datenmanagementpläne. In Handbuch Forschungsdatenmanagement. Büttner, S., Hobohm, H.-C., Müller, L. Bad Honnef: Bock u. Herchen.

Köhler, J., Labitzke, S., Simon, M., Dussa, T., Nussbaumer, M. & Hartenstein, H. (2014). bwIDM–Federated Access to IT-Based Services at the Universities of the State of Baden-Württemberg. *PIK-Praxis der Informationsverarbeitung und Kommunikation, 37*(1), 15–21.

Krauß, P., Cadolle Bel, M., Kennedy, J.A. & Jankowski, M. (2015). Eudat deliverable 9.4: Final report on efficient integrity checking. Retrieved from https://b2share.eudat.eu/records/61ba4f51655a49e59239c40035aba25a

Krauß, P., Jankowski, M. & Kennedy, J.A. (2018). Eudat deliverable 9.8: Final report on provision of staging support. Retrieved from https://b2share.eudat.eu/records/efe05b93e53849f28d49d80262b665b6

Meyer, J., Hardt, M., Streit, A., & van Wezel, J. (2014). Archival services and technologies for scientific data. *Journal of Physics: Conference Series, 513*(6). IOP Publishing.

Potthoff, J., Wezel, J.V., Razum, M., & Walk, M. (2014). Anforderungen eines nachhaltigen, disziplinübergreifenden Forschungsdaten-Repositoriums. In *7. DFN-Forum-Kommunikationstechnologien*. Gesellschaft für Informatik eV.

Schembera, B. & Bönisch, T. (2017). Challenges of research data management for high performance computing. In *International Conference on Theory and Practice of Digital Libraries* (pp. 140-151). Springer, Cham.

Tristram, F. et al. (2016). Öffentlicher Abschlussbericht von bwFDM Communities– Wissenschaftliches Datenmanagement an den Universitäten Baden Württembergs. Retrieved from http://bwfdm.scc.kit.edu/downloads/Abschlussbericht.pdf

Vangoor, B.K.R., Tarasov, V. & Zadok, E. (2017). To fuse or not to fuse: Performance of user-space file systems. In *Fast* (pp. 59–72).

van Wezel, J., Hammad, A., Krauß, P., Kurze, T., Meyer, J., Potthoff, J. & Schembera, B. (2015). Towards an interoperable data archive. Proceedings of the 2015 EUMETSAT Meteorological Satellite Conference, 2015.

Watson, R.W. & Coyne, R.A. (1995). The parallel i/o architecture of the high-performance storage system (hpss). In *Proceedings of the fourteenth ieee symposium on Mass storage systems, 'storage-at the forefront of information infrastructures'*, (pp. 27–44).