SDS@hd – Scientific Data Storage

Martin Baumann, Vincent Heuveline, Oliver Mattes, Sabine Richling, Sven Siebler
University Computing Centre (URZ)
Heidelberg University
Germany

Abstract. SDS@hd (Scientific Data Storage) is a central storage service for hot large-scale scientific data that can be used by researchers from all universities in Baden-Württemberg. It offers fast and secure file system storage capabilities to individuals or groups, e.g. in the context of cooperative projects. Fast data accesses are possible even in case of a high number of small files. User authentication and authorization are implemented in terms of the federated identity management in Baden-Württemberg allowing researchers to use their existing ID of their home institution transparently for this service. Data protection requirements can be fulfilled by data encryption and secure data transfer protocols. The service is operated by the computing center of Heidelberg University.

I. Introduction

In many fields of research, the capacity of data that is generated in scientific projects is enormous and continuously growing. This is a consequence of technical progress in data generating devices, such as high-throughput microscopes, telescopes and genome sequencers amongst others, and also for fast computer systems such as high-performance compute clusters or cloud systems that can be used e.g. for numerical simulations of complex processes. Suitable IT services for data storage and for data analysis for large data sets are key enabler. Besides the technical prerequisite, research projects have additional requirements. These include adequate group- and access management for co-operational setups and data protection for some projects with sensitive data. In the context of typical data life cycles for research data and of good scientific practise, a subset of the data has to be stored for long time (archive) or be accessible by a group of people or the public. Additional requirements for the collection of meta data and the transfer to corresponding long-term archive and repository services exist.

The second generation hardware of the "Large Scale Data Facility" (LSDF2) addresses this need of data storage in the context of research projects. LSDF2 is financed by the Ministry of Science, Research and the Arts (MWK) and the German Research Foundation (DFG). It is part of the state of Baden-Württemberg's concept for data-intensive services bwDATA [1]. LSDF2 is a storage system and at the same time a joint project by the Computing Center of the Heidelberg University and the Steinbuch Centre for Computing at Karlsruhe Institute of Technology. In the framework of this project, storage capacity is installed on both sites and operated by the two computing centers. The cooperation of the two institutions and also the use of almost identical technologies for the two systems - both on the hardware and software level – create synergies. The data distribution across university boundaries is possible through the respective storage services. This is described in detail in section III.C. In the following, the focus is on the details and use of the LSDF2 located in Heidelberg.

A new scientific data storage service "SDS@hd" [2] has been developed with the intention to improve the value of available data storage resources in the context of research projects. The service is tailored to those phases in the research data life cycle in which frequent accesses are present, i.e. for so-called 'hot data'. The data is stored on the LSDF2 in Heidelberg and can be accessed via local systems (e.g. by using a local high-performance compute cluster) or also from remote systems and even directly from the user's desktop computer. The service is open to all scientists at Baden-Württemberg's universities in the sense of a "Landesdienst".

Subsequently in this publication, the concept and technical specifications of the storage system LSDF2 in Heidelberg will be described in detail. This is followed by a presentation of the storage service SDS@hd including its most important features such as access protocols, management system and data security. After that, the connection to local and remote high-performance compute clusters and the typical usage scenarios are outlined. Finally, support and consulting activities are described.

II. LSDF2 Hardware

SDS@hd is a service which runs on the storage system LSDF2 located at Heidelberg University. The operational concept considers failover events. Therefore the storage system consists of two standalone systems – primary and secondary system – which are housed in different fire compartments of the Heidelberg University Computing Centre. The structure of the LSDF2 (light gray parts) as well as the connection to other systems at Heidelberg University and other external partners is shown in Fig. 1 LSDF2 was put into operation in 2016.

The primary LSDF2 storage system in Heidelberg is currently based in the back-end on a HPE Seagate G200 system built by a cooperation of the companies HPE and Seagate (nowadays Cray). The secondary system is based on Dell hardware. These main storage building blocks are extended by additional network components, as well as access- and administration servers. As file system software "IBM Spectrum Scale (Advanced Edition)" was chosen for both the primary

and secondary system. This file system provides high performance and good scaling characteristics and offers many valuable features like snapshotting, information lifecycle management, cluster export services for NFS, SMB, and S3 for concurrent accesses to the same namespace. By this, IBM Spectrum Scale allows for an efficient and reliable operation of the productive storage system LSDF2.

In addition to the separation into two fire compartments, both stand-alone system parts are high-available. There is no single point of failure due to a two- or many-times redundancy of all components – including the network, switches and external connection – depending on their criticality. The power supply is backed with an uninterrupted power supply (UPS). As internal storage network an Infiniband FDR network is utilized, to be able to provide the required bandwidth and access times in the back-end.

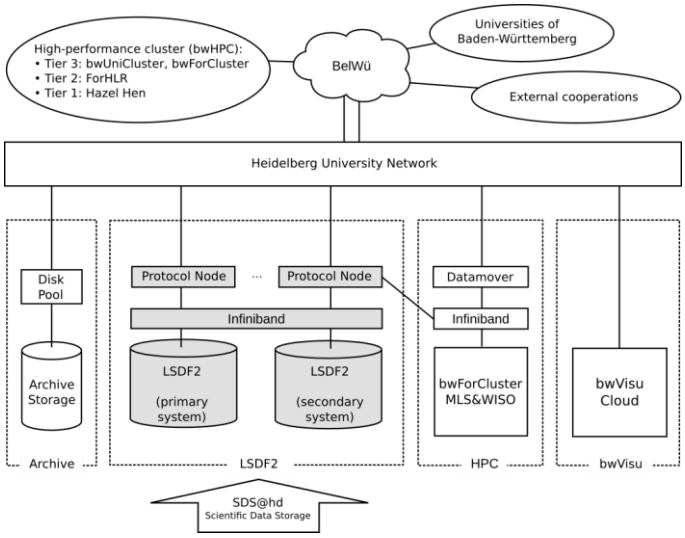

Fig. 1. LSDF2 concept at Heidelberg University and its connection to other data-intensive services at URZ and Baden-Württemberg.

The storage system is highly scalable, designed for a staged gradual expansion. The current first expansion stage of the primary system is equipped with a usable capacity of 5.8 PB. As storage media in total more than 1000 hard disk drives (HDD) and solid-state drives (SSD) are installed. The SSDs are used to store the metadata and also small files to enhance the performance. The data storage is organized in a parity declustered RAID protection scheme called GridRAID [3], which distributes the parity information over groups of 42 drives. In addition, the spare space is distributed over all drives in a storage group. In case of a disk drive failure, the parity is automatically reconstructed on all remaining drives during the recovery phase. With the distribution of the reconstructed data across all surviving drives, the important MTTR (mean time to recover) is no longer depending on the I/O performance bottleneck of a single drive and is reduced to less then a quarter compared to the excessive rebuild times of classical RAID 6 on disk drives with multiple TB space [4]. After replacing the failed disk drive, only a low-priority rebalancing needs to be done.

A scalable number of protocol nodes is used to provide the data access for the users via different file system protocols. These protocol nodes have Infiniband connections to the internal network as well as 40 Gbit/s Ethernet (40GE) connections directly to the backbone of the university and finally to the BelWü network [5] which connects the universities of the state of Baden-Württemberg. Currently the primary system can be reached via six 40GE network connections, so that data access with high bandwidth can be assured to users from Heidelberg as well as other universities of Baden-Württemberg. If needed, the aggregated bandwidth can be scaled up by increasing the number of protocol nodes and network connections.

For fail-over reasons data is stored on both systems, the primary and the secondary system, which are directly connected. In case of an outage of the primary system, the standalone secondary system can be accessed over a similar setup of network connections and protocol nodes. The secondary system provides a spatially separated copy of the data, which is available instantaneously without long rebuilding times from additional backup solutions. The throughput is reduced but the user data access can be directly enabled with a reduced performance. The purpose of the second storage system is not performance, but protection against loss of data and maintenance of service- and data-availability. The technical concept of the second storage includes one tier of online storage (ca. 50% of the primary system's capacity) and one tier of offline storage which leads to a good cost-effectiveness ratio. Since the second storage system contains only copies of data, its capacity does not lead to an increased usable total storage capacity of the service SDS@hd, but to a higher service quality.

III. SDS@HD SERVICE

A. *Service description*

The storage service "SDS@hd – Scientific Data Storage" serves as central and secure data storage of large-scale scientific research data. High performance data access in combination with high availability, data security and encryption for storage of "hot data" are the main properties of this service. "Hot data" means scientific data, which is often accessed and part of current active research, e.g. of data-driven sciences like life sciences, astrophysics or hydromechanics.

In the typical life cycle of scientific data, SDS@hd takes the role of the central storage which is accessed in different steps from data creation – e.g. from microscopy, gene sequencing or high performance computing (HPC) simulations – to data processing, analysis or visualization e.g. with compute clusters. When research activities with certain data are finished, the concerned data then can be handed over to repositories or archive solutions.

B. Access protocols

SDS@hd provides access via different data access protocols in order to enable data access from all commonly used operating systems which are used in the research community.

TABLE I. DATA ACCESS PROTOCOLS

Protocol	Properties	Recommended OS
SMB 2/3	high availability, fault tolerance and load balancing	Windows, OS X
NFSv4	high availability, fault tolerance and load balancing	Linux, *nix
SSHFS / SFTP	easy access without special installation and configuration	all

As can be seen in TABLE I. currently the three protocol types SMB, NFS and SSHFS/SFTP are provided. Both NFSv4 as well as SMB are intended for a stable connection of data sources like microscopes or gene sequencers, compute clusters, servers and workstations. The connection is highly available, fault tolerant and of high performance. SMB is intended for access from Windows and OS X, NFS for connections from Linux (or other Unix-like) systems. For both protocols the installation needs a minimal effort and support of the network firewalls. In case of the NFS connection also a handling of Kerberos tickets is necessary. However, the usage of the SSHFS/SFTP protocol is intended for accesses from all operating systems especially from desktop or mobile systems and from systems in restrictive or external networks. Additional protocols like S3 will be complemented based on the user's demands.

As mentioned in Section II user access of SDS@hd is managed by the protocol nodes, which easily can be scaled in number to handle further throughput demands of the data storage service.

C. Management and registration

For the management of the storage resources and accesses, so-called Speichervorhaben (SV) are introduced as the organizational units of SDS@hd. One SV consists of one storage share with a defined quota and it corresponds to a dedicated group of users. Members of the same SV are able to share data easily while the access of non-members by default is not possible. There is a responsible person for each SV who takes responsibility for the data stored in the SV and also for the reporting (e.g. for evaluations by the DFG). The SV responsible can manage members and roles for his or her SV via a web management tool. A fine-grained access management within one SV on the level of user roles is currently in development.

The use of SDS@hd is in general possible for all bwIDM member organizations and involves a registration procedure on the user level. bwIDM [6] is the federated identity management of Baden-Württemberg's universities which is realized as sub-federation of the DFN-AAI [7]. This technology allows the researchers to use their ID of their home institution when using SDS@hd. The access is controlled by the entitlements "sds-hd-user" and "sds-hd-sv" which are granted by each organization for their own staff and students.

Opening a new SV for a scientific project requires the entitlement "sds-hd-sv" that allows to initiate the creation of a contract between the SV responsible person and the SDS@hd operating institution which regulates usage, billing, and reporting issues. The entitlement "sds-hs-user" is required to use SDS@hd and to join an existing SV.

Before the first access to SDS@hd can be done, an additional registration step is needed in which a dedicated service password is set. At the same time, the user's account on the storage system is created.

D. Data security

In several data-intensive research areas like medicine and life science, there are high demands in data protection, especially when data from humans has to be stored (e.g. human gene sequencing or microscopy data). An overview of data protection concepts and technologies in the exemplary context of life science research is given in [8]. The authors conclude and summarize what is in principle well-known in the community: data security is related not only to technology or infrastructure – instead processes have to be designed in a way that the demands on data security can be fulfilled.

The service attributes of SDS@hd are such that particular data protection requirements can be met that might exist for various projects. In the following, relevant technologies, processes and concepts in that context are listed: The user access utilizes kerberized authentication and secure transport protocols. The owner of a share can independently manage the membership of a group of users to access the share and participate as a joint research project. In the backend of the storage service, an automatic and hardware assisted data encryption of all files is used, thus only encrypted data is stored on the disk drives. In case of failures, the defect drives remain and are shredded in the building. As usual, the access to the facilities and systems is restricted with personal access control. Specific demands on data security that users might have for certain projects can be checked individually by the operators of the service.

IV. ACCESS FROM BWHPC CLUSTERS

Many scientific communities need high performance computing (HPC) resources for data analysis. To facilitate this process, it is planned to provide access to SDS@hd from all bwHPC clusters [1] in Baden-Württemberg via datamover nodes. The connection can be established using any protocol that SDS@hd offers (see TABLE I.), whereby NFS is the recommended protocol for a production operation. Both the cluster file system and SDS@hd are connected with high bandwidth to the datamover nodes to allow fast copy processes and integration of the remote storage system into the cluster job management via data staging (yellow lines in Fig. 2). This concept is supported by the upgrade of the Baden-Württemberg scientific network BelWü to 100 Gbit/s. Datamover nodes are

already in production for the bwForCluster MLS&WISO (Production and Development) in Heidelberg [9] and the bwForCluster BinAC in Tübingen [10].

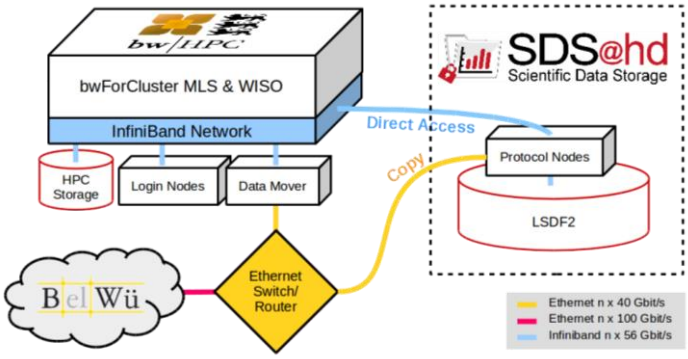

Fig. 2. Access to SDS@hd from bwForCluster MLS&WISO direct access (blue) and copy access (yellow). The latter is possible for other HPC clusters (e.g. bwHPC resources).

Due to the spatial proximity between the cluster and the storage system in Heidelberg, it is possible to provide a direct connection to SDS@hd on the compute nodes of bwForCluster MLS&WISO Production via InfiniBand (blue line in Fig. 2). This allows the direct analysis of SDS@hd data on the cluster without data staging or manual data synchronization. However, also for this direct network connection, currently only NFSv4 connections with Kerberos tickets are used. Other and potentially faster protocols for direct accesses are in preparation.

The management of Kerberos tickets on the cluster is done on the datamover nodes. Tickets are automatically renewed as long as possible. If security settings require manual ticket prolongation with password authentication, the user is notified by e-mail in time [11]. Direct access to SDS@hd on the compute nodes is used with growing interest.

The achievable performance depends on the workflow and the access patterns and is evaluated and improved in collaboration with the users. Currently, a series of performance measurements for different use cases and setups are made. The performance is analyzed systematically and compared to parallel scratch storage systems available in HPC clusters. The results will be publicated in a scientific contribution.

A close connection to SDS@hd is also in preparation for the remote visualization service bwVisu [12], which is located in Heidelberg. bwVisu will enter production state in 2018 and will provide resources for scalable and flexible data analysis and visualization.

V. User Meetings, Support and Consulting

The service SDS@hd is designed as generic as possible to meet the needs of a wide class of research projects. A flexible group, right and role management is possible that can be setup to fit to different workflows and multi-user and multi-role environments. However, to obtain the best result for specific problem classes, the capabilities of the storage service and the characteristics and limitations of the underlying storage system have to be well-understood. In some cases, the way how the research data is managed needs to be adapted for increased performance (e.g. conversion of many small files into fewer larger files). In regular user meetings the features and characteristics of SDS@hd and LSDF2 are presented to users and people who are interested in these topics. Showcasing the use of SDS@hd for representative project setups seems to be particularly beneficial for the users and is one main focus in these meetings. The use cases range from a single-user direct access scenario to setups with multi-users having different status, sharing parts of their data and using specific external analysis systems. These meetings have proven to be very valuable both for the service providers (to get feedback from the users) and for users (to discuss needs for their specific data-intensive projects).

The feedback from users and the utilization statistics are considered in the planning of further expansions. As outlined before, the storage system has a modular structure where capacity and performance can be extended and increased on a demand-driven way. Optimizations on the level of the network connection to local systems on site and also of the uplinks for cross-locational use are possible and will be implemented when needed. For the further development of the storage service, several ideas exist (e.g. access via S3, utilization of user datagram protocol (UDP) or data staging for accesses from high-performance compute clusters) which are prioritized by the expected added value for the users.

VI. Summary

The service SDS@hd for hot large-scale scientific data was introduced. The technical aspects of the underlying storage system with its components and characteristics were outlined. Subsequently, the storage service was described including the available access protocols, the management concept, the registration process, and data security aspects. The connection and interplay between Baden-Württemberg's HPC resources and SDS@hd was described and also the dedicated support that is given to users.

With the storage system LSDF2, the storage service SDS@hd and the accompanying activities the operators hope to provide a valuable service and support for the data-intensive scientific community in Baden-Württemberg.

VII. Acknowledgments

The SDS@hd is funded by the state of Baden-Württemberg and by the German Research Foundation (DFG) through grant INST 35/1314-1 FUGG.

References

[1] Rahmenkonzept der Hochschulen des Landes Baden-Württemberg für datenintensive Dienste – bwDATA (2015-2019): http://dx.doi.org/10.15496/publikation-21187

[2] SDS@hd – Scientific Data Storage. https://sds-hd.urz.uni-heidelberg.de

[3] Seagate. “Seagate ClusterStor G200 with IBM Spectrum Scale” (Datasheet), 2015

[4] Roskow, M. “The Unique Technical Benefits of an Engineered Solution for GPFS” (Presentation), 2015

[5] BelWü - das Landeshochschulnetz. https://www.belwue.de/

[6] Föderiertes Identitätsmanagement der baden-württembergischen Hochschulen. http://bwidm.de/

[7] DFN-AAI - Authentication and authorization infrastructure. https://www.aai.dfn.de/en/

[8] Helvey, T., R. Mack, S. Avula and P. Flook. “Data security in Life Sciences research”, Drug Discovery Today: BIOSILICO, Volume 2, Issue 3 (2004): 97-103, https://doi.org/10.1016/S1741-8364(04)02403-5

[9] bwForCluster MLS&WISO, http://www.bwhpc-c5.de/wiki/index.php/Category:BwForCluster_MLS&WISO

[10] bwForCluster BinAC. http://www.bwhpc-c5.de/wiki/index.php/Category:BwForCluster_BinAC

[11] Richling, S., M. Baumann, S. Friedel and H. Kredel. “bwForCluster MLS&WISO” in Proceedings of the 3rd bwHPC-Symposium: Heidelberg 2016, eds. S. Richling, M. Baumann, M. and V. Heuveline, heiBOOKS, http://dx.doi.org/10.11588/heibooks.308.418

[12] Schridde, D., M. Baumann and V. Heuveline. “Skalierbare und flexible Arbeitsumgebungen für Data-Driven Sciences” in E-Science-Tage 2017: Forschungsdaten managen, eds. J. Kratzke and V. Heuveline, heiBOOKS, http://dx.doi.org/10.11588/heibooks.285.377