Perspective of the Large Scale Data Facility (LSDF) supporting nuclear fusion applications

Rainer Stotzka, Volker Hartmann, Thomas Jejkal, Michael Sutter
Karlsruhe Institute of Technology (KIT)
Institute for Data Processing and Electronics (IPE)
Karlsruhe, Germany
Email: rainer.stotzka@kit.edu

Jos van Wezel, Marcus Hardt, Ariel García, Rainer Kupsch, Serguei Bourov
Karlsruhe Institute of Technology (KIT)
Steinbuch Centre for Computing (SCC)
Karlsruhe, Germany

Abstract—To cope with the growing requirements of data intensive scientific experiments, models and simulations the Large Scale Data Facility (LSDF) at KIT aims to support many scientific disciplines. The LSDF is a distributed storage facility at Exabyte scale providing storage, archives, data bases and meta data repositories. Open interfaces and APIs support a variety of access methods to the highly available services for high throughput data applications. Tools for an easy and transparent access allow scientists to use the LSDF without bothering with the internal structures and technologies. In close cooperation with the scientific communities the LSDF provides assistance to efficiently organize data and meta data structures, and develops and deploys community specific software on the directly connected computing infrastructure.

Keywords—Large Scale Data Facility; LSDF; data intensive science; data; meta data; services; Grid; Cloud

I. DATA INTENSIVE SCIENCE

Many scientific communities, the European Commission [1] and national funding agencies identified the importance to establish sustainable infrastructures to store, to archive and to process huge amounts of data.

In the last decade, several communities have approached this data challenge by community specific solutions, e.g. the Worldwide LHC Computing Grid (WLCG) [2], [3]. The data management in the Grid using Storage Elements [4] and in the Cloud [5] provide in principle general solutions. Contrary to that transparency and easy access for scientific users is not necessarily provided. In most Grid and HPC architectures the data has to be copied to the working nodes before it can be processed. The data transfers over the network represent often a tight bottleneck if many Gigabytes or Terabytes have to be processed. Some communities use digital object repositories [6], e.g. in humanities. These repositories also introduce meta data models to describe the data and its usage. In microscopy the Open Microscopy Environment [7] is providing a standard for handling stacks of microscopy images and their meta data.

However, the existing solutions are hardly interoperable and do not allow transparent access from one system to another.

II. REQUIREMENTS

Nowadays science and large scale scientific experiments are organized in virtual scientific communities crossing national borders. Collaboration, communication and the exchange of knowledge, methods and data are essential for their success.

The Large Scale Data Facility project [8] was started in 2009 to offer storage and data management services for scientific experiments and applications. To support a wide range of scientific communities (see Fig. 1) a broad variety of requirements need to be fulfilled:

- **Sustainability**: Most scientific data needs to be archived over a period of at least 10 years. A data facility must preserve the data and guarantee the preservation over a long period. Furthermore tools and software for access must be available to access and interpret antique data and meta data formats.
- **Availability and reliability**: A data facility must be available at a 24/7 level without interruption.
- **Scalability**: Because data and storage technologies are rapidly evolving, the facility must be designed to handle storage on an Exabyte scale, integrating heterogenous storage technologies running in parallel as distributed systems.
- **Performance and throughput**: We plan to support
applications and experiments processing up to a Terabyte per hour. Running several storage and computing systems in parallel connected by high-speed networks will be necessary to fulfill the requirements.

**Security:** Secure access must be guaranteed by authentication and authorization standards.

**Accessability:** Access must be provided by a variety of transfer protocols and application programming interfaces, which must be regularly extended by new technologies. Access via respective state-of-the-art Grid, Cloud, and Web protocols must be guaranteed.

**Tools:** A variety of tools must be provided for easy access and modification/versioning of data and meta data. The tools must be easily adaptable to the requirements of the different user communities and their specific data and meta data organization.

**Flexibility:** Each scientific community will define its own unique requirements regarding data and meta data organization and software tools. The structure of a data facility must be flexible in order to cope with these requirements without impacting the data integrity of other communities.

**III. LSDF LOGICAL ORGANIZATION**

Internally the LSDF will handle data and its meta data in a hierarchy of three different levels of catalogs.

The **Logical File Catalogs (LFCs)** translate logical file names to physical file names. This allows to hide the storage technologies from the user and to handle internally a variety of storage systems. Each new file will be registered in the LFCs data base returning only the logical file name to the user. All internal data movements (e.g. from disk storage to tape, or data replication) must be recorded in the LFCs. A variety of systems like iRods [9], dCache [10], and the LHC Grid [11], [12] have already implemented LFCs, which can be used by the LSDF.

Systems like the Consistent Physical Object (CPO) [13] for the fusion community, Fedora Commons [6] for arts and humanities, OMERo [7] for microscopy, or other proprietary software can be set up on top of the LFCs as well to meet the requirements of different scientific communities.

The **Logical Directory Catalogs (LDCs)** will organize the data in a virtual file system with virtual directory names.

The **Logical Project Catalogs (LPCs)** will describe projects by their meta data. For each type of scientific projects a meta data schema will be stored in a meta data repository. Most user interactions, e.g. search for a special data set, will take place in the LPCs. Once a data set is found, the logical directory name is returned allowing the user to access the files.

**IV. LSDF COMPONENTS**

**A. Hardware**

The core components consist of several disk systems connected via servers to a high speed Ethernet infrastructure. Directly connected is a set of computing nodes that serves as analysis farm.

The servers are DELL and IBM machines with dual quad-core Intel processors and 32 GB memory. All servers have 10 Gbit/s Ethernet (GE) interfaces and run Scientific Linux. The storage consists of DDN SFA100000 and IBM DS5300 disk systems with 8 Gbit/s fibre channel (FC) connections, holding 600 TB and 1.4 PB respectively.

Data intensive computing is provisioned with a cluster of 58 HP nodes with dual quad-core Intel processors and 36 GB memory each. Each node has two 1 TB disks and runs the Hadoop [14] environment on top of Scientific Linux 5.5. The resulting Hadoop distributed filesystem has a capacity of approximately 110 TB.
sustain a throughput of more than 1 GB/s from the data acquisition systems as well as to and from each storage instance within the LSDF. GPFS [17, 18] is used as a underlying file system. Part of the storage is configured with the Scale–out File System (SoFS) [19], a derivative of GPFS which integrates high availability services for CIFS and NFS [20].

Archival storage is provided with STK8500 and IBM TS3500 tape libraries. Data is stored on LTO media using CIFS and NFS [20].

At the end of 2010 a total of 2.0 PB disk and 2.5 PB tape storage are on–line. More than 80% of the storage was procured in cooperation with BioQuant of the University of Heidelberg [21] through special funding by the state of Baden–Württemberg. In middle of 2011 the disk space will increase to 4.5 PB.

B. Networks

The project aims to build a fully dedicated 10 Gbit/s network backbone running on the IPv6 protocol and interconnecting the different storage components with the institutes within KIT providing and consuming the scientific data.

Two redundant Cisco Nexus 7000 series routers operate at the core of the backbone, that interconnect the different institutes at layer two level by means of Virtual LANs. At each institute dedicated Cisco Catalyst 4900M switches allow the connection of the experimental data acquisition devices, supporting up to 10 Gbit/s links. Currently the network layer is based on the standard IPv4 protocol, but the project plans to start supporting the IPv6 protocol in the mid–term. The infrastructure for the core storage operations is completely separate; no “desktop” traffic is allowed. Connections to partner institutes and experiments for data taking and analysis partly use dedicated switches but also the existing KIT backbone.

All storage systems are connected to the backbone router via multiple 10 Gbit/s links, one per storage host. The network address–space is also dedicated to the LSDF project, allowing to avoid intra-campus firewalls and a more direct and faster routing. Moreover, a separate private network is dedicated to management of the core infrastructure.

In the first half of the year 2011 all the involved institutes will be joined with a 10 Gbit/s backbone, and the connection to the partner center BioQuant/Heidelberg will be upgraded to a 40 Gbit/s link. This high speed connection will enable back–up and tape–archiving of the partner’s data in the local tape facility. Connections to Heidelberg and the internet are provided for by the BelWü net that connects scientific institutes in the state of Baden–Württemberg.

C. Software and tools

1) Application programming interfaces: The LSDF will provide access via several protocols and storage methods. However not all storage instances of the LSDF will offer the same protocols. Also it is to be expected that protocols will change, new protocols will be supported and existing ones decommissioned. The evolution in protocol technologies would incur a steady adaptation for users of the LSDF. This would significantly lower acceptance by user groups as well as software developers, because every access technology has its own APIs. Software developers are forced to use those APIs and to be familiar with the underlying technologies. If a user group needs to access different parts of the LSDF as a requirement of different projects, the software developers of the corresponding group are forced to understand at least two APIs. In addition, if security features are included, this will even be more complicated.

Such behaviour would lead to a disaffirmation of the LSDF by user groups as well as software developers. Especially scientists are not interested in learning various APIs for accessing their data. They only want to store, access and process on the provided infrastructure.

To overcome this problem the access to the LSDF is possible using the Abstract Data Access Layer API (ADALAPI), providing platform and technology independent data access. The ADALAPI implements a logical software layer and hides specific details of the underlying technologies. From the logical point of view the ADALAPI is a "low–level interface” to the LSDF allowing operations with files and directories. The registration of a new data set in the LFC has to be done in another software layer, which is able to handle meta data. Like access to ADALAPI, access to the meta data level is based on secure authentication and available world–wide.

If a software developer wants to access data in the LSDF, the easiest way is to submit the request using the provided interfaces of the logical software layer. For successful access the ADALAPI prompts for the required details after the request is submitted. Such details can be e.g. a user name with password or valid credentials of a X.509 certificate. After successful submission of the required privileges access to data is granted.

The prompt for required information in the current implementation of the ADALAPI is implemented in three different flavors, which can be configured for every connection. One way is via a graphical user interface, another one via a command line and the third one via a configuration file. Especially a configuration file might be a security risk. The password, if required, is stored as plain text. Nevertheless, this method is necessary as the access is also possible from various computing infrastructures. Especially thereby it is not possible to provide the necessary information in a graphical manner or from the command line. In future versions authentication using eduGAIN [22], when available, will be considered.

After successfully establishing the connection, interaction with the LSDF is possible by calling the common methods of the ADALAPI. The common methods hide the details of the implementation, allowing to list folders, to store and to retrieve data for every implemented protocol. Additionally it is possible to retrieve the size of files, to check if a file is read- or writeable, if a file exists, to create
directories, and to query the date when the file was last modified. Another feature of the implementation is that it is possible to upload, download and to delete folders recursively, which is decided dynamically depending on the type of input (file or folder).

From the developers point of view, this behaviour allows the handling in an easy way, as he only has to connect via a specified method and is able to access data via the same interfaces regardless of the implemented technology. If an other access method is needed e.g. for a different part of the LSDF, only the configuration for the part to access has to be adapted. Everything else will be the same for all implemented technologies, making the adaption of software very easy and straight forward.

The current implementation of the ADALAPI allows the access via four technologies, which can be classified for KIT internal and worldwide external access to the LSDF. Internal methods are:

- A POSIX like access, where it is possible to access data from the local file system. This is necessary as the LSDF allows mounting of parts of it via Samba [23] over the CIFS protocol and via Network File System (NFS).
- The access to the Hadoop Distributed File System (HDFS) [14]. HDFS is based on the Google file system [24] and accessible over the APIs of the Hadoop project.

External methods are:

- An access via GridFTP [25]. The implementation is based on the CoG Kit [26] of the Globus Toolkit [27].
- Access via a Web browser over the http protocol. This allows access to the data via a Web server. The access does not need any configuration, as Web browsers are already installed on nearly every computer system. Currently access is read-only, the upload of data is not provided.

The ADALAPI is designed to be generic and can be extended by new technologies, which will be integrated in the LSDF in the future. Also the design allows to reuse the ADALAPI in several contexts, e.g. by graphical user interfaces or computing infrastructures.

2) The DataBrowser: At a first view the DataBrowser looks like a graphical user interface similar to the Windows Explorer. It provides easy access into the LSDF for storing and accessing data including user authentication using X.509 certificates based on the Globus Security Infrastructure [27]. Furthermore it provides the management of meta data according to a Logical Project Catalog.

The DataBrowser is an application programming interface and a graphical user interface for managing and analyzing large amounts of data and their meta data. Meta data schemes for various projects are implemented in the meta data repository. For uploading meta data entries have to be registered by the user and the data is uploaded using the ADALAPI. In order to access data sets first the meta data base is used. Then the requested data set is located. After that the data can be accessed and interpreted, e.g. displayed as an image.

For each scientific community and their specific needs the graphical user interface as well as the data and meta data organization can be easily adapted. Community specific services can be interfaced within the DataBrowser allowing convenient use of the LSDF’s functionalities.

![Figure 4. The DataBrowser graphical user interface.](image)

Figure 4 shows the DataBrowser. At the head information and the lifetime of the X.509 certificate is displayed. The Project Browser on the left allows the user to select a specific project or experiment. The right window allows a variety of views to the project data organization, as shown, or to the project meta data or results of community specific services. Selected files, e.g. images, can be directly interpreted and displayed.

In future a graphical user interface with full functionality running within a Web browser will be provided.

D. Community specific services

1) Data and meta data organization as a service: Certainly, there are many different preconditions for different communities in the field of data and meta data handling. A few communities already use advanced technologies and schemas for data and meta data handling. These technologies and schemas are integrated into the LSDF, either by complete adoption of technologies or by appropriate transformation of data and meta data in both directions.

The more challenging task in terms of a service idea is offered by communities that only have rough ideas and specifications how they want to store their data and which meta data they want to link to this data. In most cases these communities have no expertise on how to implement an entire standard for (meta) data organization. To bring these communities’ data into the LSDF we employ experts providing data and meta data organization as a service. This kind of service embraces the analysis of the communities’ data structures, the definition of appropriate meta data schemas with respect to existing, community specific meta data standards and the integration of data and meta data organization into LSDF components like LPC and LDC.

2) Integration and deployment of community services: Apart from data and meta data organization there is also the need for adapting the way how data is processed. To benefit from the advantages of the LSDF to a great
extend processing must be performed as close as possible to
the data. The logical connection between meta data,
data and data intensive applications is realized via special
APIs, e.g. the ADALAPI, which works hand in hand with
meta data repositories and high performance data access
technologies.

To enable existing algorithms to take advantage of
these technologies, data intensive applications are adapted
to (meta) data access APIs and computing infrastructure
in close cooperation with the respective code owners.
In most of all cases these code owners are algorithm
developers working on behalf of the addressed community,
in some cases algorithm developers are community
members themselves. To support these different developers
we have experts giving guidance and assistance during the
porting of existing applications and the development of
new applications connected to the LSDF. These experts
also provide help for the definition of workflows which
are necessary to connect data management and processing
in a feasible and performant way.

After the communities’ application is LSDF-enabled it
is deployed on the computing infrastructure as a service
within the computing center. Afterwards it is accessible by
the community and its partners. In fact, these applications
are subject of individual software life cycles. Hence, there
are mechanisms that allow to update deployed applications
at any time and to reflect different software versions within
processing meta data schemas.

For the case, that there are any problems with an application
or access to the LSDF, communities and community
developers have direct access to the Global Grid User
Support (GGUS) [28], which is hosted and developed at
KIT.

V. FUSION COMMUNITY SUPPORT

The first users of the LSDF are the Institute of Toxicology
and Genetics (ITG) at KIT that is writing data for
processing and long time storage. ITG research develops
and uses High Content Screening platforms for biomedical
research. Their experiments closely observe Zebrafish
embryos to detect specific changes during growth. Data
from several camera mounted automatically operating
microscopes is stored in the LSDF and tagged using the
DataBrowser. A lightweight batch system copies data to
the hadoop cluster and prepares it for processing. The
hadoop storage in effect is used as scratch space. Results
are copied back to permanent storage and included in the
relevant meta data description. During the first tests a total
of 1000 microscope images were processed with a speedup
of 100 compared to a state-of-the-art workstation. The
high throughput microscopy data of ITG is growing from
50 TB in 2009 to 4 PB in 2013. Early 2011, data from
the ANKA synchrotron lightsource at KIT will also be
stored and processed at the LSDF. The amount of data
from several synchrotron experiments is estimated to grow
from 50 TB to 1 PB in 2015.

Among the involved scientific communities, the fusion
community differs because of two points. One stems from
the fact that the simulation for fusion science is old when
compared to other simulation disciplines. Thus, some of
the codes which are required to run come with a wealth
of requirements on clusters where they can run, dependent
of e.g. operating systems and available compilers. The
other point lies in the distributed nature of the fusion
community. Not only are the scientists authors of the code
and located around the world, also the input data and
access is distributed.

While the first point challenges the flexibility in setting
up LSDF resources for fusion, the second point is more
demanding and challenges many of the requirements de-
scribed in section II.

One example for this is the connection to HPC–FF [29],
[30] at the Jülich Supercomputing Center (JSC). Although
local storage for this supercomputer is provided by JSC,
the input data is located at LSDF. Performant and transpar-
ent access to such data sets are the most important features,
as fusion scientists have neither time nor knowledge to
focus on the underlying technologies. LSDF will provide
the solutions required for seamless access of files stored
remotely.

This tightens the requirements made on one hand on
accessibility as well as on performance, because HPC–
FF would idle in case of delays in data delivery or even
downtimes. On the other hand, the seamless use of repli-
cation technologies, as suggested by LSDF, will improve
reliability and performance without an additional burden
on the user. Furthermore, the involved security aspects
will have to be addressed. This is in particular challenging
with regard to different security policies implied by both
centers, KIT and JSC.

Despite the fact that some fusion simulations have old
computational cores, also modern computing concepts are
incorporated by the fusion community. The resources and
the models offered by Grid computing as well as Cloud
computing, fit well to the requirements of the fusion
community in general. Both paradigms require adapted
access methods. While Grid computing usually accesses
files via GridFTP, this point is more complex for Clouds.
There, operating system images have to be stored and
access to data has to be provided. Although GridFTP can
be accessed from inside the Cloud instances, it is more
performant to provide access to storage via the Cloud
layer. This is to provide direct access to a virtual hard drive
(just like for the operating system) and thereby sparing
overhead otherwise introduced by the network (such as
TCP/IP stack on both sides).

However, fusion computing requires more than just
resources. We need to run many simulation codes, each
simulating a different aspect of the physics involved in
the reactor. The codes can roughly be subdivided into
two groups. One group runs on supercomputers, while the
second group runs on clusters (e.g. via the Grid) requiring
communication during their runtime.

Fusion community has defined the format of the data
to exchange, the “Consistent Physical Object”, CPO [13].
Together with the CPOs the transport method “Universal
Access Layer”, UAL was specified. The UAL relies on a centralistic approach, because consistency is valued over performance in the fusion simulations. Fusion has started with a dedicated cluster to serve UAL requests. Meanwhile the LSDF concept represents a more generic approach which suits very well the requirements imposed by the architecture of UAL. Hence, it is currently considered to use the LSDF services for future UAL provisioning.

VI. CONCLUSION

Since the start of the Large Scale Data Facility project in 2009, we attracted plenty of different scientific communities: from biology, over synchrotron light sources, nuclear fusion, material testing up to humanities. Beneath the provisioning of storage and access methods our user centric orientation and the ability to integrate new community services and software seamlessly have been major success points.

For a better data exchange in the fusion modeling community the LSDF with its projected strong network connection to the Jülich Supercomputing Center and the HPC–FF supercomputer will extend the productivity. Since the LSDF offers also strong support to high throughput experiment data management, in future also the analysis of interlocked experiment measurements on fusion reactors with simultaneously produced simulation results are feasible.

The basic services, storage hardware and networks are set up and running. We currently implement and extend the software services to allow transparent access to a variety of systems.

VII. ACKNOWLEDGEMENTS

This work has been supported by Helmholtz Research Programme Super Computing and the European Community Framework Programme 7 — Research infrastructures projects — EUFORIA — contract no. RII1804. We would like to thank the researchers of the Institute for Applied Computer Science, the Institute of Toxicology and Genetics and the Institute for Synchrotron Radiation (all KIT) for their contributions.

REFERENCES


