Journal of large-scale research facilities

Journal of large-scale research facilities, 5, A136 (2019) http://dx.doi.org/10.17815/jlsrf-5-172

Published: 25.03.2019

JUST: Large-Scale Multi-Tier Storage Infrastructure at the Jülich Supercomputing Centre

Forschungszentrum Jülich, Jülich Supercomputing Centre *

Instrument Scientists:
- Supercomputing Support, Jülich Supercomputing Centre, Forschungszentrum Jülich, phone: +49(0)2461 61 2828, sc@fz-juelich.de
- Helpdesk Distributed and Data Services, Jülich Supercomputing Centre, Forschungszentrum Jülich, phone: +49(0)2461 61 2828, ds-support@fz-juelich.de

Abstract: JUST is a versatile storage infrastructure operated by the Jülich Supercomputing Centre at Forschungszentrum Jülich. The system provides high-performance and high-capacity storage resources for the supercomputer facility. Recently, additional storage and management services, addressing demands beyond the high-performance computing area, have been added. In support of its mission, JUST consists of multiple storage tiers with different performance and functional characteristics to cover the entire data lifecycle.

1 Introduction

The JUST (Jülich Storage) cluster is a large-scale storage infrastructure operated by the Jülich Supercomputing Centre (JSC) at Forschungszentrum Jülich (Forschungszentrum Jülich, 2019). Currently installed in its fifth hardware and software generation, JUST has evolved from a dedicated storage cluster for JSC's leadership-class system to the central storage provider for the modular supercomputing facility at JSC. As such, it provides a variety of data-storage and management related services and is to be seen as a research infrastructure on its own.

The JUST investment and operational costs are covered by funding from different sources. A share of the system is funded by the German Ministry of Education and Science (Bundesministerium für Bildung und Forschung – BMBF) and the Ministry for Culture and Science of the State North Rhine-Westphalia (Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen – MKW) via the Gauss Centre for Supercomputing (GCS). A share of the system is funded by the Helmholtz Association (Helmholtz Association, 2019b) through the program "Supercomputing & Big Data" and the Helmholtz Data Federation initiative (Helmholtz Association, 2019a).

* **Cite article as:** Jülich Supercomputing Centre. (2019). JUST: Large-Scale Multi-Tier Storage Infrastructure at the Jülich Supercomputing Centre. *Journal of large-scale research facilities, 5*, A136. http://dx.doi.org/10.17815/jlsrf-5-172

2 JUST System Details

JUST is a heterogeneous storage infrastructure encompassing various types of storage media and utilizing multiple software technologies for data management. Different storage media feature different capacity and performance characteristics and offer unique economically optimal design points. By using them along with customized system designs for separate service classes – or tiers – the overall capacity and performance of the system can be optimized, albeit at the expense of a more demanding data management by users.

Figure 1: Disk-arrays in the storage cluster JUST in the facility of Jülich Supercomputing Centre. Copyright: Forschungszentrum Jülich GmbH.

The different tiers in JUST, along with their relative capacity and performance metrics, are schematically shown in Figure 2. The details are provided in the following sections.
Please note: Due to a continuously growing demand for storage capacity and access performance, the hardware configuration of JUST is subject to regular changes that are mostly performed during operation. In the following, the configuration at the time of this writing is specified. The article will only be updated when changes are performed that significantly affect the functionality or performance of the offered service.
Access to JUST services (in particular, storage space on the file systems) is provisioned and managed on a per-project basis. Compute projects on the supercomputers JUWELS (Jülich Supercomputing Centre, 2019) and JURECA (Jülich Supercomputing Centre, 2018) are automatically granted access to storage space on the HPC storage tier (see Section 2.3). Additional storage resources are made accessible to data projects, which may or may not be linked to compute projects.

2.1 `ARCHIVE` Storage Tier

The **`ARCHIVE`** storage tier is the lowest tier in the JUST hierarchy. It is designed to provide long-term storage for "cold" data, i.e., data that will likely not be accessed again in the near future. In particular, it allows for the storage of raw data underlying published scientific results. **`ARCHIVE`** is architected to

technically allow for storing data for more than one decade. Due to the funding scheme for the service, no guarantee about the availability of the service over such a long period can be given, though.

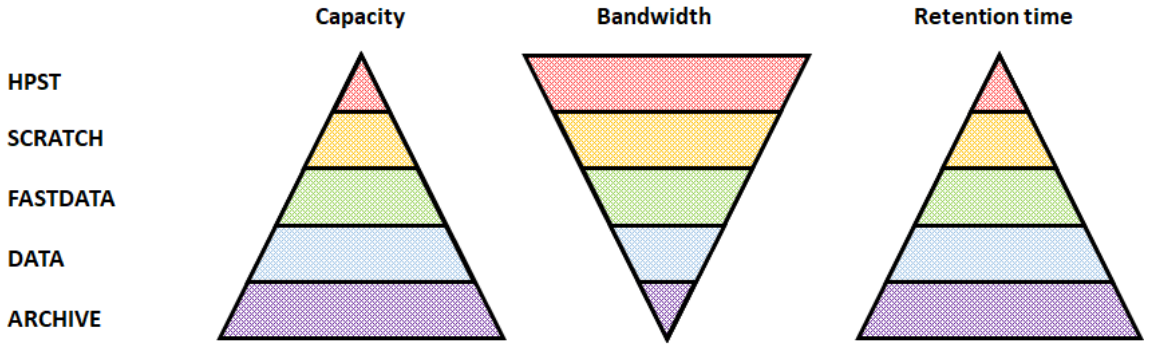

Figure 2: Schematic representation of the storage tiers in JUST with their capacity, bandwidth and (anticipated) data retention time characteristics. For simplicity, the small **HOME** file system as well as the **PROJECT** file system, that is solely accessible for compute projects, are not shown in the sketch.

ARCHIVE is accessible on the login nodes of the supercomputers JUWELS (Jülich Supercomputing Centre, 2019) and JURECA (Jülich Supercomputing Centre, 2018) as well as the JUDAC data access nodes (see Section 4). It is not accessible on the compute nodes of the supercomputers.
The performance of the system for an individual user depends strongly on overall system utilization and data location. Data access latencies can be very large. The average bandwidth for an individual file access will be in the range of 200 – 400 MB/s. Files in the **ARCHIVE** file system should be several Gigabytes up to ca. 8 TB in size. Smaller files must be packed using, e.g., the tar utility. Strict limits on the number of files (i-nodes) per data project apply. Movement of large data sets (e.g., renaming of directories) within the file system should be avoided due to the consumed resources by such operations.

2.1.1 Infrastructure Details

ARCHIVE is implemented as a two-tier storage system utilizing hierarchical storage management (HSM) techniques. The **ARCHIVE** is accessible as an IBM Spectrum Scale (IBM, 2019b) file system. Recently accessed data is stored in a disk pool and automatically migrated to/from a storage pool based on tape media. The system utilizes tape storage as the underlying medium in order to store the highest possible capacity with low footprint, minimal operational costs for electricity and cooling, and long durability. The data movement is managed by the "IBM Tivoli Storage Manager for Space Management" software based on IBM Spectrum Protect (formerly Tivoli Storage Manager) (IBM, 2019a). Two data copies are stored on tape.
The disk pool for the **ARCHIVE** system is located on a Lenovo DSS-G 26 (Lenovo, 2019) storage system with 10 TB disk drives. The accessible disk pool is regularly adjusted in size in the range between 1.5 to 3 PB. The DSS-G 26 servers are connected with 100 Gb/s Ethernet to the facility Ethernet fabric (see Section 3). In addition to the **ARCHIVE** disk pool, the underlying IBM Spectrum Protect infrastructure accesses a disk pool with a capacity of ca. 3.3 PB based on a Lenovo GSS-26 and IBM ESS GL6, both with 6 TB disk drives, for caching.
JSC operates three tape storage libraries: Two Oracle StorageTek SL8500 (Oracle, 2019) libraries with 6,600 and 10,000 slots, respectively, and an IBM TS4500 library (IBM, 2019c) with 21,386 slots. In the two SL8500 libraries, Oracle STK T10000 T2 tape cartridges are used. Each library is equipped with 20 T10000D tape drives that can write up to 8.5 TB per cartridge at up to 252 MB/s (uncompressed). The TS4500 library is equipped with LTO7-M8 tape cartridges and 20 IBM TS1080 Ultrium 8 tape drives capable of writing 9 TB per cartridge with up to 360 MB/s (uncompressed). All drives connected to one of four central switches of the Storage Area Network (SAN) with 8 Gb/s Fibre-Channel (FC) links.

The IBM Spectrum Protect server instances are hosted on eight Power 8 servers, each connected with four 16 Gb/s FC links to the central SAN switches. The servers are used for the **ARCHIVE** service as well as for external backup of other JUST file systems and auxiliary backup clients. For this reason, only about one eigth of the accumulated SAN injection bandwidth of the infrastructure is available for **ARCHIVE** (nominal 8 GB/s).

2.2 DATA Storage Tier

The **DATA** storage tier is intended for mid-term storage of data that is used as part of a scientific study over a multi-year period and re-used periodically. The system is designed for high capacity with reasonable bandwidth and predictable performance. In contrast to the **ARCHIVE** service, **DATA** is suitable for storing very large data sets that will be accessed again in, e.g., a month's time.
DATA is accessible on the login nodes of the supercomputers JUWELS (Jülich Supercomputing Centre, 2019) and JURECA (Jülich Supercomputing Centre, 2018) as well as the JUDAC data access nodes (see Section 4). It is not accessible on the compute nodes of the supercomputers. Data access latencies on **DATA** are low. Since **DATA** capacity will grow continuously with demand, file system performance will vary over time. A minimal accumulated file system bandwidth of at least 10 GB/s is maintained. The system's nominal bandwidth under optimal conditions is several times higher.
Regular file system snapshots of the **DATA** file system serve as an internal backup mechanism and enable to recover recently deleted files. The **DATA** file system is not externally backed up. Please note that the snapshot mechanism does not protect against the very unlikely case of a file corruption due to, e.g., file system software bugs or a full system failure.
DATA is the primary storage tier for data projects. It enables data sharing across different (compute) projects. It also enables data sharing with communities and the creation of additional data-based services through a connection with the HDF cloud infrastructure (see Section 5).
In the future, an object-storage service based on the **DATA** hardware and system software is planned.

2.2.1 Infrastructure Details

The **DATA** storage tier is implemented as a dedicated IBM Spectrum Scale (IBM, 2019b) cluster build on a scalable hardware infrastructure organized in building blocks. Each building block consists of two servers (Lenovo SR650 with Intel Xeon Gold CPUs), acting as Network Shared Disk (NSD) servers in a high-available mode, connected to four Lenovo ThinkSystem DS2600 storage controllers via 12 Gb/s SAS. Each DS2600 controller connects to three Lenovo D3284 expansion enclosures via SAS. Each building block contains 1,008 disks. The "ADAPT" controller firmware feature provides a declustered redundant array of disks (RAID) functionality for fast rebuild times after disk failure. The NSD servers each connect with four 100 Gb/s Ethernet to the facility Ethernet fabric (see Section 3).
The system will be enlarged in yearly increments until 2021. The initial deployment in 2018 consists of four building blocks with 10 TB drives (ca. 30 PB usable disk space) and one building block with 12 TB drives (ca. 9 PB usable disk space). The available disk space for **DATA** is lower due to the collocated services on the same hardware.
To enable data access from external sources, in particular the HDF OpenStack cluster (see Section 5), the IBM Spectrum Scale Cluster Export Service (CES) feature is used. This enables NFS re-exports of the Spectrum Scale **DATA** file system based on the Ganesha NFS fileserver. The **DATA** storage tier includes eight IBM Power 8 servers that are three-way virtualized as 24 systems for the CES export. Each virtualized host connects with two 100 Gb/s Ethernet links to the facility Ethernet fabric.

2.3 Large-Capacity HPC Storage Tier

The large-capacity HPC storage tier is the primary storage provider for the supercomputing systems in the high-performance computing facility at Jülich Supercomputing Centre. It provides several file systems with different purposes and characteristics.

The **HOME** file system stores the users' home file systems and user-specific configuration files (including public keys for the secure shell login). The file system offers low performance and minimal capacity. All data in **HOME** is externally backed up regularly to tape.
The **PROJECT** file system provides persistent storage space for compute projects. It is intended to store application codes, compiled software and input or output data of moderate size. The file system offers moderate performance and capacity. For high-bandwidth access to, e.g., simulation input data, staging to the **SCRATCH** file system is advised. All data in **PROJECT** is externally backed up regularly to tape.
The **SCRATCH** file system provides temporary storage space for compute projects. It is intended to store (large) simulation input and output data (including, checkpoints) and is optimized for high bandwidth. Data should not remain longer than necessary on the **SCRATCH** file system. **SCRATCH** is not backed up and older files are regularly deleted.
In addition to the above-mentioned file systems that are exclusively available for compute projects, the large-capacity HPC storage tier offers the **FASTDATA** service. **FASTDATA** provides persistent high-performance storage for data projects. It augments the **DATA** service for use cases where the combined use of **DATA** and **SCRATCH** is not feasible, e.g., because very large data sets are repeatedly processed in their entirety on the supercomputers in regular time intervals. Prerequisite for a **FASTDATA** data project is an active compute project and a use case that falls in the mentioned category. Regular file system snapshots of the **FASTDATA** file system serve as an internal backup mechanism and enable to recover recently deleted files. The **FASTDATA** file system is not externally backed up. Please note that the snapshot mechanism does not protect against the very unlikely case of a file corruption due to, e.g., file system software bugs or a full system failure.
The **SCRATCH** and **FASTDATA** file systems can provide up to 380 GB/s bandwidth for optimized parallel I/O. Please note that the file system bandwidth currently exceeds the maximal bandwidth that the individual supercomputers JURECA and JUWELS can attain.
All four file systems are accessible on the login and compute nodes of the supercomputers at JSC as well as the JUDAC cluster.

2.3.1 Infrastructure Details

The JUST large-capacity HPC storage tier is implemented with 21 Lenovo DSS-G 24 (Lenovo, 2019) storage building blocks. Of these, 18 building blocks serve the **SCRATCH** and **FASTDATA** file systems. **HOME** and **PROJECT** are provided by three storage units. Each building block contains two NSD servers (Lenovo SR650 with Intel Xeon Gold CPUs), four Lenovo D3284 enclosures and provides a raw capacity of 3.3 PB with 10 TB disk drives. The NSD servers connect with three 100 Gb/s Ethernet to the facility Ethernet fabric (see Section 3).

2.4 High-Performance HPC Storage Tier

The high-performance storage tier (HPST), a performance-optimized low-capacity tier that will utilize non-volatile memory technologies, is scheduled for a production start in Q4 2019. Further details will be provided in an update of this article once the HPST enters its production phase.

3 Network Infrastructure

In order to optimally support the different JUST services a highly flexible, scalable and performing network infrastructure is required. To this end, JSC is using Ethernet as the core network technology for the JUST services due to its high standardization, flexibility and cost efficiency.
The central Ethernet fabric in the supercomputing facility is based on three Cisco Nexus 9516 modular switches. The highest bandwidth utilization is due to the storage connection of the supercomputing systems (see Section 2.3 and Section 2.4). For this reason, the architecture of the fabric is designed primarily to optimize the performance of the IBM Spectrum Scale software. Specifically, the fabric is

designed for high north-south bandwidth (for client/storage server communication). It provides, by design, only limited performance for east-west traffic patterns.
The majority of the JUST components connect to the fabric with 100 Gb/s Ethernet. Most of the supercomputers are currently connected via 40 Gb/s Ethernet to the switches.

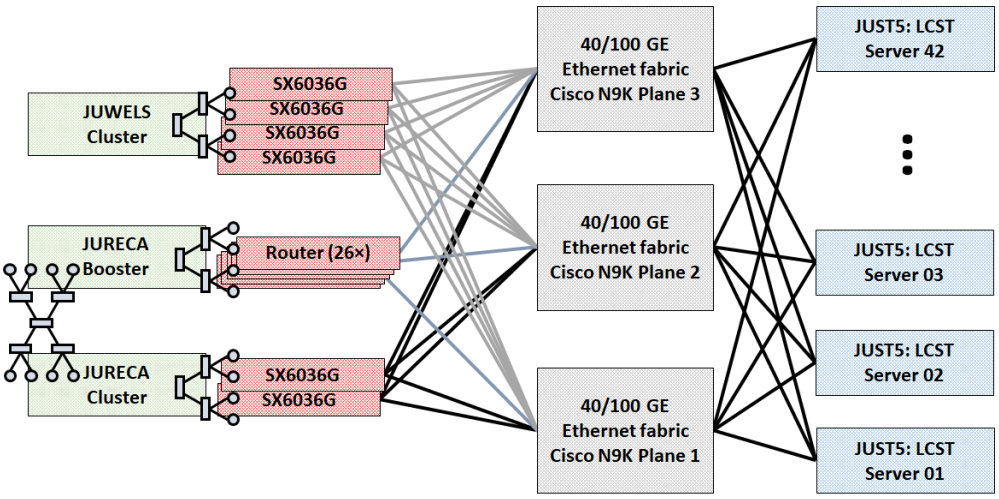

Figure 3: Schematic of the facility Ethernet fabric with a focus on the connection between supercomputing clients and the JUST large-capacity HPC storage tier (LCST). Please note that the sketch is rotated by 90° for the sake of presentation, i.e., the north-south traffic corresponds to a horizontal data flow.

4 JUDAC Cluster

The JUST Spectrum Scale file systems are accessible on the production supercomputers in the facility at JSC, in particular, on JURECA (Jülich Supercomputing Centre, 2018) and JUWELS (Jülich Supercomputing Centre, 2019). The login nodes of the supercomputers can be used for data pre- and post-processing as well as data management. In order to provide a consolidated platform for data transfer in and out of the facility, as well as to provide data access for users during maintenance windows of the supercomputers, the Jülich Data Access (JUDAC) service was created. JUDAC in particular provides common tools for wide-area network data transfer, such as GridFTP (Grid Community Forum, 2019) and Unicore UFTP (UNICORE Forum e.V., 2019), along with the necessary firewall configuration settings.
The JUDAC cluster consists of three nodes with X86-64 CPUs and two 100 Gb/s Ethernet connections to the facility Ethernet fabric. Two nodes are user accessible. The third node is reserved for hosting data transfer service processes.

5 HDF Cloud Infrastructure

The **DATA** storage service facilities data sharing and exchange across compute projects within the JSC supercomputing facility. In order to be able to participate in this data sharing, an active user account is required. In order to enable data sharing with external users, e.g., via a web service, additional infrastructure is required.
The HDF cloud infrastructure is a virtual machine hosting infrastructure based on OpenStack (OpenStack community, 2019). It allows provisioning and management of user-controlled virtual machines with a Linux operating system. The terms and conditions for services provided by virtual machines on the cloud infrastructure are regulated by an acceptable use policy.
Portions on the **DATA** file system can be exported via NFS into virtual machines hosted on the cloud infrastructure. This enables the creation of community-specific, data-oriented services such as web-accessible databases. Since the service-providing virtual machines are community-managed they are

isolated from the user management of the supercomputing facility. For this reason, all NFS-exports are protected with an "uid mapping" that alters the visible data ownership to typically a single user for read and write accesses. In particular, fine grained access control capabilities through the file system layer itself are limited and, if required, need to be implemented on the service layer.

References

Forschungszentrum Jülich. (2019). *Forschungszentrum Jülich webpage.* Retrieved from http://www.fz-juelich.de

Grid Community Forum. (2019). *GridCF webpage.* Retrieved from https://gridcf.org

Helmholtz Association. (2019a). *Helmholtz-Datafederation (HDF).* Retrieved from https://www.helmholtz.de/en/research/information_data_science/helmholtz_data_federation

Helmholtz Association. (2019b). *Helmholtz-Gemeinschaft Deutscher Forschungszentren e.V. (HGF) webpage.* Retrieved from http://www.helmholtz.de

IBM. (2019a). *IBM Spectrum Protect product webpage.* Retrieved from https://www.ibm.com/de-en/marketplace/data-protection-and-recovery

IBM. (2019b). *IBM Spectrum Scale product webpage.* Retrieved from https://www.ibm.com/de-en/marketplace/scale-out-file-and-object-storage

IBM. (2019c). *IBM TS4500 Tape Library webpage.* Retrieved from https://www.ibm.com/de-en/marketplace/ts4500

Jülich Supercomputing Centre. (2018). JURECA: Modular supercomputer at Jülich Supercomputing Centre. *Journal of large-scale research facilities*, *4*, A132. http://dx.doi.org/10.17815/jlsrf-4-121-1

Jülich Supercomputing Centre. (2019). JUWELS: Modular Tier-0/1 Supercomputer at the Jülich Supercomputing Centre. *Journal of large-scale research facilities*, *5*, A135. http://dx.doi.org/10.17815/jlsrf-5-171

Lenovo. (2019). *Lenovo DSS-G product webpage.* Retrieved from https://www.lenovo.com/gb/en/data-center/servers/high-density/Distributed-Storage-Solution-for-IBM-Spectrum-Scale/p/WMD00000275

OpenStack community. (2019). *OpenStack project webpage.* Retrieved from http://www.openstack.org

Oracle. (2019). *Oracle StorageTeck SL8500 Modular Library System webpage.* Retrieved from https://www.oracle.com/storage/tape-storage/sl8500-modular-library-system

UNICORE Forum e.V. (2019). *Uniform Interface to Computing Resources (UNICORE) webpage.* Retrieved from http://www.unicore.eu