



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, 2021). 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

*Cite article as: Jülich Supercomputing Centre. (2021). JUST: Large-Scale Multi-Tier Storage Infrastructure at the Jülich Supercomputing Centre. *Journal of large-scale research facilities*, 7, A180. <http://dx.doi.org/10.17815/jlsrf-7-180>



(Helmholtz Association, 2021b) through the program "Supercomputing & Big Data" and the Helmholtz Data Federation initiative (Helmholtz Association, 2021a). One JUST subsystem was funded by the "Interactive Computing E-Infrastructure for the Human Brain Project" project (Human Brain Project, 2021).

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), JURECA (Jülich Supercomputing Centre, 2018) and JUSUF (Jülich Supercomputing Centre, 2021) 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.

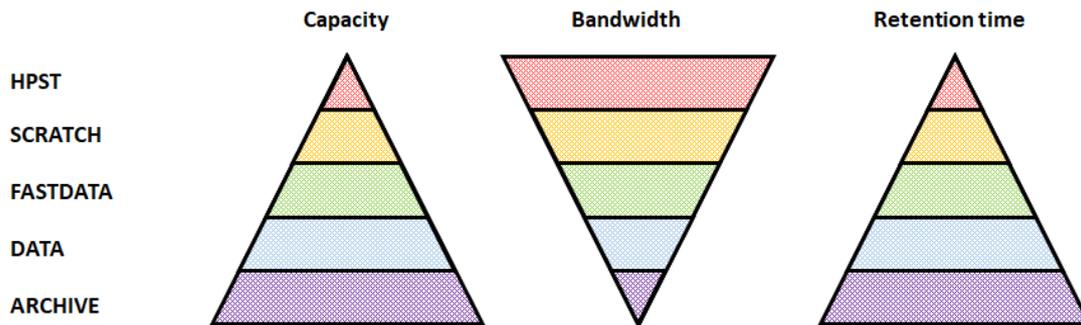


Figure 2: Schematic representation of the storage tiers in JUST with their capacity, bandwidth and (anticipated) data retention time characteristics. For simplicity, the small **HOME** file system as well as the **PROJECT** file system, that is solely accessible for compute projects, are not shown in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

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.

ARCHIVE is accessible on the login nodes of the supercomputers JUWELS (Jülich Supercomputing Centre, 2019), JURECA (Jülich Supercomputing Centre, 2018) and JUSUF (Jülich Supercomputing Centre, 2021) 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, 2021b) 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, 2021a). Two data copies are stored on tape.

The disk pool for the **ARCHIVE** system is located on a Lenovo DSS-G 26 (Lenovo, 2021) 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. 8.3 PB based on two Lenovo DSS-240 with 16 TB disk drives for caching.

JSC operates three tape storage libraries in two buildings to support different fire sections: Two Oracle StorageTek SL8500 (Oracle, 2021) libraries with 6,600 and 10,000 slots, respectively, and an IBM TS4500 library (IBM, 2021c) 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). In addition the 10,000 slot library has 18 IBM LTO8 tape drives installed using LTO7-M8 (9 TB) and LTO8 (12 TB) tape cartridges. These tape drives can read and write data with a bandwidth of up to 360 MB/s (uncompressed). The TS4500 library is equipped with LTO7-M8 tape cartridges and 20 IBM TS1080 Ultrium 8 (LTO8) tape drives capable of writing 9 TB per cartridge with up to 360 MB/s (uncompressed). All drives are 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 IBM Power 8 and one Intel x86 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 eighth 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), JURECA (Jülich Supercomputing Centre, 2018) and JUSUF (Jülich Supercomputing Centre, 2021) 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 also backed up by IBM Spectrum Protect but only on a weekly base due to the huge amount of data.

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 and the JUSUF-Cloud infrastructure (see Section 5 and (Jülich Supercomputing Centre, 2021)).

Starting 2021, a separated partition of the **DATA** hardware and system software is used to provide object-storage service. In the beginning only selected projects will get preliminary access.

2.2.1 Infrastructure Details

The **DATA** storage tier is implemented as a dedicated IBM Spectrum Scale (IBM, 2021b) cluster built on a scalable hardware infrastructure organized in building blocks. The initial deployment in 2018 consists

of four building blocks with 10 TB drives (ca. 30 PB usable disk space). In yearly increments the system will be enlarged until September 2021.

Each building block consists of two servers (Lenovo SR650 with Intel Xeon Gold CPUs), acting as Network Shared Disk (NSD) servers in a high-availability mode. Each NSD server is connected with four 100 Gb/s Ethernet links to the facility Ethernet fabric (see Section 3).

In the first 5 building blocks the NSD servers are connected to four Lenovo ThinkSystem DS6200 storage controllers via 12 Gb/s SAS. Each DS6200 controller connects to three Lenovo D3284 expansion enclosures via SAS. One building block contains 1,008 disks using 10 TB disks in the first 4 systems and 12 TB disks in the 5th one. The “ADAPT” controller firmware feature provides a declustered redundant array of disks (RAID) functionality for fast rebuild times after disk failure.

For building block 6, 7 and 8 the storage subsystem was replaced. The NSD server pairs are connected to four Lenovo DE6000-4U storage controllers via 12 Gb/s SAS. Each controller connects to four Lenovo DE600 Expansion Enclosure JBODs via SAS. The building block 6 has in total 1,040 disks using 12 TB, building block 7 and 8 consist of 928 disks each using 16 TB disks. The Lenovo DE6000-U controller firmware provides a declustered redundant array of disks (RAID) functionality for fast rebuild times after disk failure named “Dynamic Disk Pooling (DDP)”.

In 2021, the **DATA** storage tier provides roughly 75 PB usable disk space in total and is shared across different services.

To enable data access from external sources, in particular the HDF OpenStack cluster (see Section 5) and the JUSUF-Cloud (see (Jülich Supercomputing Centre, 2021)), 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.

Since Q3 2021 a separate Spectrum Scale file system has been used and exported as an object-store service using the CES feature, too. It provides the Amazon S3 and the SWIFT object data access protocol which can be used from the JUDAC data access nodes (see Section 4).

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 directory and user-specific configuration files (including public keys for the secure shell login). This 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 also backed up by IBM Spectrum Protect but only on a weekly base due to the huge amount of data.

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 JUSUF, 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, 2021) 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 four 100 Gb/s Ethernet links to the facility Ethernet fabric (see Section 3).

2.4 High-Performance HPC Storage Tier

The high-performance storage tier (**HPST**) is a performance-optimized low-capacity tier that utilize non-volatile memory technologies (NVMe). It is a cache layer on top of the **SCRATCH** file system providing high bandwidth and low latency access. Its purpose is to speed up I/O in compute jobs. One single client node can get up to 8 GB/s. Overall the **HPST** provides up to 2 TB/s bandwidth for optimized parallel I/O. Please note that the cache bandwidth exceeds the maximal bandwidth that the individual supercomputers JUSUF, JURECA-DC and JUWELS can attain.

The **HPST** provides one global namespace for all clients. Data can be staged in and staged out the underlying **SCRATCH** file system by special commands. Writing data to the cache will be stored in the cache. Reading data which are not yet staged in will not be automatically copied to the cache.

HPST is accessible on JUSUF, JURECA-DC and JUWELS login and compute nodes. Prerequisite to get access to **HPST** is an active compute project and a use case that benefits significantly from the I/O enhancement.

2.4.1 Infrastructure Details

Hardware The JUST high-performance storage tier is a cluster of 110 DDN IME-140 servers (DataDirect Networks (DDN), 2021). Each server is equipped with two dual port Mellanox HDR Infiniband Connect-X6 cards where one of the ports is connected to the EDR Infiniband interconnect of the **HPST**

cluster. Two Ports are used to connect the server directly to JUWELS', JURECA-DC's or JUSUF's Infini-band network. The fourth port is configured to 100 Gb/s Ethernet and connected to the facility Ethernet fabric (see Section 3). Each server contains 10 NVMe drives with 2 TB capacity each.

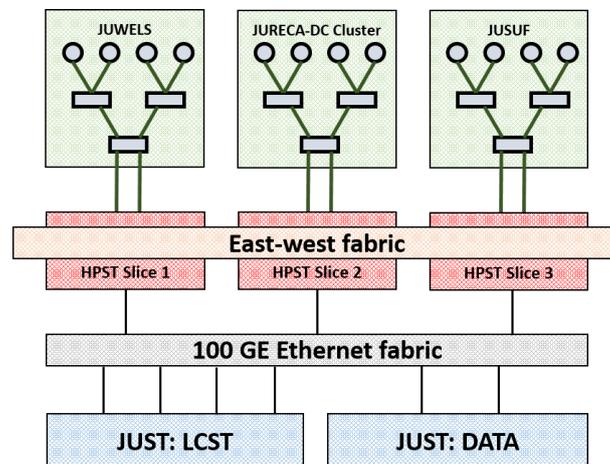


Figure 3: Schematic of the **HPST** setup and network design. One single cache divided in three slices and their connection to the assigned compute cluster.

Software The **HPST** is based on the “Infinite Memory Engine (IME)” (DataDirect Networks (DDN), 2021). It is a client-server architecture where the server creates a cache on top of the file system **SCRATCH** which is mounted on all servers. The cache itself uses the 10 NVMe disks on each server.

There are two technologies to access the **HPST** on JUSUF, JURECA-DC and JUWELS. The IME cache is mounted at `/p/ime-scratch/fs` by the IME FUSE client to get POSIX access, alternatively applications can use the IME native interfaces to use the cache directly.

Design One single cache that is offering a single, global namespace is created and accessible on the connected supercomputers. This namespace is divided into slices. The JUWELS slice contains 54 servers, the JURECA-DC slice 44 servers and the JUSUF slice 10 servers as shown in Figure 3. Client nodes can only communicate with the assigned slice servers. This is to ensure the separation of the supercomputers' interconnect and data locality for best performance. Write requests will store data on the slice servers, read requests will be forwarded to another slice if needed.

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, reliability and cost efficiency.

The central Ethernet fabric in the supercomputing facility is based on four 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. While JUSUF nodes are using plain 40 Gb/s Ethernet to connect to the Ethernet fabric and with this to JUST, all other HPC clusters use different technologies. The JUWELS Cluster is using Mellanox SX6036 gateway switches to translate from the cluster-internal Infiniband to the Ethernet fabric. The JURECA Booster is using gateway nodes that translate from OmniPath to Ethernet (2*40 Gb/s per gateway) The JUWELS Booster and JURECA-DC are using a new solution called NVIDIA Mellanox Skyway to translate from Infiniband to Ethernet (8*100 Gb/s per gateway)

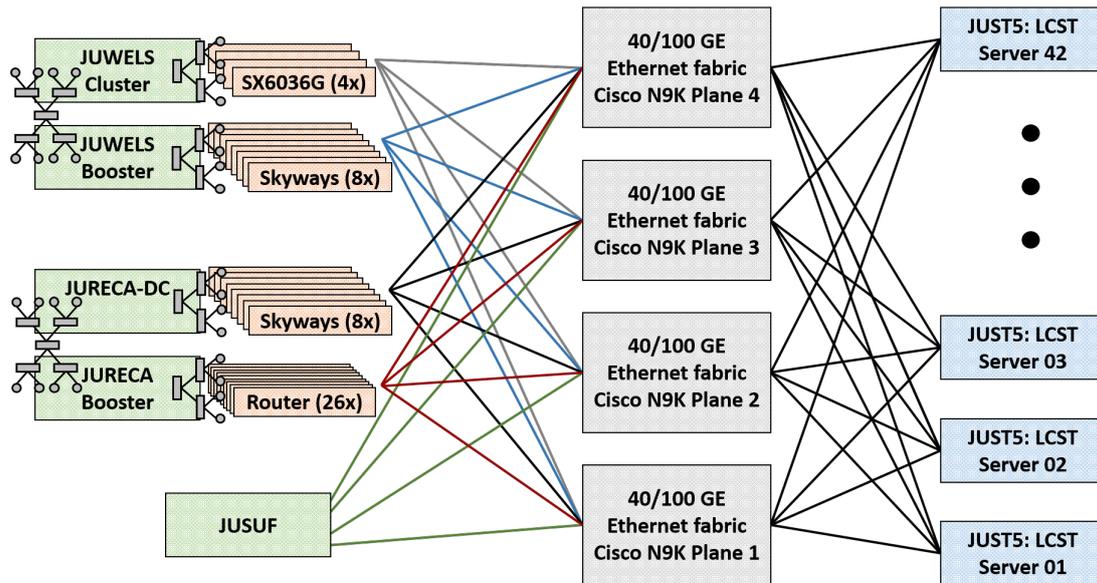


Figure 4: Schematic of the facility Ethernet fabric with a focus on the connection between supercomputing clients and the JUST large-capacity HPC storage tier (LCST). Please note that the sketch is rotated by 90° for the sake of presentation, i.e., the north-south traffic corresponds to a horizontal 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 JUWELS (Jülich Supercomputing Centre, 2019), JURECA (Jülich Supercomputing Centre, 2018) and JUSUF (Jülich Supercomputing Centre, 2021). 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, 2021) and Unicore UFTP (UNICORE Forum e.V., 2021), 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 services.

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, 2021). 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. The service-providing virtual machines are community-managed and 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 level.

References

- DataDirect Networks (DDN). (2021). *Infinite Memory Engine webpage*. Retrieved from <https://www.ddn.com/products/ime-flash-native-data-cache>
- Forschungszentrum Jülich. (2021). *Forschungszentrum Jülich webpage*. Retrieved from <https://www.fz-juelich.de>
- Grid Community Forum. (2021). *GridCF webpage*. Retrieved from <https://gridcf.org>
- Helmholtz Association. (2021a). *Helmholtz-Datafederation (HDF)*. Retrieved from <https://www.helmholtz.de/forschung/information-data-science/helmholtz-data-federation-hdf>
- Helmholtz Association. (2021b). *Helmholtz-Gemeinschaft Deutscher Forschungszentren e.V. (HGF) webpage*. Retrieved from <https://www.helmholtz.de>
- Human Brain Project. (2021). *HBP-ICEI webpage*. Retrieved from <https://www.humanbrainproject.eu/en/massive-computing/fenix-icei/>
- IBM. (2021a). *IBM Spectrum Protect product webpage*. Retrieved from <https://www.ibm.com/products/data-protection-and-recovery>
- IBM. (2021b). *IBM Spectrum Scale product webpage*. Retrieved from <https://www.ibm.com/products/spectrum-scale>
- IBM. (2021c). *IBM TS4500 Tape Library webpage*. Retrieved from <https://www.ibm.com/products/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>
- Jülich Supercomputing Centre. (2021). JUSUF: Modular tier-2 supercomputing and cloud infrastructure at Jülich Supercomputing Centre. *Journal of large-scale research facilities*, 1.



Lenovo. (2021). *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. (2021). *OpenStack project webpage*. Retrieved from <https://www.openstack.org>

Oracle. (2021). *Oracle StorageTek SL8500 Modular Library System webpage*. Retrieved from <https://www.oracle.com/storage/tape-storage/sl8500-modular-library-system>

UNICORE Forum e.V. (2021). *Uniform Interface to Computing Resources (UNICORE) webpage*. Retrieved from <https://www.unicore.eu>