ISBN: 978-3-95450-255-4 ISSN: 2226-0358 doi: 10.18429/JACoW-ICALEPCS2025-THCR001

DATA MANAGEMENT INFRASTRUCTURE FOR EUROPEAN XFEL

J. Malka*, S. Aplin, D. Boukhelef, K. Filippakopoulos, L. Maia,
T. Piszczek, G.Previtali, J. Szuba, K. Wrona,
European XFEL, Schenefeld, Germany
S. Dietrich, M. Gasthuber, J. Hannappel, K. Hoyos, M. Karimi, Y. Kemp, R. Lueken,
T. Mkrtchyan, K. Ohrenberg, M. Sahakyan, F. Schluenzen, K. Schwarz, S. Sternberger,
P. Suchowski, Ch. Voss, Deutsches Elektronen-Synchrotron (DESY), Hamburg, Germany

Abstract

Effective data management is critical to ensuring that research data remains readily accessible, efficiently processed, and optimally usable. In this article, we will describe the design and implementation of the data management infrastructure at the European XFEL, which supports high-level data management services. The system architecture is organised into four layers of storage, each addressing specific challenges associated with data handling.

The first layer, known as the Online storage, acts as a high-speed cache to accommodate the extremely high data rates generated during experiments, up to 15 GB/s from a single scientific instrument. The second layer, High-Performance Storage, supports both real-time data processing during experiments and offline post-experiment analysis. These two layers are interconnected through a single InfiniBand fabric, which is linked by a 4.4 km long, 1 Tb/s connection, facilitating rapid data transfer from the European XFEL experiment hall to the DESY computing centre.

The third layer, Mass Storage, significantly expands the system's capacity, enabling mid-term data access for more detailed analyses. Finally, the Tape Archive layer ensures secure, long-term preservation of data for a minimum retention period of 10 years. The high-performance and mass storage systems are connected to computing cluster, allowing users to conduct both near-online and offline data analysis, or export data from the European XFEL facility as needed.

The European XFEL's data management infrastructure is capable of handling and processing up to 2 PB of data per day, showcasing the exceptional performance and reliability of the system and its associated sub-services.

INTRODUCTION

The European XFEL Facility is among the world's most advanced sources of pulsed, extremely intense, and coherent radiation in both the hard and soft X-ray regimes. These unique characteristics make it highly attractive to a wide range of scientific communities, enabling diverse experimental research. The facility spans a total length of 3.4 km, stretching from the DESY campus in Hamburg to the town of Schenefeld in Schleswig-Holstein, where the experimental stations are located. At present, experiments can be conducted using one of seven[1] available instruments, each powered by one of three self-amplified spontaneous emission (SASE) X-ray sources. The configuration of the scientific instruments and SASE sources is shown in Fig. 1. It is important to note that, at any given moment, only a single instrument is in operation on a SASE. Depending on the specific SASE, experiments are conducted either in a continuous 24-hour mode or in a 16-hour experimental mode followed by 8 hours dedicated to commissioning or in-house research on other instruments. This means that all resources are dedicated exclusively to the current experiment, and the facility operates 24 hours a day.

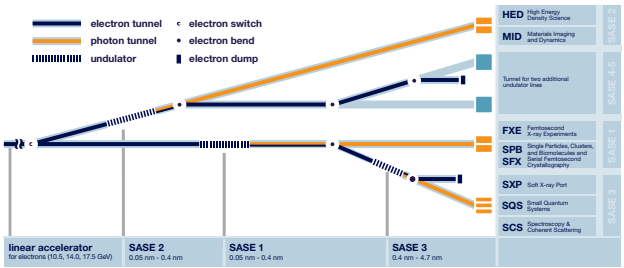

Figure 1: Beamlines and instruments of the European XFEL.

EXPERIMENT DATA FLOW

At the European XFEL, a wide variety of data sources generate data. These sources include sensors, cameras, motors, and both 1D and 2D detectors. Detector data is transferred over a dedicated network to data acquisition (DAQ) servers, where it is aggregated by separate DAQ devices called data aggregators. Typically, multiple data aggregators run on a single server, with each aggregator producing one HDF5 [1] file that is written to disk. In parallel with writing to storage, the data is also sent over the network to the online computing cluster for real-time processing. The period during which data is collected is called a run. Each experiment consists of multiple runs, and the duration of a run depends on the experimental technique being used. Typically, a run lasts a few minutes, but runs longer than 30 minutes also occur. Breaks between runs can be used to change a sample or adjust other critical parameters. Once the setup is complete and stable, the pause between runs can be reduced to just a few seconds. As a result, the high load generated by the DAQ system on the file system becomes nearly continuous during active data collection. For example, when the AGIPD [2] detector is used in a run with other smaller detectors and data sources, a throughput of about 15 GB/s is generated from all servers

* janusz.malka@xfel.eu
[1] The eighth instrument is under construction.

ISBN: 978-3-95450-255-4 ISSN: 2226-0358 doi: 10.18429/JACoW-ICALEPCS2025-THCR001

to the underlying storage system. This often results in a few petabytes of data generated from a single experiment.

Each run is registered at its start in the European XFEL Data Management Portal (myMdC) [3], along with descriptive metadata such as sample name and run type, as well as transient status information like the run state. After completion, myMdC is updated with further details, including run size and a reference to the current online data repository. The team performing the experiment can then assess the data quality of the run guided by various monitoring tools such as online previews and real-time data analysis running on the online computing cluster, which receives data from the DAQ system over the network. Once the experimental setup is fully established and data collection is proceeding reliably, automatic run assessment may be configured to facilitate and enhance the process.

A positive data quality assessment in myMdC triggers, via a messaging system, the transfer of raw data files from the online storage system in the European XFEL experiment hall to the offline storage at the DESY Data Centre, across a 4 km network link. This data transfer enables further data processing, data analysis and long-term data preservation. The file copy service utilises the native data management policy engine of the underlying file system. Once the complete run is transferred to the offline storage, the copy service notifies myMdC and registers the run in the offline repository. A notification is also sent to a message broker, which is consumed by the data processing service [4] to trigger the calibration pipeline. The status of the calibration process is also reported back to myMdC. The output of the calibration pipeline is stored in two different file systems: processed (proc) data are stored together with the raw data on a file system optimised for storage of large files; calibration reports and auxiliary data, which are required, for example, to support data reproducibility, are stored separately for performance reasons. These reports can be directly downloaded from the specific run metadata page in myMdC.

In addition to triggering offline data processing, the copy service also initiates the creation of a third copy of the raw data in a mass storage system. Once this process is completed, the third copy is also registered in myMdC. Finally, each raw file is stored in the tape archive through an asynchronous process governed by mass storage policies. Registration of the run in myMdC is performed only after successful completion of the tape copy for all files associated with the run.

Data Retention

The European XFEL follows a two-stage data copy principle within its four-layer data management model. According to this model, raw data is not removed from the online storage (Layer 1) until it has been successfully copied, with checksum validation, to the mass storage system (Layer 3). Likewise, data is not removed from offline storage (Layer 2) until it has been copied to the tape archive (Layer 4). This principle applies to raw data under the embargo period, which is defined as three years. As a result, the online storage must have sufficient capacity to: host the data during the experiment, typically one week, and retain the data until it has been successfully transferred to mass storage. The asynchronous process that handles data copying to tape must be fast and reliable enough to allow the timely deletion of raw data from the offline storage layer.

The overall data retention model is based on the European XFEL's Scientific Data Policy [5]. In addition to the rules defined in this policy, other factors also influence data retention, such as the specific needs of individual experiments. These needs can be fully addressed within the Data Management Plan (DMP), provided they remain within the boundaries set by the scientific data policy.

Other Data

In addition to the file system spaces used for raw and proc data, both of which are writable only by facility services, two dedicated storage spaces are available for users: usr and scratch.

The usr space is intended for the reliable storage of essential data, including analysis scripts, tools, and results. The available capacity is constrained by filesystem quotas, and data stored in this space is safeguarded through regular filesystem snapshots and backups. Prior to beamtime, when access to online resources is not yet available, users can test their analysis workflows and tools within the offline computing environment. During beamtime, data are synchronised between the offline storage at the DESY Data Centre and the online storage at the European XFEL experiment hall. Following the experiment, the usr space on the online system is cleared to free resources for subsequent experiments; however, the offline usr space is retained for the duration of the embargo period and subsequently made immutable to ensure the integrity of its contents.

In addition to the usr space, there is a scratch space available for users. This area is intended only for storing temporary files. No data protection mechanisms (e.g., backups or snapshots) are provided, and the offline scratch space is not synchronised with the online environment.

Together with the raw and proc spaces, the usr and scratch spaces are automatically created by the system based on entries in myMdC, usually a few weeks before the experiment. However, if access is needed earlier, it can be arranged in advance through an agreement in a Data Management Plan.

Access to experiment data spaces is granted using group-based access control lists (ACLs). Experiment team groups are managed through myMdC, while facility support and service groups are centrally managed outside of myMdC.

DATA ARCHITECTURE MODEL

The data architecture's four layers mentioned above are shown in Fig. 2, and will be discussed in more detail in the sections below.

ISBN: 978-3-95450-255-4 ISSN: 2226-0358 doi: 10.18429/JACoW-ICALEPCS2025-THCR001

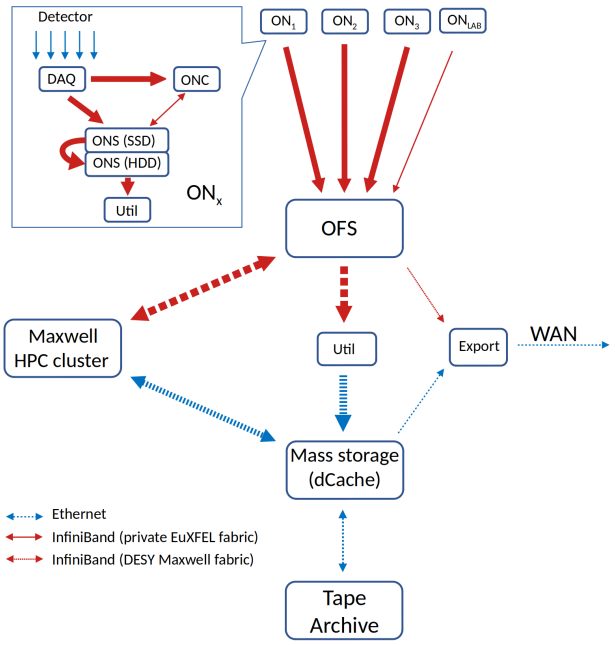

Figure 2: Storage and computing infrastructure diagram. The red arrows show the direction of the data flow in EuXFEL private InfiniBand fabric, the dotted red arrows represent the data flows in DESY Maxwell InfiniBand fabric and the blue arrows data traffic over Ethernet. DAQ - Data Acquisition cluster, ONC - Online Computing cluster, ONS - Online Storage cluster, OFS - High-performance Offline Storage, Mass storage (dCache), Tape archive, and Maxwell HPC cluster are indicated by the blue rounded rectangles, together with utility nodes performing administration tasks and actual data movement.

Online

Each SASE is equipped with a dedicated data infrastructure. Since only one experiment is scheduled per SASE at any given time, exclusive access to computing and storage resources is granted for all active experiments.

The Online Storage (ONS) clusters constitute the foundation of the online environment. Each ONS cluster is based on the IBM Storage Scale System [6], and is composed of two subsystems: one built on high-speed NVMe drives, and the other on HDD drives. The NVMe-based subsystem offers high performance, ideal for handling bursts of I/O operations and simultaneous accesses, while the HDD-based subsystem provides the capacity to store data from at least one full experiment day, with the target capacity of one week.

Several file systems are hosted on the ONS cluster, each configured for specific use cases, including: user home directories, software repositories, calibration constants, raw experiment data and intermediate results from online or near-online processing and analysis. The file system for experiment data is backed by three separate storage pools: System Pool – Stores file system metadata and uses NVMe drives for high responsiveness; Cache Pool – Holds the most recent experiment data, also on NVMe drives, enabling high-speed data ingest and read access; Data Pool – Stores less frequently accessed (colder) data, built on HDD drives for higher capacity. An automatic migration policy process moves data from the cache pool to the data pool to prevent the cache from filling up. In parallel, the raw data collected during experiments is also copied to a high-performance offline storage cluster at the DESY Data Centre. At any moment, the ONS cluster may simultaneously be handling: data ingest from the experiment; cache-to-data pool migration; data copy to offline storage. These processes compete for available I/O resources, but the system is designed to prioritise data ingest, limiting IOPS available to the other two processes to maintain real-time performance. Typical measured data rates under simultaneous load are: 15 GB/s for data ingest, 20 GB/s for data migration (drain), 30 GB/s for data copy to offline storage. The file systems hosted on the ONS cluster are exported to two client clusters: the Data Acquisition (DAQ) cluster and the Online Computing (ONC) cluster.

The DAQ cluster allows the data acquisition system to aggregate and store experiment data directly on the ONS cluster. In parallel, the DAQ streams data over the InfiniBand network to the ONC cluster, which consists of several GPU and CPU-only nodes. If needed, preliminary data corrections are performed on ONC nodes, enabling live data previews and fast feedback analysis.

High-Performance Storage (Offline)

Much larger than the Online Storage Clusters (ONS), the Offline Storage Cluster (OFS) is the core high-performance data storage system and plays a central role in the facility's data management architecture. Unlike a ONS, which serves as an exclusive resource for the active experiment, the OFS is shared among the ongoing experiment, experiments for which beamtime has concluded and analysis is still in progress, as well as activities related to experiment preparation and commissioning. The OFS hosts both raw and processed data, as well as user-generated data essential for analysis workflows, making it the most heavily used and performance-critical system in the entire storage hierarchy. Designed for scalability, speed, and resilience, the OFS enables high-throughput, low-latency access to frequently used datasets. It supports concurrent high-bandwidth workloads, including: copy-in from the Online Storage (ONS), data processing and user analysis jobs. Its performance is critical to enabling fast feedback during the experiment and ensuring rapid data availability post-experiment. The system has demonstrated sustained read performance exceeding 175 GiB/s, supporting highly parallel workloads across users and services. The offline storage cluster is built on IBM Storage Scale System (ESS/SSS) technology, a high-performance, scalable solution based on IBM Spectrum Scale (formerly GPFS). The OFS leverages both NVMe-based nodes for metadata operations and HDD-based capacity nodes for large-scale data storage. The system currently consists of 13 building blocks (BB), each BB consists of two head nodes

ISBN: 978-3-95450-255-4 ISSN: 2226-0358 doi: 10.18429/JACoW-ICALEPCS2025-THCR001

and several JBODs (Just a Bunch of Disks) enclosures with up to hundreds of disks, connected to them. Together, these provide a total usable capacity of 63 petabytes, distributed across multiple GPFS file systems optimised for different workloads and exported to the DESY High-Performance Computing Cluster (Maxwell).

The high-performance storage systems, both online and offline, benefit from petabyte-scale density in compact rack footprints, maximising space utilisation. With NVMe and high-speed RDMA[2]-capable networks (InfiniBand), the systems can sustain extremely high read/write speeds. Designed for distributed workloads, they allow near-linear scaling of performance and capacity by adding more building blocks. The modular nature of the system allows the OFS to evolve. When older building blocks reach end-of-life or go out of support, they are phased out and replaced with newer, more performant components, often with increased capacity and throughput. This rolling upgrade model ensures continuous availability and enables organic growth as scientific demands increase. Moreover, the system provides end-to-end data protection through checksums and ensures reliability and fast rebuild times through the dispersed data layout and erasure coding of Storage Scale RAID. It also integrates tightly with Spectrum Scale, allowing fine-grained filesystem policies, tiered storage, and Quality of Service (QoS) control, as well as integration with external compute environments through native GPFS export or via central export services, thereby enabling data export outside the facility.

Mass Storage

The mass storage infrastructure is the third layer in the data management architecture of the European XFEL facility and is based on dCache technology [7]. It currently represents the largest storage capacity (120 PB) within the system. Its primary role is to provide long-term storage for raw data and portions of processed data, as well as to serve as a staging area for archiving data onto the tape repository.

In contrast to the Online and Offline clusters, the mass storage system is built on commodity hardware, connected via an Ethernet network, and utilises nearline SAS disks. At its core, dCache is built on a micro-service architecture, with components responsible for managing: storage servers (pools), client access, internal services for authentication and authorisation, namespace management, interaction with tertiary (tape) storage systems and storage node selection.

Importantly, dCache does not manage physical storage resilience itself; rather, it relies on the underlying infrastructure to protect against hardware failures. However, it can be configured to store multiple replicas of files across different storage nodes for redundancy. Each storage node, or pool, operates independently, enabling horizontal scalability. Overall storage capacity can be increased simply by adding new pools.

The dCache instance for European XFEL is logically divided into two main sections: a disk-only area and an area connected to the DESY tape infrastructure. The disk-only area is intended to store additional files required for data analysis. To ensure reliability, each file is replicated on another storage server, providing resilience in the event of a complete server failure. However, this area constitutes only about 1% of the total dCache capacity.

The tape-backed area is primarily used for storing raw experiment data. Each storage pool in this area is connected to the DESY tape system and follows a classical Hierarchical Storage Management (HSM) model. However, unlike traditional HSM systems, users cannot initiate restores themselves. For example, clients accessing files via NFS 4.1 will receive a permission denied error if a file resides only on tape. Staging requests must be triggered by European XFEL data managers or the DESY dCache operations team, typically after a preliminary assessment.

Each file is assigned a storage class based on specific tags applied to its directory, such as the instrument type and run period. These tags allow the tape system to logically organise data, grouping files from the same experiment onto a common tape set, which improves space efficiency and simplifies future retrievals.

File locality and retention are controlled using Quality of Service (QoS) levels, as implemented in dCache. By default, all raw data files are set to a QoS of disk+tape, meaning they reside on both disk and tape. If disk space needs to be reclaimed, European XFEL data managers can change the QoS to tape, marking the files as cacheable and allowing dCache to remove them from disk when needed.

If users later request these files, the QoS can be reset to disk+tape, which triggers a restore from tape and ensures the files are retained on disk once again. These transitions are handled programmatically via the dCache REST API.

Tape Archive

The European XFEL uses tape storage as the fourth layer in its data management architecture, serving as the primary solution for long-term archival and preservation of scientific data. Tape technology remains a strategic choice in large-scale scientific facilities due to its unique combination of benefits: high capacity, low cost per terabyte, long durability, and energy efficiency – especially compared to disk-based systems – albeit at the cost of increased access latency.

At the hardware level, the facility relies on the IBM TS4500 tape library, a highly scalable and modular enterprise solution capable of supporting thousands of tape cartridges. The library hosts multiple generations of tape media, including: LTO-8 cartridges with up to 12 TB native capacity, LTO-9 cartridges offering 18 TB native capacity, and IBM 3592 JE cartridges providing 20 TB native capacity. Each tape drive offers transfer rates of approximately 400 MB/s. By scaling the number of available tape drives the system can be adapted to meet the high-throughput demands of data archiving workflows while keeping operational costs and power consumption low.

The tape archive is orchestrated by the CERN Tape Archive (CTA) [8], a modern, high-performance software

[2] Remote Direct Memory Access

ISBN: 978-3-95450-255-4 ISSN: 2226-0358 doi: 10.18429/JACoW-ICALEPCS2025-THCR001

solution developed by CERN to manage large-scale archival systems. CTA is designed to optimise tape usage efficiency, improve file retrieval times, and support intelligent data placement and scheduling across available tape resources. CTA seamlessly integrates tape archive with dCache, the mass storage system used in the third layer. This integration ensures a smooth and automated transition of data from disk-based storage to tape. Once files are copied to the dCache tape-backed pools, CTA handles the migration to tape, monitors the status and integrity of tape copies, and coordinates recall (staging) operations when data needs to be brought back to disk. This setup allows European XFEL to:

- preserve raw and processed data for the long term in a cost-effective and environmentally sustainable way,
- ensure data integrity, as files written to tape undergo checksum verification, support flexible retrieval policies, managed by data managers using QoS controls in dCache,
- scale archiving capacity over time by adding new cartridges and drives without significant architectural changes.

By combining high-performance hardware with robust software orchestration, the tape storage infrastructure forms a critical component in the facility's strategy for secure, efficient, and policy-compliant scientific data preservation.

Maxwell High-Performance Computing Cluster

Data analysis, processing and simulations related to experiments at the European XFEL are almost exclusively performed on the Maxwell HPC cluster. The cluster has all the ingredients of a typical HPC platform with SLURM-scheduling, low-latency InfiniBand backbone, cluster file-systems, and a total of 940 compute nodes with a theoretical peak performance of 4000 TFlops.

The Maxwell cluster is, however, in contrast to conventional HPC systems, extremely heterogeneous, thereby reflecting the heterogeneity of the DESY campus, which hosts a very large number of independent institutions covering a wide area of scientific fields. Rather than each institution operating its own compute cluster, the Maxwell cluster enables institutions to contribute compute resources, which are then conveniently shared by all users of the Maxwell cluster. This cooperative model leads to better resource utilisation, allowing users to allocate a substantially larger number of compute nodes than an institutional platform could offer, but comes at the price of a more heterogeneous compute environment.

European XFEL is one and by far the largest member of the Maxwell platform, contributing 439 CPU-only and 30 GPU nodes or roughly 1200 TFlops, which account for about 30% of the cluster's compute power. The compute nodes are essentially distributed over two general partitions serving staff members and users, and three partitions dedicated to compute tasks and calibrations during user experiments. The Slurm partitions are overlapping, and the resources can be reserved dynamically to cope with the needs of running experiments.

The primary focus of the Maxwell cluster to serve primarily photon science data processing leads to a somewhat unusual HPC workload distribution. Out of roughly 4 million batch jobs per year, more than 90% of the jobs are single-node jobs. Jobs running in the European XFEL context contribute more than 50% of jobs and, quite remarkably, for almost all many-node jobs (jobs with more than 16 nodes). Serving running experiments with the aim to enable large-scale data processing requires immediate availability of sufficient compute resources. Guaranteed immediate access to compute resources is not perfectly compliant with the scheduling configuration (FIFO+Backfilling). Compute resources for running experiments are hence allocated from highly prioritised partitions, using node-reservations to distribute resources across concurrent experiments according to individual computational requirements.

The Maxwell cluster also offers a customised JupyterHub instance, which allows selecting Slurm partitions and reservations to launch a Jupyter server as regular batch jobs in customizable environments. Not surprisingly, Jupyter notebooks became quite popular among scientific users: more than 500 of the 2800 Maxwell users occasionally launch Jupyter notebooks. However, Jupyter jobs account for only 1% of all batch jobs and consume less than 1% of compute resources on average. The growing use of Jupyter notebooks for online analysis during experiments makes the Jupyter ecosystem a critical infrastructure.

Sustainability

In addition to ongoing research on computing efficiency, supported by complementary activities across various scientific communities, such as Particle Physics and High-Performance Computing (HPC), recent initiatives have focused on raising awareness of energy consumption. This includes generating per-user and per-job summaries of power usage and the associated greenhouse gas emissions, which are currently being prepared for deployment on the Maxwell HPC cluster.

By maintaining a single, shared compute cluster for distinct workloads, such as real-time data processing during experiments and traditional batch processing, as described in the previous section, we avoid the need to deploy large-scale, dedicated resources for each use case. This strategy enhances overall utilisation efficiency and reduces power consumption.

Plans are underway to implement a comprehensive power metering system for all data centre infrastructure, with work initiated in 2024 through externally funded projects. An example is the RF2.0 project (Research Facilities 2.0, 2024) [9], which develops sustainable, energy-efficient solutions for particle accelerators and research infrastructures. Within this framework, the DESY IT department is investigating dynamic resource provisioning that adapts to user demand and green energy availability, with future workshops expected to provide training on preparing jobs for such environments.

ISBN: 978-3-95450-255-4 ISSN: 2226-0358 doi: 10.18429/JACoW-ICALEPCS2025-THCR001

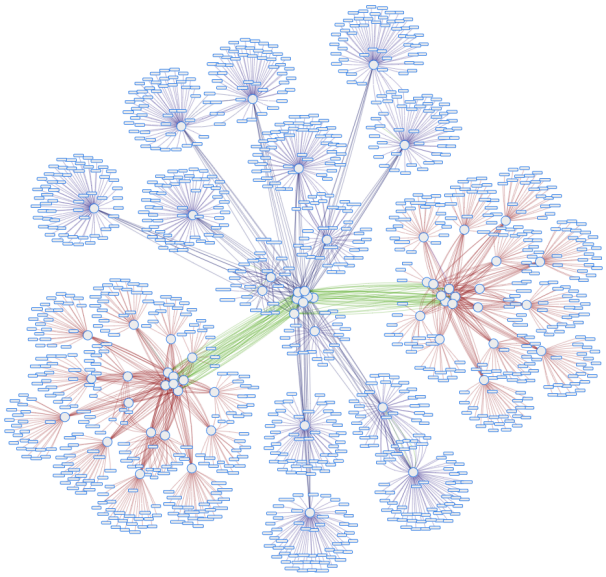

Figure 3: InfiniBand fabric of the Maxwell cluster. An HDR (violet) root layer connects two EDR (green) layers (blobs at top and bottom), which then connect nodes via FDR (red) leaf switches. The smaller "branches" are HDR switches that connect nodes via HDR100 splitter cables.

Network

The IBM Storage Scale clusters, namely the ONS and OFS storage systems, as well as the Maxwell compute cluster, the online compute clusters, and the DAQ clusters, all utilise InfiniBand as a high-speed, low-latency networking fabric with RDMA enabled.

To isolate the time-critical, online data acquisition environment from the shared Maxwell compute cluster (used by multiple scientific communities), two independent InfiniBand fabrics are deployed. The DESY Maxwell fabric (Fig. 3) connects all Maxwell compute nodes and all OFS storage nodes. A separate private InfiniBand fabric links the online clusters to the OFS storage nodes, providing a high-speed connection with an aggregate bandwidth of approximately 1 TB/s, ensuring low-latency and high-throughput data transfer over 4 km, for time-critical operations. In this way, the OFS storage nodes function as a bridge between the online and offline environments. Each node is connected to a separate Infiniband fabric, allowing data stored on the nodes to be accessed from both fabrics without directly linking the fabrics themselves.

Similar to the storage infrastructure, the InfiniBand network is also a "rolling" installation. As components are gradually upgraded or added, the fabric comprises multiple generations of InfiniBand technology, ranging from FDR[3] and EDR[4] to HDR[5]. The introduction of adaptive routing has significantly improved performance, particularly on the private fabric, by dynamically balancing traffic across multiple connections between switches and ensuring efficient utilisation of available links.

While the primary workload over the InfiniBand fabrics is high-throughput file I/O, other performance-critical applications also benefit. On the private fabric, data is streamed directly from DAQ nodes to online compute nodes, enabling fast, low-latency data preview. In the Maxwell fabric, MPI-based jobs use InfiniBand for efficient inter-node communication. Meanwhile, administrative tasks, user logins, and dCache-related traffic are handled via a separate Ethernet network.

Monitoring

Effective service and performance monitoring is essential for detecting and minimising outages, as well as for guiding future infrastructure scaling. This section outlines the monitoring strategies applied across the system.

A central Icinga 2 [10] installation in DESY Data Centre is responsible for service monitoring of the EuXFEL data management infrastructure. The Icinga 2 Agent executes service checks to determine the health of the service and reports the result to the Icinga 2 instance. The following service checks monitor the infrastructure:

- System Health: CPU, memory and disk usage checks
- Hardware Health: Memory, power supply and hard drive failure checks
- IBM Storage Scale Health: Daemon availability, long-running waiters (Deadlock detection), pool usage

Service checks are written in various programming languages (C, Bash, Perl, Python).

Hardware-related failures occurring during normal business hours are handled by the local operations team. Critical service failures outside of business hours are addressed by on-call duty personnel. More complex issues, such as Storage Scale-related interruptions, are escalated to the storage administration team.

For performance metrics gathering, two different tools are used:

- IBM Storage Scale Performance Monitoring Tool,
- Telegraf for metric collection and Graphite for storing numeric time-series data.

The IBM Storage Scale Performance Monitoring Tool consists of two components:

- Collector: Store and query of IBM Storage Scale metrics,
- Sensor: Collect IBM Storage Scale metrics from a host.

Sensors collect various Storage Scale-related metrics, such as I/O throughput, latency, IOPS, block usage for filesets and storage pools, as well as the number of waiters. These metrics are recorded by a collector and made available for querying.

Telegraf [11] is responsible for collecting several standard Linux metrics, like CPU, memory and disk usage metrics, throughput of network interfaces and also the power consumption. While there is an overlap between the Icinga 2 system health checks and system metrics, they are stored in different databases with different capabilities. The Graphite

[3] FDR - 56 Gb/s
[4] EDR - 100 Gb/s
[5] HDR - 200 Gb/s

ISBN: 978-3-95450-255-4 ISSN: 2226-0358 doi: 10.18429/JACoW-ICALEPCS2025-THCR001

[12] time-series database is specialised for numeric data and supports aggregation of metric values. This allows for reducing the storage consumption at the price of reduced precision for historic values.

In order to visualise all collected metrics, Grafana [13] is used. Although both Graphite and the IBM Storage Scale Collector offer their own visualisation capabilities, using Grafana provides significant advantages. It enables the display of metrics from both time-series databases on a single dashboard. Grafana natively supports Graphite, while visualising IBM Storage Scale metrics requires the IBM Spectrum Scale Bridge for Grafana [14]. This bridge translates OpenTSDB queries from Grafana into the query language understood by the collector.

In addition to the central monitoring provided by the DESY Data Centre, all components located in the European XFEL experiment hall are also monitored by a dedicated local monitoring system based on Zabbix [15]. This local monitoring is crucial, as it enables the on-site team to quickly detect and respond to issues that could immediately impact user experiments. By providing real-time, targeted oversight, the local system ensures faster intervention and minimises the risk of an experiment downtime.

OUTLOOK

Taking advantage of the ongoing long-term maintenance period at the European XFEL, a comprehensive review of the data management system is now underway. This review aims to ensure that the infrastructure and workflows remain robust, scalable, and capable of supporting upcoming operational challenges, including the integration of new high-performance detectors and more demanding modes of operation. The review is being conducted from two key perspectives:

- Data Management Workflows – Evaluating and optimising end-to-end workflows to enhance efficiency, scalability, and user experience, while ensuring data integrity, traceability, and compliance with the facility's scientific data policy.
- Hardware Infrastructure – Investigating new hardware solutions to meet increasing demands for data throughput, storage capacity, and real-time processing required by future experimental configurations.

This forward-looking effort will lay the foundation for a next-generation data management ecosystem, designed to meet the evolving needs of users, experiments, and facility operations in the years to come.

CONCLUSION

The infrastructure and services presented in this paper have been in production since 2017. Since then, continuous optimisation and automation have significantly reduced the administrative burden of operating the system. At the same time, the proven rate and capacity for storing and processing detector data have increased by a factor of 2–3, closely following advancements in the underlying technologies.

While the current system has demonstrated reliable performance and scalability, the evolving scientific landscape, characterised by increasing data volumes, more complex experiments, and higher detector performance, demands a proactive approach to future readiness.

ACKNOWLEDGEMENTS

We wish to acknowledge the help provided by the instrument scientists and data experts of European XFEL GmbH and DESY-IT colleagues not mentioned in the author lists. We would also like to show our deep appreciation to our business partners who are helping us provide an excellent data service for users of our facility.

REFERENCES

[1] The HDF Group. Hierarchical Data Format, https://www.hdfgroup.org/HDF5

[2] A. Allahgholi *et al.*, "The Adaptive Gain Integrating Pixel Detector at the European XFEL", *J. Synchrotron Radiat.*, vol. 26, no. 1, pp. 74–82, Jan. 2019. doi:10.1107/s1600577518016077

[3] L. Maia *et al.*, "Integrated and automated data management at European XFEL empowered by myMdC metadata catalogue", presented at ICALEPCS'25, Chicago, USA, Sep. 2025, paper THCR005, this conference.

[4] P. Schmidt *et al.*, "Turning European XFEL raw data into user data", *Front. Phys.*, vol. 11, Jan. 2024. doi:10.3389/fphy.2023.1321524

[5] Scientific Data Policy, https://www.xfel.eu/users/policies/index_eng.html

[6] F. B. Schmuck, and R. L. Haskin, "GPFS: A Shared-Disk File System for Large Computing Clusters", In *Proc. FAST 2002 Conf. File Storage Technol.*, Monterey, CA, USA, Jan 2022, pp. 231—244.

[7] Tigran Mkrtchyan *et al.*, "dCache: Inter-disciplinary storage system", EPJ Web of Conferences, vol. 251, p. 02010, 2021. doi:10.1051/epjconf/202125102010

[8] The CERN Tape Archive, https://cta.web.cern.ch/cta

[9] Research Facilities 2.0., https://rf20.eu

[10] Icinga, https://icinga.com

[11] Telegraf, https://www.influxdata.com/time-series-platform/telegraf

[12] Graphite, https://graphiteapp.org/

[13] Grafana, https://grafana.com/

[14] IBM Spectrum Scale Bridge for Grafana, https://github.com/IBM/ibm-spectrum-scale-bridge-for-grafana

[15] Zabbix, https://www.zabbix.com