frontiers | Frontiers in Physics

TYPE Technology and Code
PUBLISHED 08 January 2024
DOI 10.3389/fphy.2023.1321524

OPEN ACCESS

EDITED BY
Iain Sedgwick,
Rutherford Appleton Laboratory,
United Kingdom

REVIEWED BY
Naoto Yagi,
Japan Synchrotron Radiation Research Institute, Japan
Suren Chilingaryan,
Karlsruhe Institute of Technology (KIT),
Germany

*CORRESPONDENCE
Philipp Schmidt,
philipp.schmidt@xfel.eu

RECEIVED 14 October 2023
ACCEPTED 14 December 2023
PUBLISHED 08 January 2024

CITATION
Schmidt P, Ahmed K, Danilevski C, Hammer D, Rosca R, Kluyver T, Michelat T, Sobolev E, Gelisio L, Maia L, Manetti M, Malka J, Wrona K, Sztuk-Dambietz J, Rovensky V, Ramilli M, Duarte N, Lomidze D, Dourki I, Yousef H, Senfftleben B, Meyer O, Turcato M, Hauf S and Aplin S (2024), Turning European XFEL raw data into user data.
Front. Phys. 11:1321524.
doi: 10.3389/fphy.2023.1321524

Turning European XFEL raw data into user data

Philipp Schmidt*, Karim Ahmed, Cyril Danilevski, David Hammer, Robert Rosca, Thomas Kluyver, Thomas Michelat, Egor Sobolev, Luca Gelisio, Luis Maia, Maurizio Manetti, Janusz Malka, Krzysztof Wrona, Jolanta Sztuk-Dambietz, Vratko Rovensky, Marco Ramilli, Nuno Duarte, David Lomidze, Ibrahym Dourki, Hazem Yousef, Björn Senfftleben, Olivier Meyer, Monica Turcato, Steffen Hauf and Steve Aplin

European XFEL, Schenefeld, Germany

The European X-ray Free Electron Laser is a research facility located close to Hamburg, offering X-ray pulses with ultra-high brilliance and femtosecond duration at megahertz repetition rates. The detection systems necessary to unlock the full scientific potential made possible by this machine poses considerable challenges both in terms of data volume and rate, as well as the interpretation of their recorded signal. To provide optimal data quality, expert and detector-specific knowledge not easily accessible to external facility users is essential, and its implementation must cope with the generated volumes. We therefore aim to perform these preparatory processing steps and offer users a dataset suitable for further analysis as the primary data product. This work describes the machinery and workflows providing this data to users in an automatic, configurable and reproducible manner, both online during the experiment, and offline for scientific analysis afterward on the way to publication.

KEYWORDS

photon science, free electron laser, big data, data analysis, data quality, detector correction, HPC

1 Introduction

The advent of X-ray free electron laser sources and in particular their recent advance into data rates in the kHz regime continues to push the boundaries of data analysis techniques. The *European X-ray Free Electron Laser* (*European XFEL*) [1, 2], in operation since 2017, is such a facility located in the area of Hamburg, Germany. Its superconducting linear accelerator produces electron bunches with an energy of up to 17.5 GeV in a unique burst mode time structure as shown in Figure 1. The resulting X-ray pulses are arranged in trains of up to 2,700 pulses, with trains arriving at a rate of 10 Hz. Within each train, the pulses are separated by as little as 222 ns, which is equivalent to an intra-train repetition rate of up to 4.5 MHz. They are currently delivered to three beamlines in parallel covering the soft X-ray to hard X-ray photon energy regime. At each beamline, up to three instrument endstations are installed spanning a large range of different experiment techniques.

This unique train-pulse time structure offers the benefit of high pulse energies and small wavelengths at comparably high repetition rates, but incurs additional challenges in terms of detector technologies able to keep up with this intra-pulse distance and duty cycle. These challenges led to the development of multiple custom X-ray 2D imaging cameras—*AGIPD*

[3], *LPD* [4], and *DSSC* [5]—capable of capturing up to 8,000 frames per second at the pulse repetition rate of 4.5 MHz, by using large memory cell arrays, while also being able to cover large dynamic ranges of photon intensities. Achieving optimal data quality with these detectors requires intimate technical knowledge and the sheer data volume they produce means that processing must be highly scalable. For users, this complexity can impose a high barrier of entry to make use of their data and to achieve scientific results for their proposals.

We are aware of the impact of this complexity on data analysis for users, and aim to offer the data taken during a proposal in a form useful in the scientific context of its experiment. This form of data we call *user data*, and it is provided in the same data format alongside the original *raw data* as it was acquired by detectors. What constitutes user data can be highly variable from experiment to experiment and depends on the technique, experimental conditions, and of course detectors used. It may range from image corrections per pixel, over clustering or integration of neighbouring intensities, to event reconstruction across correlated signal sources. To this end, established and essential data preparation steps are offered as a service running on the facility infrastructure, where they can be efficiently and reproducibly applied at scale. Data processing is provided both for real time applications during an experiment—delivering data streams at latencies of a few seconds or less—as well as for exhaustive processing of data recorded to disk with a focus on completeness, precision, and reproducibility, scaling to up the petabyte regime for single experiments. Rather than replacing the raw data product, however, these systems are designed to maintain configurability and integrate into user workflows with custom adaptions for each scientific application. This article reports on the general infrastructure and systems developed for this purpose as well as the specific detectors and methods it was applied to over the past 6 years of facility operation. Their impact on user experiments and the facility is discussed, leading up to a comparison with the originally envisioned concept and an outlook into upcoming developments.

2 Methods

Data processing at European XFEL is generally separated into the two paradigms of *online* and *offline*. This separation is reflected in the facility-side machinery and tools that provide user data.

Online processing happens during the experiment on the direct data streams from the detectors and other acquisition devices, it is near real-time and provides immediate feedback and monitoring to steer the experiment. Given these requirements, and the key role online analysis plays in the success of an experiment, its primary focus is low latency to provide analysis results within a few seconds or less with high reliability for the operator. To this end, it may only operate on a relevant subset of data and employ less sophisticated algorithms to guarantee a result at the highest possible throughput, potentially at the cost of accuracy.

Offline processing on the other hand operates on data stored in persistent files for deeper data exploration and analysis. This may take place minutes to hours after acquisition to guide experimental decisions and extend for months after the experiment is concluded until a clear scientific picture emerges. Such analyses aim at accuracy, completeness, and reproducibility and, as such, they are generally more efficient at scales which enable the use of computationally expensive methods.

Distinct solutions have been developed to optimally serve both of these requirements, built around streams and files respectively, with a common ecosystem for tracking metadata. Both operate on the same input of raw data, but are strictly split on the volatility of their results. Those obtained from online processing are generally not stored to disk to prevent any compromise in data quality or reproducibility, which may result from their *a priori* configuration or performance requirements. Instead, any permanent results are produced through the offline processing system to allow for continuous tuning of parameters and behaviour for optimal and traceable results.

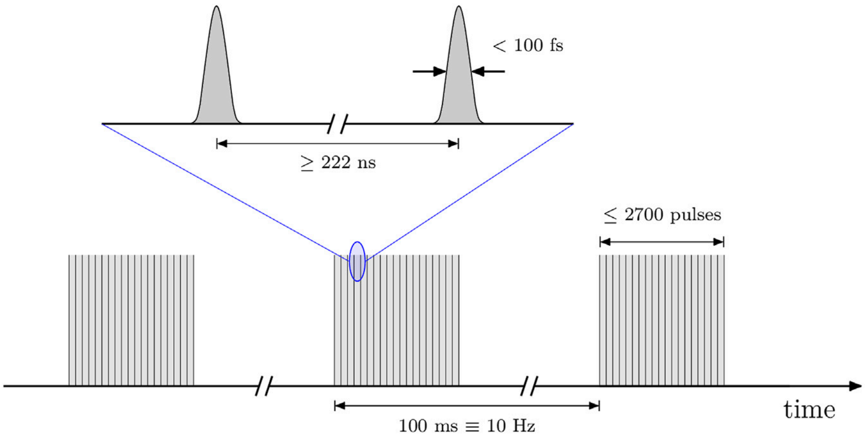

FIGURE 1
The time structure of European XFEL consisting of pulse trains with up to 2,700 individual pulses at a train repetition rate of 10 Hz. Within one train, the spacing between bunches is in the order of several hundred nanoseconds, while each bunch by itself has a length of typically less than 100 fs [1].

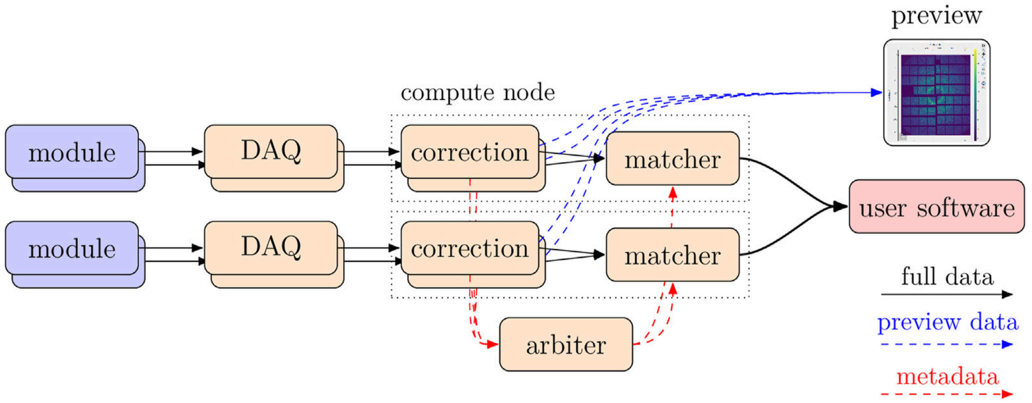

FIGURE 2
Data flows in the online processing pipeline for multi-module 2D pixel detectors with up to 16 modules. Detector hardware sends raw data through the data acquisition system (DAQ) to correction software devices. Each compute node typically hosts up to four correction devices called a group, each processing the data of a single module. From the correction devices, the low-latency preview stream is limited to a single frame and provided for immediate operator display. The full data stream for further downstream analysis can include additional metadata introduced by calibration add-ons running in the correction device. This metadata across all modules may be used for data reduction decisions in an arbiter device. In each group, a matcher aggregates the individual data streams taking this data reduction feedback into account.

2.1 Online data processing

Providing an interpretable result in near real-time during an experiment is often essential to performing successful user beamtimes in photon science, in particular, to make the most efficient use of the short time allocated to each experiment. In addition, it serves a monitoring role for the experimental hardware and environment, helping to ensure safe and effective operation.

The facility-provided online processing system is integrated into *Karabo* [6], a distributed control system developed at European XFEL to address three main challenges: i) acquisition and processing of the large data volumes generated at the facility, ii) provision of global time synchronisation across most control variables, and iii) the flexibility required to efficiently control both static and highly dynamic setups in prototype and facility scale. Physical hardware is represented by corresponding Karabo devices written in C++ or Python, which may communicate with each other via a central message broker and direct point-to-point connections. Additional functionality can be provided in the form of pure software devices, as is the case with the online data processing system. The processing is performed in the online computing cluster (ONC), which consists of dedicated nodes for each of the three beamlines operating in parallel and located physically close to their endstations. This is equipped with datacenter-grade graphics processing units (GPUs), which can accelerate the most demanding processing steps.

The topology for online image correction for the large-area multi-module 2D imaging detectors is illustrated in Figure 2, as this presents the most demanding application due to the high data rates involved. A single Karabo endpoint device per physical detector module feeds data to a single correction device which performs processing per module. Multiple correction devices are grouped together on a single computing node to balance bandwidth and computing capabilities. From the correction devices, two separate types of output streams are provided: preview output and full data output.

The preview path is limited to at most a single frame per train via configurable reduction methods and provides quick feedback with minimal latency. It is geared for direct monitoring on screen and includes assembly steps of the individual detector modules into their physical geometry. The latency of the assembled preview—including corrections as well as assembly—is typically a few hundred milliseconds.

The full data path, on the other hand, delivers the complete data stream and can be tuned to best match the experimental analysis requirements to the available network performance. It may carry only a few specific detector modules of interest or assemble entire detector frames in a single stream, with modules grouped and processed together on the same machine as required to make optimal use of network bandwidth. The highest data rate among the currently used detectors is 9 Gbits per second for a single module and a total bandwidth of 140 Gbits per second for the entire assembly consisting of 16 modules. This output path is generally used for specialized real time analysis suites provided by facility users, which are tailored for each instrument via interfaces into the Karabo control system.

In addition to the built-in correction methods, custom processing code may be injected into the high-performance paths via so-called *correction add-ons*. In particular for implementations running on GPUs, this may take advantage of data already being present in device memory to perform further analysis after corrections. It should be noted that to preserve the monitoring aspect, the preview result always remains unchanged. An important application of the correction add-on mechanism is for the purpose of online data reduction. Any metadata generated by correction add-ons across all correction devices of a detector can be transported to a central arbiter device ahead of the actual detector data, where further custom code is executed in so-called reduction kernels. Here, the final decision can be made as to which data to include in the data stream available further downstream and make optimal use of the available bandwidth by minimizing the amount of data to be transferred and processed.

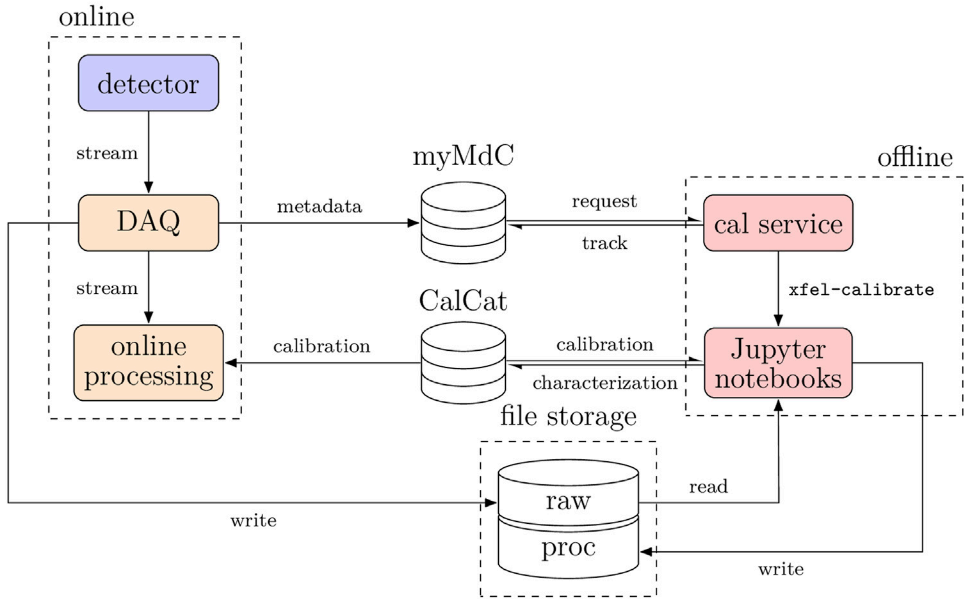

FIGURE 3
Overview of the services involved in the facility processing system at European XFEL. Detector data sent to the data acquisition system (DAQ) is sent as a stream to online processing systems and written to disk as raw data. For the files tracked by the myMdC application, requests can be sent to be calibration service to perform offline processing. This configures and runs Jupyter notebooks on the Maxwell compute cluster, interacting with the CalCat database to retrieve necessary calibration data or add new data after characterization. The resulting processed data is then stored and accessible alongside the raw data.

2.2 Offline data processing

Data saved to files constitutes the primary data product of beamtimes with more than 100 PB generated since European XFEL began user operation in 2017. It serves a critical role from early data exploration for decision-making during the beamtime to forming the basis of scientific publication. To address the FAIR [7] data principles and commitments of the common framework for scientific data management at photon and neutron facilities laid out by the PaN-data Europe Strategic Working Group [8], raw data is generally acquired via the facility data acquisition system (DAQ) and stored in the HDF5 format [9].

The raw data is complemented by data processed by the facility and saved in additional files alongside. This data is generated by a dedicated system either automatically upon the end of acquisition, or by explicit requests made through the data management portal for European XFEL users (myMdC). The portal is implemented as a web application and tracks all scientific data saved to disk with its physical storage location and metadata such as samples, techniques, and experiment types. Additionally, it includes administration of the experimental team, the electronic logbook, as well as management of digital object identifiers (DOIs) to the scientific data.

An overview of the service interactions and flows to generate this facility-processed data is provided in Figure 3. The requests triggered through *myMdC* are tracked and managed by the calibration service and run on the offline computing cluster Maxwell [10], which updates their status through the *myMdC* interface. The actual processing code at the heart of this system is implemented in *Jupyter* notebooks [11]. This allows the same code to scale from processing entire beamtimes automatically to manual, interactive execution on a selected subset of data for exploration or development. Each notebook is identified by an *action* it performs on data of a particular *detector* and is written in such a way to receive input values using nbparameterise [12]. These specify the input data and the intended output location, as well as any other parameters in terms of format or scientific context. The xfel-calibrate runtime is then used to divide the workload and run several copies of the processing at once via the SLURM workload scheduler [13], each copy running on a subset of the data, spread across multiple compute nodes to maximise the efficient use of time and resources. At the end, the executed notebooks are compiled into a report documenting the processing, including plots intended for diagnostics and to monitor data quality. Additional single notebooks may be run before and after this central processing step to prepare the environment or reduce the results further.

Next to the automatic processing steps performed on acquired data, this system is also used to implement the characterization and generate the necessary calibration data for some of the aforementioned processing steps, e.g., image corrections. This processing differs in the degree of automation and interactivity depending on how often it has to be repeated and its robustness. The calculation of gain factors to calibrate intensity in absolute units is generally done by manual invocation of xfel-calibrate with suitable parameters, often running the underlying processing notebook manually first to exploring the parameter space. An example of fully automated characterization is the determination of baseline offsets from *dark* data, i.e., data in the absence of an external stimulus on the detector. It is performed at least daily during operation and triggered through the *myMdC* web application.

In both cases, the location of the generated calibration data is centrally indexed in a database called the Calibration Catalogue (CalCat). It is queryable by the detector identifier, the point in time the characterization took place and the conditions the data is applicable for, such as the sensor temperature, bias voltage or integration time. This enables the retrieval of the most suitable available calibration data anytime a given detector was used. Queries to *CalCat* happen both during the experiment using the current detector conditions, as well as for any data taken prior, with the respective conditions at the time of measurement. The conditions are described by key-value pairs, which are generally scalar numbers and have been assigned an allowed deviation at characterization time. Detectors are uniquely labelled throughout the facility by the physical detector unit (PDU) identifier, which is independent of their physical location at a particular experiment. This allows calibration data to seamlessly follow a detector to wherever it is used at a particular time, as long as calibration data does not depend strongly on the environment. In the case of multi-module detectors, each of these modules represents a single *PDU* to facilitate maintenance or exchange of individual modules. The calibration data stored in this database can be readily compared to past values for the same conditions to allow for regular monitoring by experts.

An important aspect of facility-processed data is reproducibility, which in this context denotes the ability to recreate the same output given the same input at a later point in time. Reproducibility aims to ensure a level of confidence in the scientific results derived from such data. It also alleviates the need to archive processed data in the longer term, as it can be recreated from archived raw data if needed. Here, it is important to acknowledge that in general, running the same arbitrary code irrespective of the software environment will not result in an identical result. Both changes to configuration and external services, e.g., calibration data received from *CalCat*, as well as differences in the lower lying soft- and hardware can lead to a numerically different result. Furthermore, as the processing code is developed further, its application to previous data may yield a different output than an earlier version.

For the offline processing system, reproducing an earlier result is therefore considered a distinct action from reprocessing it. It is implemented at the `xfel-calibrate` level, where for every invocation a special *metadata folder* contains all necessary parameters about the computation itself, the executed notebook with concretized parameters, the software environment it ran in as well as the captured responses from external services. A second command `xfel-calibrate-repeat` then uses this metadata directory to re-run the same code as before, with the same parameters, in a similar Python software environment, with the same external service responses. Some lower-level factors are not tracked in this implementation, such as the type of CPUs running the code or the compiler used for dependencies.

Essential for the data quality aspect of facility-provided processing is the continuous verification of its results. On a purely software engineering level, this is achieved by a wide coverage of unit tests [14] to test components individually. These tests are triggered automatically on every code change as part of a continuous integration workflow. In addition, an end-to-end approach from a scientific perspective is used, which processes data taken during regular user operation and compares the output against the expected result. For every supported processing task, a collection of such reference data is curated alongside the intended and verified product. As part of ongoing improvements, this reference result is regularly replaced after manual examination. The list of configurations is also extended to cover significant or incompatible changes, e.g., a different data structure on the detector side, to ensure that newer and enhanced processing code also works on older data.

3 Results

The described machinery for facility-provided processing has been used for a wide range of actions, chief among them characterization and image corrections of the custom large-area 2D imaging detectors. Recently, this has been extended to special operating modes for these as well as entirely different but essential pre-processing steps for types of experiments not involving pixel-based detectors (see below).

3.1 Supported detectors and actions

A primary data driver of several instruments at European XFEL are the *AGIPD*, *DSSC*, and *LPD* detectors, which are developed specifically to exploit the unique burst mode time structure. As such, their uniqueness necessitated establishing new characterization and correction methods, and we consider it critical to offer an implementation ourselves. Common to all these systems, and a particular challenge for any applied processing method, is their very high data rate on the order of 100 Gbit/s for a Mpixel detector.

The *Adaptive Gain Integrating Pixel Detector* (AGIPD) [3] is a fast, integrating detector in the hard X-ray regime with adaptive gain. It offers single photon sensitivity at 12 keV and a dynamic range of up to 10^4 12 keV photons while being able to take up to 352 consecutive images at the facility's pulse repetition of 4.5 MHz. This image burst is then read out at 10 Hz between pulse trains (compare Figure 1), resulting in a total frame rate of up to 3,520 Hz. There are currently two 1 MPixel installations consisting of 16 modules each in use at the SPB/SFX [15] and MID [16] instruments, as well as another 0.5 MPixel prototype system with 8 modules and an upgraded version of the readout ASIC (application-specific integrated circuit) at the HED instrument [17]. A rich set of image corrections is implemented for this detector. First, the gain stage each pixel was recorded in is chosen through a threshold procedure followed by offset subtraction. Both the required threshold and offset values are inferred from dark image characterization automatically performed during operation in regular intervals. In certain scenarios, baseline shifts and common mode effects, both spatially per ASIC and temporally across multiple trains, can be accounted for. Finally, gain calibration converts pixel amplitudes to intensity in units of absolute energy. Several methods have been established to obtain the necessary slope characterization data and implemented as part of the facility-processing package, and are generally invoked manually. A detailed description of this detector and its calibration can be found in [18].

The *Large Pixel Detector* (LPD) [4] is another fast detector system acquiring up to 512 images at 4.5 MHz in three parallel gain stages. From these stages, an auto-gain mode can choose the optimal signal to resolve up to 10^5 12 keV photons. A Mpixel installation with 16 modules is in use at the FXE hard X-ray instrument [19] alongside several smaller single-module detectors called *LPD Mini*. The image corrections consist of the basic steps of offset subtraction based on automatically characterized dark images and subsequent gain calibration by pre-determined slopes obtained by manual analysis [20].

At the soft X-ray instruments SCS [21] and SQS [22], the *DEPFET Sensor with Signal Compression* (DSSC) [5] is available. This camera operates at a peak frame rate of 4.5 MHz and features on-chip digitization of 1-Mpixel images for up to 800 images. There are two versions of this camera, each employing different sensor technologies. The first version, which uses *MiniSSD* technology, has been in user operation since 2019. It offers several gain configurations to accommodate a broad range of soft X-ray photon energies. The camera exhibits a linear intensity response and thus, only offset subtractions are necessary before proceeding with further analysis. The second version of the camera utilizes *DEPFET* sensors and is currently in the commissioning phase. This technology offers superior noise performance, with an equivalent noise charge averaging 16 el rms [23]. It enables single-photon imaging capabilities down to a photon energy of 0.25 keV, all while maintaining a dynamic range of up to 10^4. However, its nonlinear response necessitates additional correction steps to convert ADC counts into photon energy. These are currently in development.

In addition to these large-area burst mode detectors, several other X-ray pixel detectors are used across the instruments with corresponding support for characterization and image corrections provided by the facility. The *JUNGFRAU* [24], *ePix100* [25] and *pnCCD* [26] are 2D frame-based detectors known from other facilities with robust and mature processing methods available in literature and upon which dark image pedestal subtraction, common mode, and gain calibrations are based. In the burst mode operation at European XFEL, these detectors generally are unable to record the intra-train pulses, hence they are only operated at the train repetition rate of 10 Hz. The *Gotthard-II* [27] is a 1D strip detector developed at the Paul Scherrer Institute for use at European XFEL capable of matching the pulse repetition rate. It is particularly suited to spectroscopic measurements. Here, the corrections also include an essential linearization of the raw output of the analog-to-digital converter (ADC) before subtracting offset and calibrating intensity to absolute units. Common to these detectors are considerably lower bandwidth requirements either due to their operation at only the 10 Hz train repetition rate or due to the smaller data volume of a single frame.

A different set of processing actions is available for a detector built on *Timepix3* [28], a time-resolved and event-driven pixel read-out chip. One such device is in use at the SQS instrument and is primarily used for electron and ion spectroscopy. Rather than full frames, it acquires individual time-over-threshold events for each of its pixels. A time walk correction is offered alongside centroiding to group neighbouring pixels illuminated at the same time into single particle impacts, if applicable. The calibration data required for the former correction process is currently acquired and prepared manually in a similar fashion to gain calibrations for frame-based detectors, but planned to be further automated in the future like dark image characterization is.

The SQS instrument also employs time and position sensitive delay line detectors [29] for charged particle and photon spectroscopy and imaging techniques such as REMI (reaction microscope) [30, 31]. Here the reconstruction process to assemble concrete particle impacts on the detector is entirely implemented as part of the facility-based processing systems, starting from digitized traces in the acquired raw data. After common mode correction of the analog data and discrimination to pulse arrival times, these digital signals on each channel are sorted into tuples corresponding to the same detector hit. For optimal resolution and reconstruction quality, further time sum and position correction on the digital signals and sophisticated event sorting based on components of the vendor-provided *CoboldPC* package can be included. In this application, the output format involving time and position events differs entirely from the initial input of analog voltage signals.

3.2 Special operating modes

In general, the core functionality of the implemented processing actions aims to be generic and experiment-agnostic apart from tuning parameters. Over the course of facility operation, however, more and more toggle-able operation modes have been added to aid users in data preparation procedures particular to their beamtime or technique.

In the case of offline analysis, this allows us to automatically enjoy the same benefits of reproducibility and scalability for these steps. While these special operating modes are in most cases not specific to a detector, their exclusive use at a particular instrument typically ties them to one detector and is thus implemented as part of its processing notebook.

One example of this is the generation of *virtual CXI* files for serial femtosecond crystallography (SFX) experiments [32] with the *LPD* detector after image corrections. The native data format of the Coherent X-ray Imaging Data Bank [33] specifies a particular layout of HDF5 files for SFX experiments, and analysis software developed in this scientific community can often use these files directly. By generating these files using HDF5 *virtual datasets* to refer to the corrected result in European XFEL's data format, this is possible without the need for an additional full copy on disk while being immediately available for users after acquisition and processing.

Another example influencing the actual data result is *photonization* available for the *AGIPD* detector [34]. Under certain illumination conditions commonly present at the MID instrument, pixel intensity after gain calibration can be interpreted as singular photon events of a particular photon energy and represented by an integer count. Performing this operation during image corrections can be implemented particularly efficiently for immediate analysis based off it. Furthermore, it serves as a data reduction technique, as the resulting integer representation is significantly more compressible, resulting in space savings of up to 97% at a negligible runtime cost [35].

Some operating modes are exclusively for data reduction purposes before processing takes place, e.g., limiting the trains or frames within a train to be included in the result. Even in the case of the *AGIPD* detector, different methods are employed depending on the experimental environment. At the SPB and MID instruments, the so called *LitFrameFinder* software automatically aligns the X-ray pulse pattern with the detector frame pattern to discard any frames in the processed data output not directly illuminated by X-rays. At the HED instrument, however, an optical chopper device is used to pick out entire pulse trains irrespective of the actual pulse filling pattern, and the processing result is thus reduced by exploiting this pulse picker information.

In the context of online processing, the main goal of offering special purpose analysis directly into the facility-provided processing system is exploiting data locality, especially in the case of high-bandwidth data accessed on GPUs. Currently, this is primarily done for real-time image corrections of large-area detectors via the mechanism described in 2.1, with the first implementations covering the computation of integrated intensity, counting lit pixels, and performing peak-finding in the context of SFX. In most other cases, special requirements in an online setting are often not mature or standard enough to warrant an application within these systems, and are left to more flexible online analysis solutions developed in-house [36] or those from the corresponding scientific communities.

3.3 Processing performance

For all detector implementations listed in 3.1, the online image corrections performed in real-time are able to cope with incoming data rates delivered as data streams. In the case of the *AGIPD*, *DSSC*, and *LPD* large-area detectors however, certain rate limitations still remain when actually moving the full data through the network.

These occur in particular when all modules are desired on a single online cluster machine, entirely exhausting its network links. A single group of correction devices each processing a single module on the other hand can generally be transported at full rate. For a given experiment, the ideal distribution can therefore be chosen, e.g., concentrating modules critical for analysis in the same group. Further optimizations are possible by reducing the precision, e.g., to half-precision `float16` per pixel as well as applying frame selections. Statically frames can be selected either by skipping entire trains or choosing fewer frames in each train, while calibration add-ons allow dynamic frame selection based on user-implemented criteria. Typically, up to 500 frames per second are currently achievable when assembling all modules to full frames for a single destination.

The achievable performance for basic image corrections is generally comparable on CPUs and GPUs when the total turnaround including memory transfer costs is considered. Still, these devices can perform these tasks more energy efficient and furthermore offer a surplus of computing resources that can be exploited for analysis or transformation steps These can be integrated much easier with the data already local in device memory at that time. Currently it includes computational tasks like peak finding and azimuthal integration described in 3.2 as well as memory-intensive axis stacking and reordering when requested by downstream user applications using this data. At the same time, the CPUs remain available to perform processing tasks not well suited or not yet implemented on GPU architectures.

To feed this to analysis suites running outside of Karabo, the lightweight Karabo-bridge protocol is available (in the illustrations, each “matcher” can have one or more bridge outputs), with Python [37] and C++ [38] bindings for convenient integration with existing software. This has been used, for example, to feed data to Hummingbird [39] or OnDA [40], two packages in the field of X-ray imaging experiments developed by the scientific community and widely used across facilities.

The offline processing system aims to operate as efficiently as possible on data already fully present in files. Generally, these files are grouped into datasets called *runs*, which contain all the data collected during manually operated triggers. The data volume is automatically split into independent parts to parallelize the work in multiple jobs across the Maxwell computing cluster. The automatic creation of PDF reports and capturing other metadata relevant to reproducibility typically takes less than 1 minute in a trailing job after all processing jobs have returned. The distribution of resources on the Maxwell compute cluster prioritizes currently ongoing experiments to guarantee near-immediate allocation of nodes without any delay. At this time, swift access to processed results for interpretation is particularly valuable to make optimal use of beamtime. After the experiment has concluded, processing requests are queued with the same priority as regular users and may incur waiting times. On average, this results in about 650 jobs per day but also peaked at 2,200 jobs over a single day in the current year of operations. The per-week job statistics for the current year of operations are visualized in Figure 4.

For image corrections and other automatic data pre-processing tasks, each job is assigned a set of trains along file boundaries. The image corrections for the large-area pixel detectors with burst mode generally achieve on the order of 1,500–5,000 frames per second of a single module in each job, compared to data acquisition rates of up to 3,520 frames per second in the case AGIPD, for example,. The pixel detectors not operated in burst mode and hence acquired at 10 Hz typically reach up to 140 frames per second. For the non-frame based detectors performance depends significantly on the experimental event rates, but train centroiding for *Timepix* is generally processed at least with 60 Hz, while *REMI* reconstruction varies between 40 and 80 Hz. By parallelization of larger datasets across multiple SLURM jobs, this generally matches or exceeds real-time across all processing options factoring in disk I/O and the constant costs of set-up and tear-down cluster jobs and processes.

In the case of the automatic dark characterization of pixel detectors, however, jobs are generally assigned all data belonging to a single detector module, if applicable. For burst mode detectors, these generally have runtimes of 10 min or less depending on the number of memory cells and trains considered for statistics. The runtimes of characterizing dark data of other detectors are negligible and complete within a minute or less. The non-automatic characterization tasks for gain calibration are not written with performance but completeness and traceability in mind with runtimes on the order of hours, as they are only repeated a few times per year.

For special circumstances where exceptionally fast file-based feedback is required, the offline processing machinery is capable of running on the online cluster usually reserved for streaming applications. This can take advantage of extremely fast disk I/O and avoid delays until data is accessible on the Maxwell compute cluster, but is limited to a too small number of machines to warrant the same level of parallelization. It is therefore typically restricted to cases with a high level of data reduction, e.g., pulse on demand techniques, that can take advantage of fast file-based data exploration. To this purpose, many of the features geared towards reproducibility and data tracking can be turned off to minimize run time to processing only. For example, corrected data written to file for a single AGIPD train with 352 frames can be made available within 30 s of acquisition.

The reproducibility of offline processing was confirmed for four different detector types (AGIPD, LPD, JUNGFRAU and ePix100)

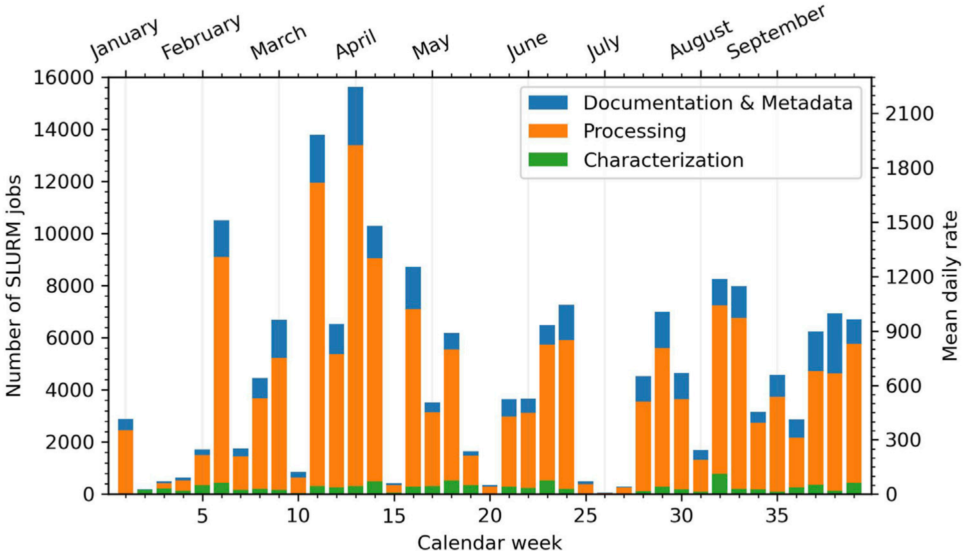

FIGURE 4
Number of processing jobs executed per week on the Maxwell SLURM cluster as part of the offline processing systems during the operations in 2023. Characterization jobs evaluate detector performance and generate calibration data, such as noise and pixel offsets. Processing jobs then transform experimental datasets, for example, by applying image corrections to 2D X-ray detector data. Documentation and metadata jobs run alongside all these actions to compile PDF reports with their results and capture metadata important for reproducibility.

using data from just after the reproducibility work was completed for each detector. This data was 18 months old for AGIPD at the time of testing, and 12 months for other detectors. In each case, the code used in the previous processing ran successfully and produced output data identical to the original results.

4 Discussion

4.1 Evolution of concept

Prior to facility operation and facing the challenges ahead, the concept for such a system was already planned and described in [41], with a particular focus on the *AGIPD*, *DSSC* and *LPD* detectors. On a conceptual level, significant differences can be found and discussed between the current implementation and the original expectations in three areas: i) online analysis, ii) access to raw data, and iii) configurability. This stems from experience accumulated during operation and an increasing diversity of scientific applications.

The online monitoring of experimental data was not expected to deliver the full input rate, but rather a continuous and non-guaranteed stream targeted at visual monitoring by operators. While this is critical for monitoring from a technical and detector operation perspective, the diversity of metrics significant for online analysis from a user perspective has proven to be much wider. Beyond immediate and in particular visual evaluation of immediate detector signals, the near real-time evaluation of technique-specific quantities based on as many detector frames as possible can significantly enhance the efficiency of the running experiment. For example, single particle imaging (SPI) [42] typically suffers from very low hit rates for interaction of the nanoscale targets with X-rays. An estimation of this rate during the experiment is more robust with access to at least some data of every acquired frame, rather than the entire data of a fraction of frames. This is primarily addressed by a flexible topology of the data stream to tailor any necessary compromise to each use case.

With the presence of a facility-provided processing system, no direct access to raw data affected by these systems was foreseen to be necessary by users. After calibrated files are produced by these systems, it was to only serve as the main archival data product, to be used when those files are no longer present in temporary storage. However, the operational experience so far has shown that a single truth for correction methods—particularly for this custom hardware, but also generally for the plethora of different experiments—has not been found. Established methods, like the aforementioned *SFX*, can almost exclusively rely on the processed datasets already, but this is not universal for other, often still developing, or novel techniques. Access to raw data remains essential for users with different requirements, either to expand on the processing methods already offered by us or replace them entirely with custom implementations. To this purpose, the utilized calibration data is available to users through the *CalCat* database as well. The collaboration on user's data treatment methods in this way allows to continuously adapt any improvements to the facility systems for the profit of all scientific users. Here, the distinction between reproducing an earlier result and reprocessing the same raw data was also underestimated, as it is possible that the current implementation may be newer and improved, giving a potentially better, yet different result. In the context of the scientific method, however, the capability to obtain an equal result is as essential as well.

To maximize the applicability to as many experiments as possible, and enable *user data* to be the data product sufficient for scientific analysis, processing cannot be limited to standardized blocks. The

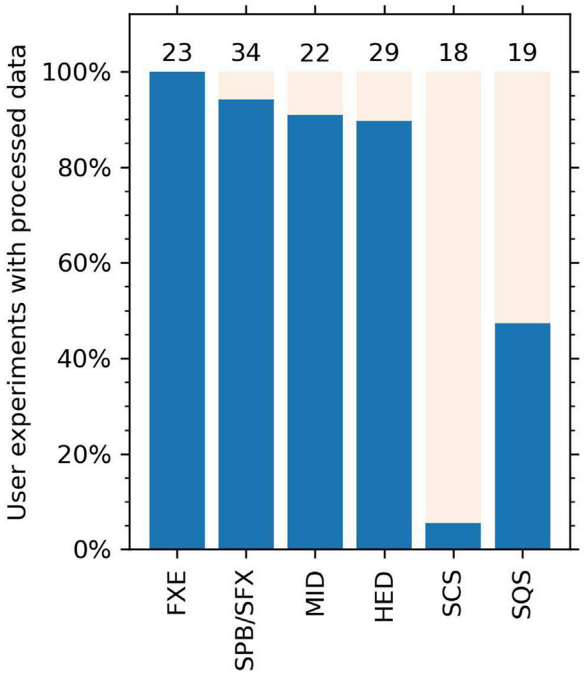

FIGURE 5
Ratio of user experiments of each instrument that generated data using the facility processing system in 2022 and 2023. The total number of experiments is shown above each bar.

emerging need for configurability covers immediate and often empiric parameters for computations, the implementation of special operating modes as discussed in 3.2, and tolerance to a constantly changing environment without stable interfaces. Even in the originally envisioned case of 2D imaging detectors, including such configurability in fundamental steps, like image correction and gain calibration, allows for greater adaptation to moving requirements. In contrast, limiting to standard methods risks focusing on the needs of established communities to the detriment of emerging fields. Instead, sensible defaults can aim to making common use cases as efficient as possible while leaving the option to expand for uncommon use cases as well. Parameters may range from flags for which algorithms are applied to tuneable numerical thresholds, but also includes technical configuration, like pipeline topology, for efficient network usage as described for 2.1. Whenever applicable, they are inferred from available metadata and hardware conditions and configured automatically. What remains may require manual tuning not just by operators and experts on the facility-side, but users as well. In line with the previous area of raw data access, this necessitates transparency and accessibility in this process. Finally, reliable operation of all these steps relies on countless implementation details of hard- and software upstream, which are themselves a moving target as the facility develops.

4.2 User operation

The facility processing system as presented here is in a mature and stable state for user operation at European XFEL. Since inception, it has seen widespread use during experiments performed at the facility illustrated in Figure 5 for 2022 and 2023. More than 90% of experiments across the hard X-ray instruments—namely, FXE, SPB/SFX, MID, and HED—are generating processed data as part of their data product over this time period. In a sharp contrast, approximately half of all experiments at SQS and only a single experiment at SCS took advantage of this system. These soft X-ray instruments differ significantly from the hard X-ray instruments in experimental techniques and thus detection methods. As discussed previously in 4.1, the standardized blocks focused originally around image corrections for 2D pixel detectors were not flexible enough for, or did not at all cover, the requirements and use cases of the soft X-ray instruments. In total however, 75% of the user experiments performed at European XFEL in 2022 or 2023 were assisted by the presented system in its current state.

A significant impact of processed data being generated seamlessly and automatically by facility systems has been observed on the storage infrastructure. The data volume of the large area detectors, like *AGIPD*, can effectively be doubled after acquisition by having both copies of raw and processed data on disk at the same time, culminating in single experiments exceeding multiple petabytes in a single week. This has been primarily mitigated by the introduction of reproducibility, which allows to only keep processing results on storage when they are actively in use for analysis, as they can be safely recreated after deletion. In addition, it stimulated the development of data reduction techniques that are applied during these processing steps. As such reduced processed data is the result of the same still unmodified raw data, these represented ideal opportunities for research of such methods and their validation.

These systems are provided for and continuously monitored by experts from the Data Analysis and Detector groups at European XFEL with expertise in software development, data analysis and detector characterization. These experts also provide 24/7 on-call support for user experiments. The established testing infrastructure verifies that expected data quality is preserved, and further quality improvements are confirmed manually in dedicated commissioning campaigns.

4.3 Outlook

Driven by the evolution of the calibration and processing concept offered by European XFEL as discussed in 4.1, further improvements are in development. These focus on increasing the flexibility and performance of both online and offline processing with data reduction in mind throughout the entire process.

While the online processing pipelines are able to process the data volume of all burst mode detectors at their full input rate, moving this data across the network still imposes limitations. On the technical side, Remote Direct Memory Access (RDMA) technology will increase the achievable bandwidth in the near future between DAQ and the processing infrastructure. This is combined with expanding on the correction add-on mechanism to exploit data being present in the memory of high-performance GPUs to benefit custom user analysis as well. In fact, these devices are increasingly used to accelerate time-consuming analysis tasks in the analysis of X-ray experiments and enable their real-time application, in particular for techniques based on machine learning [43–46]. These capabilities are also the ideal place to apply data reduction. As a result of the deep integration into the facility systems at this point, these decisions can not only be used to alleviate the bandwidth limitations in the online pipeline, but also reduce data before it hits the file storage. This removes the need for additional data reduction steps after writes have been performed.

For offline processing, the main focus is to use the existing scalable machinery for a wider catalogue of generic analysis steps beyond the facility-provided actions described in 3.1. Building on the first experiences here in the form of the special operating modes, this should expand to steps commonly re-implemented for many experiments as these are often agnostic of the underlying detector, such as azimuthal integration of detector frames for X-ray scattering experiments. Apart from the potential for highly optimized implementations, these can enjoy the same reproducibility guarantees as the existing actions. In the second step of this process, this machinery can be opened up to users to run their own, fully custom, analyses. The aim is for this to serve as a generic and accessible runtime for automatic, configurable, dataset-based offline analysis. This provides users instantly with a broad infrastructure related to managed code execution, monitoring, and parallelization. Both of these goals are contingent on further developments on the configurability and in particular interfaces first, like in the form of the myMdC web application. An important lesson already learned here for such an automated system is to clearly document and communicate the situations for which a particular analysis can be applied, for example, the photonization of absolute energy scales currently offered for the *AGIPD* detector. When the assumptions for such a method are not met, perhaps unknowingly, their application can result in a diminished data quality contrary to its intended purpose. Finally, this extends to more support for interoperability with existing solutions in the scientific community, for example, through the *NeXus* data format [47].

4.4 Summary and conclusion

A comprehensive and scalable system for processing scientific data has been developed at European XFEL. It aids users in essential preparatory processing steps, which are increasingly challenging due to high data rates and the use of custom detector technologies. Experimental data is delivered to the majority of user groups in a form suitable for further analysis at a constantly verified data quality. Future developments are focused on more support for a wider range of experimental techniques and integration of data reduction techniques.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://git.xfel.eu/calibration.

Author contributions

PS: Conceptualization, Data curation, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing–original draft, Writing–review and editing. KA: Data curation, Methodology, Software, Validation, Visualization, Writing–review and editing. CD: Data curation, Software, Writing–review and editing. DH: Data curation, Methodology, Software, Validation, Writing–review and editing. RR: Methodology, Software, Validation, Visualization, Writing–review and editing. TK: Methodology, Software, Validation, Visualization, Writing–review and editing. TM: Methodology, Software, Writing–review and editing. ES: Methodology, Software, Writing–review and editing. LG: Conceptualization, Funding acquisition, Resources, Supervision, Writing–review and editing. LM: Methodology, Software, Writing–review and editing. MM: Methodology, Software, Writing–review and editing. JM: Methodology, Resources, Software, Writing–review and editing. KW: Resources, Supervision, Writing–review and editing. JS-D: Data curation, Investigation, Methodology, Validation, Writing–review and editing. VR: Data curation, Investigation, Methodology, Software, Validation, Writing–review and editing. MR: Methodology, Validation, Writing–review and editing, Data curation, Investigation. ND: Data curation, Investigation, Methodology, Software, Validation, Writing–review and editing. DL: Data curation, Investigation, Methodology, Validation, Writing–review and editing. ID: Investigation, Methodology, Writing–review and editing. HY: Data curation, Investigation, Methodology, Software, Validation, Writing–review and editing. BS: Data curation, Investigation, Methodology, Software, Validation, Writing–review and editing. OM: Project administration, Resources, Writing–review and editing. MT: Funding acquisition, Resources, Supervision, Writing–review and editing. SH: Conceptualization, Investigation, Methodology, Software, Writing–review and editing. SA: Funding acquisition, Resources, Supervision, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The effort for the method, systems and infrastructure described in this work is carried by the entire staff of European XFEL. In particular all parts of the IT infrastructure maintained by the European XFEL ITDM and Deutsches Elektronen-Synchrotron (DESY) IT groups are critical for its operations. It relies on the reliable operation of detectors attended by instrument detector experts and other instrument staff. In addition, the authors acknowledge countless facility users that contributed feedback and ideas.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Decking W, Abeghyan S, Abramian P, Abramsky A, Aguirre A, Albrecht C, et al. A MHz-repetition-rate hard X-ray free-electron laser driven by a superconducting linear accelerator. *Nat Photon* (2020) 14:391–7. doi:10.1038/s41566-020-0607-z

2. Tschentscher T. Investigating ultrafast structural dynamics using high repetition rate x-ray FEL radiation at European XFEL. *The Eur Phys J Plus* (2023) 138:274. doi:10.1140/epjp/s13360-023-03809-5

3. Allahgholi A, Becker J, Delfs A, Dinapoli R, Goettlicher P, Greiffenberg D, et al. The adaptive gain integrating pixel detector at the European XFEL. *J Synchrotron Radiat* (2019) 26:74–82. doi:10.1107/S1600577518016077

4. Veale M, Adkin P, Booker P, Coughlan J, French M, Hart M, et al. Characterisation of the high dynamic range Large Pixel Detector (LPD) and its use at X-ray free electron laser sources. *J Instrumentation* (2017) 12:P12003. doi:10.1088/1748-0221/12/12/P12003

5. Porro M, Andricek L, Aschauer S, Castoldi A, Donato M, Engelke J, et al. The MiniSDD-based 1-mpixel camera of the DSSC Project for the European XFEL. *IEEE Trans Nucl Sci* (2021) 68:1334–50. doi:10.1109/TNS.2021.3076602

6. Hauf S, Heisen B, Aplin S, Beg M, Bergemann M, Bondar V, et al. The Karabo distributed control system. *J Synchrotron Radiat* (2019) 26:1448–61. doi:10.1107/S1600577519006696

7. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* (2016) 3:160018. doi:10.1038/sdata.2016.18

8. H2020-EU14 – EXCELLENT SCIENCE – Research Infrastructures. *Photon and neutron open science cloud* (2018). doi:10.3030/823852

9. Koranne S. *Hierarchical data format 5: HDF5.* Handbook of open source tools. Berlin, Germany: Springer US (2011). p. 191–200. doi:10.1007/978-1-4419-7719-9_10

10. Deutsches Elektronen-Synchrotron. Maxwell-cluster (2023). Available at: https://confluence.desy.de/display/MXW/Documentation.

11. Granger BE, Pérez F. Jupyter: thinking and storytelling with code and data. *Comput Sci Eng* (2021) 23:7–14. doi:10.1109/MCSE.2021.3059263

12. Kluyver T. nbparameterise (2014). Available at: https://github.com/takluyver/nbparameterise.

13. Jette MA, Wickberg T. *Job scheduling strategies for parallel processing*. Berlin, Germany: Springer (2023). p. 3–23. doi:10.1007/978-3-031-43943-8_1

14. Hulzinga D, Kolawa A. *Automated defect prevention: best practices in software management*. Hoboken, New Jersey: Wiley-IEEE Computer Society Press (2007).

15. Mancuso AP, Aquila A, Batchelor L, Bean RJ, Bielecki J, Borchers G, et al. The single particles, clusters and biomolecules and serial femtosecond crystallography instrument of the European XFEL: initial installation. *J Synchrotron Radiat* (2019) 26:660–76. doi:10.1107/S1600577519003308

16. Madsen A, Hallmann J, Ansaldi G, Roth T, Lu W, Kim C, et al. Materials imaging and dynamics (MID) instrument at the European X-ray free-electron laser facility. *J Synchrotron Radiat* (2021) 28:637–49. doi:10.1107/S1600577521001302

17. Zastrau U, Appel K, Baehtz C, Baehr O, Batchelor L, Berghäuser A, et al. The high energy density scientific instrument at the European XFEL. *J Synchrotron Radiat* (2021) 28:1393–416. doi:10.1107/S1600577521007335

18. Sztuk-Dambietz J, Klackova I, Klyuev A, Laurus T, Trunk U, Ahmed K, et al. Operational experience with adaptive gain integrating pixel detectors at European XFEL. *Front Phys* (2023).

19. Galler A, Gawelda W, Biednov M, Bomer C, Britz A, Brockhauser S, et al. Scientific instrument Femtosecond X-ray Experiments (FXE): instrumentation and baseline experimental capabilities. *J Synchrotron Radiat* (2019) 26:1432–47. doi:10.1107/S1600577519006647

20. Wheater R, Hart M, Veale M, Wilson M, Doblas-Jiménez D, Turcato M, et al. Development of data correction for the 1M large pixel detector at the EuXFEL. *J Instrumentation* (2022) 17:P04013. doi:10.1088/1748-0221/17/04/P04013

21. Tschentscher T, Bressler C, Grünert J, Madsen A, Mancuso AP, Meyer M, et al. Photon beam transport and scientific instruments at the European XFEL. *Appl Sci* (2017) 7:592. doi:10.3390/app7060592

22. Mazza T, Baumann TM, Boll R, De Fanis A, Grychtol P, Ilchen M, et al. The beam transport system for the Small Quantum Systems instrument at the European XFEL: optical layout and first commissioning results. *J Synchrotron Radiat* (2023) 30:457–67. doi:10.1107/S1600577522012085

23. Castoldi A, Ghisetti M, Guazzoni C, Aschauer S, Strüder L, Hansen K, et al. Qualification of the X-ray spectral performance of the DEPFET pixels of the DSSC imager. *Nucl Instr Methods Phys Res Section A: Acc Spectrometers, Detectors Associated Equipment* (2023) 1057:168686. doi:10.1016/j.nima.2023.168686

24. Mozzanica A, Andrä M, Barten R, Bergamaschi A, Chiriotti S, Brückner M, et al. The JUNGFRAU detector for applications at synchrotron light sources and XFELs. *Synchrotron Radiat News* (2018) 31:16–20. doi:10.1080/08940886.2018.1528429

25. Blaj G, Caragiulo P, Dragone A, Haller G, Hasi J, Kenney CJ, et al. X-ray imaging with ePix100a: a high-speed, high-resolution, low-noise camera. In: James RB, Fiederle M, Burger A, Franks L, editors. *Hard X-ray, gamma-ray, and neutron detector Physics XVIII.* Bellingham, Washington USA: International Society for Optics and Photonics (2016). doi:10.1117/12.2238136

26. Meidinger N, Andritschke R, Hartmann R, Herrmann S, Holl P, Lutz G, et al. pnCCD for photon detection from near-infrared to X-rays. *Nucl Instr Methods Phys Res Section A: Acc Spectrometers, Detectors Associated Equipment* (2006) 565:251–7. doi:10.1016/j.nima.2006.05.006

27. Zhang J, Andrä M, Barten R, Bergamaschi A, Brückner M, Chiriotti-Alvarez S, et al. Design and first tests of the gotthard-II readout ASIC for the European X-ray free-electron laser. *J Instrumentation* (2021) 16:P04015. doi:10.1088/1748-0221/16/04/P04015

28. Poikela T, Plosila J, Westerlund T, Campbell M, Gaspari MD, Llopart X, et al. Timepix3: a 65K channel hybrid pixel readout chip with simultaneous ToA/ToT and sparse readout. *J Instrumentation* (2014) 9:C05013. doi:10.1088/1748-0221/9/05/C05013

29. Jagutzki O, Mergel V, Ullmann-Pfleger K, Spielberger L, Spillmann U, Dörner R, et al. A broad-application microchannel-plate detector system for advanced particle or photon detection tasks: large area imaging, precise multi-hit timing information and high detection rate. *Nucl Instr Methods Phys Res Section A: Acc Spectrometers, Detectors Associated Equipment* (2002) 477:244–9. doi:10.1016/S0168-9002(01)01839-3

30. Ullrich J, Moshammer R, Dorn A, Dörner R, Schmidt LPH, Schmidt-Böcking H. Recoil-ion and electron momentum spectroscopy: reaction-microscopes. *Rep Prog Phys* (2003) 66:1463–545. doi:10.1088/0034-4885/66/9/203

31. Boll R, Schäfer JM, Richard B, Fehre K, Kastirke G, Jurek Z, et al. X-ray multiphoton-induced Coulomb explosion images complex single molecules. *Nat Phys* (2022) 18:423–8. doi:10.1038/s41567-022-01507-0

32. Barends TRM, Stauch B, Cherezov V, Schlichting I. Serial femtosecond crystallography. *Nat Rev Methods Primers* (2022) 2:59. doi:10.1038/s43586-022-00141-7

33. Maia FRNC. The coherent X-ray imaging Data Bank. *Nat Methods* (2012) 9:854–5. doi:10.1038/nmeth.2110

34. Dallari F, Reiser M, Lokteva I, Jain A, Möller J, Scholz M, et al. Analysis strategies for MHz XPCS at the European XFEL. *Appl Sci* (2021) 11:8037. doi:10.3390/app11178037

35. Sobolev E, Schmidt P, Malka J, Hammer D, Boukhelef D, Möller J, et al. Data reduction activities at European XFEL: early results. *Front Phys* (2023).

36. Fangohr H, Aplin S, Barty A, Beg M, Bondar V, Boukhelef D, et al. Data analysis support in karabo at European XFEL. Proceedings of the 16th International Conference on Accelerator and Large Experimental Control Systems (2018) Barcelona, Spain, October 2018.

37. European-XFEL. karabo-bridge-py (2018). Available at: https://github.com/European-XFEL/karabo-bridge-py.

38. European-XFEL. karabo-bridge-cpp (2018). Available at: https://github.com/European-XFEL/karabo-bridge-cpp.

39. Daurer BJ, Hantke MF, Nettelblad C, Maia FRNC. Hummingbird: monitoring and analyzing flash X-ray imaging experiments in real time. *J Appl Crystallogr* (2016) 49:1042–7. doi:10.1107/S1600576716005926

40. Mariani V, Morgan A, Yoon CH, Lane TJ, White TA, O'Grady C, et al. OnDA: online data analysis and feedback for serial X-ray imaging. *J Appl Crystallogr* (2016) 49:1073–80. doi:10.1107/S1600576716007469

41. Kuster M, Boukhelef D, Donato M, Dambietz JS, Hauf S, Maia L, et al. Detectors and calibration concept for the European XFEL. *Synchrotron Radiat News* (2014) 27:35–8. doi:10.1080/08940886.2014.930809

42. Chapman HN. X-ray free-electron lasers for the structure and dynamics of macromolecules. *Annu Rev Biochem* (2019) 88:35–58. doi:10.1146/annurev-biochem-013118-110744

43. Wang C, Florin E, Chang HY, Thayer J, Yoon CH. SpeckleNN: a unified embedding for real-time speckle pattern classification in X-ray single-particle imaging with limited labeled examples. *IUCrJ* (2023) 10:568–78. doi:10.1107/S2052252523006115

44. Ekeberg T, Engblom S, Liu J. Machine learning for ultrafast X-ray diffraction patterns on large-scale GPU clusters. *Int J High Perform Comput Appl* (2015) 29:233–43. doi:10.1177/1094342015572030

45. Ignatenko A, Assalauova D, Bobkov SA, Gelisio L, Teslyuk AB, Ilyin VA, et al. Classification of diffraction patterns in single particle imaging experiments performed at x-ray free-electron lasers using a convolutional neural network. *Machine Learn Sci Tech* (2021) 2:025014. doi:10.1088/2632-2153/abd916

46. Zhang Y, Yao Z, Ritschel T, Villanueva-Perez P. ONIX: an X-ray deep-learning tool for 3D reconstructions from sparse views. *Appl Res* (2023) 2:e202300016. doi:10.1002/appl.202300016

47. Könnecke M, Akeroyd FA, Bernstein HJ, Brewster AS, Campbell SI, Clausen B, et al. The NeXus data format. *J Appl Crystallogr* (2015) 48:301–5. doi:10.1107/S1600576714027575