Online & Offline data storage and data processing at the European XFEL facility

IOP Conf. Series: Journal of Physics: Conf. Series **898** (2017) 062049 doi:10.1088/1742-6596/898/6/062049

Online & Offline data storage and data processing at the European XFEL facility

Martin Gasthuber[1], Stefan Dietrich1, Janusz Malka[1], Manuela Kuhn[1], Uwe Ensslin[1], Krzysztof Wrona[2] and Janusz Szuba[2]

[1] DESY, Notkestrasse 85, 22607 Hamburg, Germany
[2] European XFEL GmbH, Holzkoppel 4, 22869 Schenefeld, Germany

Abstract.

For the upcoming experiments at the European XFEL light source facility, a new online and offline data processing and storage infrastructure is currently being built and verified. Based on the experience of the system being developed for the Petra III light source at DESY, presented at the last CHEP conference, we further develop the system to cope with the much higher volumes and rates (50GB/sec) together with a more complex data analysis and infrastructure conditions (i.e. long range InfiniBand connections). This work will be carried out in collaboration of DESY/IT, European XFEL and technology support from IBM/Research. This presentation will shortly wrap up the experience of 1 year runtime of the PetraIII ([3]) system, continue with a short description of the challenges for the European XFEL ([2]) experiments and the main section, showing the proposed system for online and offline with initial result from real implementation (HW & SW). This will cover the selected cluster filesystem GPFS ([5]) including Quality of Service (QOS), extensive use of flash based subsystems and other new and unique features this architecture will benefit from.

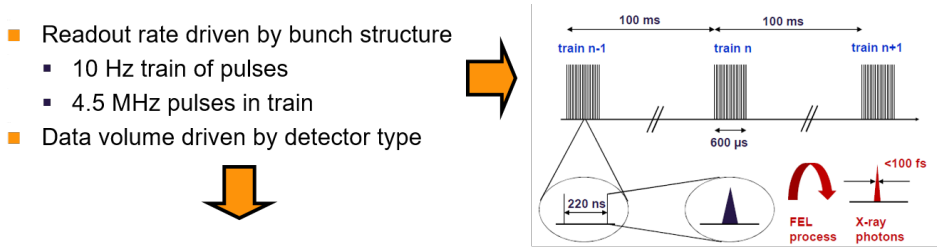
1. Introduction

The European XFEL is located mainly in underground tunnels which can be accessed on three different sites. The 3.4-kilometre-long facility will run from DESY ([7]) in Hamburg to the town of Schenefeld (Schleswig-Holstein). The Schenefeld site hosts the research campus, where international teams of scientists will carry out experiments using the X-ray flashes. Using the X-ray flashes of the European XFEL, scientists will be able to map the atomic details of viruses, decipher the molecular composition of cells, take three-dimensional images of the nanoworld, film chemical reactions and study the processes in the interior of planets.

2. The Challenge

Although the aggregates bandwidth is in a common range with other HEP like experiments, the main difference is in the bandwidth demands for single stream (one node to/from storage). For certain streams the required rate is up to 6 GiB/sec (InfiniBand FDR line speed) aggregating to more than 50 GiB/sec. The majority of data challenges depends on the used detector capabilities and the x-ray beam characteristics, figure 1 shows a summary of that.

To minimize the time where data is at risk, the automated data flow makes sure we generate two independent copies as soon as possible - minimizing the time the data is only in online storage and delay data deletion in the online and offline storage once the dCache copy is generated -



Detector type	Sampling	Data/pulse	Data/train	Data/sec
1 channel digitizer	5 GS/s	~2 kB	~6 MB	~60 MB
1 Mpxl 2D camera	4.5 MHz	~2 MB	~1 GB	~10 GB
4 Mpxl 2D camera	4.5 MHz	~8 MB	~3 GB	~30 GB*

Figure 1. Main accelerator and detector characteristics highlighting the data challenges

including the copy to tape. Data integrity is based on native GPFS features founded especially in the GPFS native raid ([1]) implementation including the end to end checksums. The dCache instance generates checksums (and check on reads) for all disk and tape copies.

3. Architecture

The overall architecture shown in figure 2 shows the main system components and networks. The additional XFEL specific middleware to handle user interactions (i.e. metadata database, scheduling database, etc.) are not shown. A more detailed discussion could be found in section *Cluster components and configuration.*

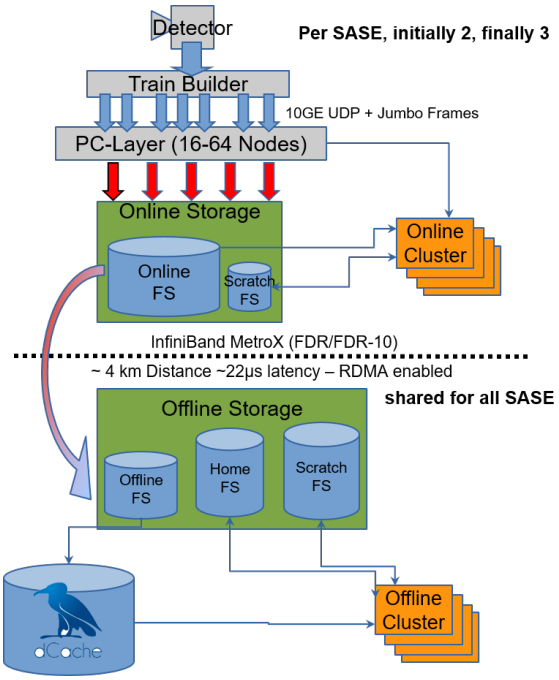


Figure 2. Main system components and networks covering the complete life cycle of experimental data

The depicted components are responsible for certain operations, namely:

Train Builder	– reshuffles picture modules into whole picture – pictures shuffled in trains – sends single trains per channel
PC-Layer	– data analysis for monitoring – data reduction, i.e. FPGA based compression – veto – file creation in memory and online filesystem, every node creates a 1GB HDF5 file every 1.6s (typical setup)
Online Storage	– dedicated instance per beamline – capture raw data stream from PC-Layer nodes – source for draining raw data & online user data to offline storage – provide generic storage resources for online user data analysis applications
Proxy Nodes - not depicted	– transfer online to offline storage (GPFS policy runs and AFM relation) – monitoring and cleanup
Online Cluster	– 10-80 nodes – dedicated instance per beamline – online data analysis and re-calibration – initial/online calibration of raw data – online data analysis, compression, high level veto
Offline Cluster	– 100-200 nodes – home of all offline user analysis – generation of calibrated data (raw + calibration) - read from dCache instance, write to offline storage for further user analysis
Offline Storage	– shared across all beamlines (SASE) – raw data arrives after delay, stored on GPFS – copy (raw) data to dCache (includes ACL config) – store calibrated data for user analysis
dCache - Raw Data	– stores all raw & calibration data – generate copies of raw and calibration data for archival storage (tape)

3.1. Cluster components and configuration

The test framework is basted on two IBM Spectrum Scale storage clusters, the on-line storage and the off-file storage cluster located in the experimental hall in Schenefeld and in the DESY computer center, respectively. The on-line storage cluster is consist of two GL4 systems, one GS1 system and four proxies servers. The off-line storage cluster consists of one GL4 system and four utility clients. In addition to the storage clusters, the three disk-less client cluster has been configured: the PC-layer cluster (19 nodes), the on-online analysis cluster (4 nodes), the off-line analysis cluster (more then 130 nodes). All installed nodes are connected to the Mellanox Infiniband high speed network - both standards: FDR (56Gb/s) and EDR (100Gb/s) are used. The experimental hall and the DESY computer center are connected using two Mellanox MetroX switches placed on both sites of the European XFEL tunnel. The solution extends our RDMA

network to 4km over dark fiber, joining two Infiniband fabrics. The Mellanox MetroX system, which has been implemented, supports up to 6 long-haul channels allows to transfer up to 240Gb/s in total and 6 download ports running 56Gb/s each. Currently only three long-haul links are used for tests. The total bandwidth can be easily increased by adding new switches to existing infrastructure. The current infrastructure is being constantly extended, in order to fulfill requirements of fast and reliable storage.

The on-line cluster is equipped with two file systems per SASE, one file system is dedicated to store all data which are common to all experiments operating on the beam-line like: calibration constants, log-files generated by devices supporting an experiment, etc., the second file system is dedicated to store data which are specific for a given experiment, concretely for an accepted proposal. The are four dedicated spaces on the on-line cluster for an experiment: "home", "data cache", "scratch", and "beam-time store". Each space is configured as a separate file system object called a fileset [1]. Similar structure has been configured on off-line cluster: "home", "data cache", "scratch", "ingest buffer". However, in contrast to the on-line cluster for the off-line cluster each space is configured as a separated file-system, this approach simplify further export of "spaces" to external clusters.

The on-line "data cache" space is introduced to store the raw data and timely migration of the good data to the off-line "ingest buffer". The data transfer is implemented by IBM Spectrum Scale policy, which allows to specify a set of the rules that describes the life cycle of the data based on the attributes of files. In this case a newly created data are identify and copied to the off-line "ingest buffer". The "ingest buffer" keeps data until migration to the dCache archive. The "beam-time store" is available for a proposal specific data and software, before, during and after scheduled experiments. Synchronization between the on-line and the off-line "beam-time store" is implemented by IBM Spectrum Scale utility called Active File Management (AFM). The on-line "scratch" space is temporary space for fast-feedback analysis and on-line calibration. The off-line "scratch" is the large space dedicated to user data analysis. The on-line and the off-line "scratch" spaces are independent from each-other. The "home" space is a system home for the on-line analysis cluster and the off-line analysis cluster. Both "home" spaces are also independent. As far the data acquisition and data analysis are concern three main processes will be present in the on-line system: the data injection, the data coping and the on-line data analysis. All processes use the same file system mounted on the on-line storage for write and read access. The data injection process is responsible for collecting the experimental data from the PC-layer and write them to the on-line storage, the data coping process transports the data from the on-line storage to the off-line storage, the on-line data analysis process perform the calculation on sub-sample of the raw data which are send directly to the on-line analysis cluster from PC-layer cluster, the results of the calculations are stored on the on-line storage cluster resulting in additional parallel data ingest. For the off-line system two processes should be taken into account: the incoming data transfer from the on-line cluster and the data archiving process to the dCache.

To verify the system bandwidth capabilities, various tests have been applied. Pseudo data is stored and the response of the storage is measured. In order to simulate the data ingest process files of 1GB were generated every 1.6 second and wrote to the on-line cluster, the resulting throughput was under investigation - see figure 3 (top). The simulation of the data writes from the on-line analysis was preformed by generating a set of files with different sizes from 1 kB to 1 GB in some pseudo-random order, the figure 3 (middle). Those two processes were run in parallel, after about 6 minutes the copy process from on-line to off-line storage was initialized, the figure 3 (bottom). In the figure 4 the write throughput for on-line (top) and off-line (bottom)

[1] A fileset is a subtree of a file system namespace that in many respects behaves like an independent file system. Filesets provide a means of partitioning the file system to allow administrative operations at a finer granularity than the entire file system

Figure 3. Throughput (GB/s) in function of time of three processes: the top picture – the simulated data ingest from PC-Layer cluster integrated over all nodes; the middle picture – simulated data ingest from on-line analysis cluster integrated over all nodes; the bottom picture - the data copy process (read) from on-line to off-line cluster, the magenta line denotes integrated throughput, other colors each node individually.

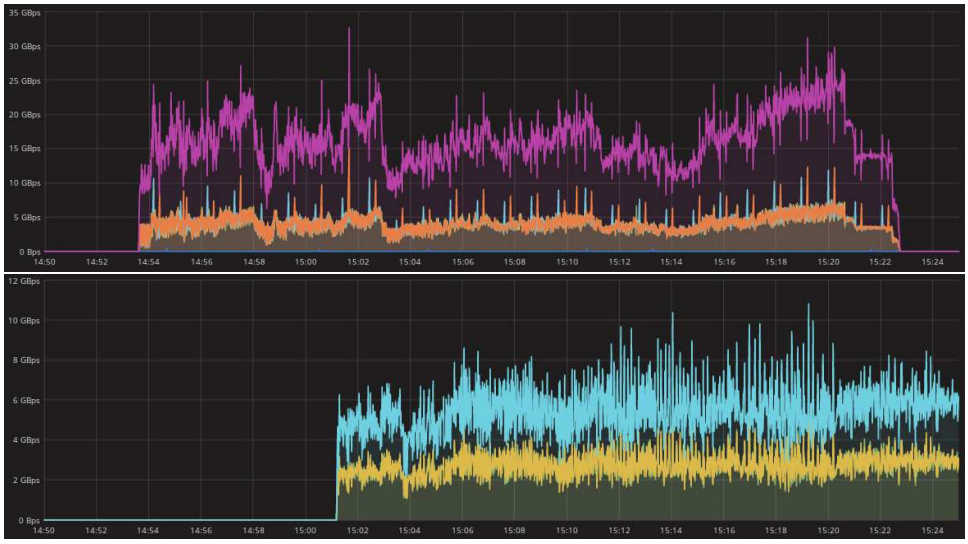


Figure 4. Throughput (GB/s) – writes in function of time for the on-line (top) and off-line (bottom) cluster. On the top picture the magenta line denotes integrated throughput over all nodes in the cluster, other colors each node individually. On the bottom picture the blue line shows integrated throughput over all nodes in the cluster, other colors each node individually.

cluster is show for completeness. It is clearly visible that the data ingest process has been not disturbed neither by on-line analysis process nor by the data drain to off-line cluster. Moreover, the writing time of single file was also measured in the test, yield the value of 0.29 ± 0.02 second,

which value is much below 1.6 second boundary. The data transfer process from the on-line to off-line cluster results with throughput about 6 GB/s, where the copy process is run by two proxies, with four proxies the throughput is about 12 GB/s. During the stress test the large number of 1 GB file were written to the on-line cluster as fast is possible over one hour resulting with the average value of about 29 GB/s, the number can be still improved by adding more capacity (additional GL4 system) resulting with more I/O operation per second available in the storage.

3.2. Quality of Service - QOS

The QoS allows to set the limit the I/O operations of given command by grouping them into classes which have assigned the maximum capacity in I/O operation per second (IOPS). We would like to use this capability to influence system performance by steering processes according to their priority. The figure 5 illustrate how the QoS works. The plots show the I/O operations of a process integrated over 5 second. On the first plot no limits on IOPS has been set, the process runs with rate of about 400 IOPS/5s and only limits due to other reasons then active QoS are present. For second plot the QoS functionality has been enabled and limit on IOPS has been set, clear throttling of I/O operations is visible.

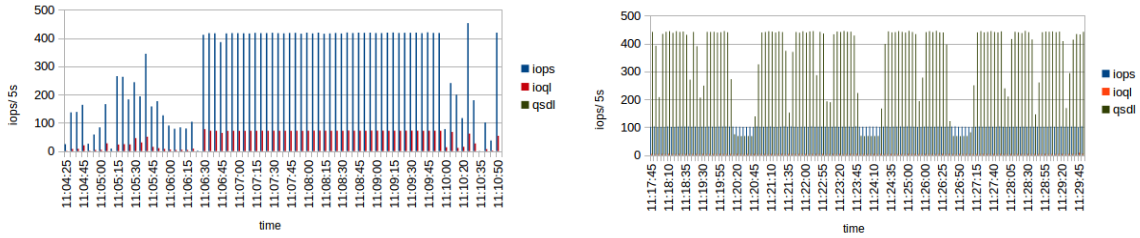


Figure 5. The number of I/O operation per second with three different IOPS groups: iops – the average number of I/O operations, ioql – the average number of I/O requests in the class that are pending for reasons other than being queued by QoS, qsdl – the average number of I/O requests in the class that are queued by QoS.

In our use case QoS will allow to control the amount of IOPS available for each of the three processes which are present in the system: the data injection, the data coping and the data analysis. The first results of tests show big potential of the QoS capability.

3.3. Light Weight Events - LWE

The current implementation makes extensive usage of GPFS policy runs to control the automated data flow between online, offline storage and finally the dCache instance. In order to support shorter delay times between data copies, we are looking into an event based model. This is based on new services in GPFS resulting in an equivalent of a cluster wide 'inotify' feature, allowing to capture and process filtered filesystem events and trigger the appropriate action.

3.4. All Flash for Online storage

In order to support even faster detectors, initial tests are done to verify the technical capabilities and integrity of powerfull all-flash based storage systems with capacities of more than 1PB and random access bandwidth of more than 30GiB/sec per building block. Recent market surveys indicates that economically solutions are on the horizon resulting in a much better 'burst' capturing characteristic compared to classic disk based systems.

3.5. Raw data and calibration constants repository - including archival
A dCache [6] instance has been set up for EuXFEL to act as raw data and calibration constants repository. With it's scaleout capabilities proven in LHC computing [4] , dCache allows for mass deployment of storage hardware that, while beeing cost effective, can be aggregated to keep up with expected ingest rates and also provides excellent data protection. dCache has proven interfaces to tertiary storage implementing the migration of *cold data* to archive media like tape or object stores.

Within the EuXFEL data flow, raw data is migrated from the ingest buffer (GPFS) to dCache starting already during the experiment, after initial quality control. It is then available as input to the calibration pipeline discussed below.

EuXFEL experiment data (raw data) needs to be calibrated to be usefull for user analysis. A set of up to 10^9 calibration constants is therefore prepared before starting the experiment or collected during data taking. As calibration constants can change over time, calibration is a repetitive process.

The EuXFEL dCache instance also implements the calibration constants repository. Versioned sets of calibration constants together with raw data are made available to the calibration pipeline using the dCache implementation of the NFS 4.1 (pNFS) protocol. Resulting calibrated data produced by the pipeline is written to the GPFS offline cluster to enable HPC style user analysis.

To provide the required bandwidth, dCache aggregates IO capabilities of storage building blocks. Initial tests using a single dCache building block consisting of two frontend machines attached to one disk array show a throughput rate of above 1 GB/s can be sustained for hours. Figure 6 shows IO bandwith achieved as a function of time in a $\frac{hour}{day}$ notation. Data was sent to dCache from a single host with a bonded 2 port 10 GE network connection. Write accesses simulating raw data taking are concurent to reads from processes on the frontends that copy data to tertiary storage. Additional reads starting near the end are due to NFS accesses coarsely simulating the calibration pipeline. The noticeable dip in the read rates is due to a temporary space shortage on the tertiary storage.

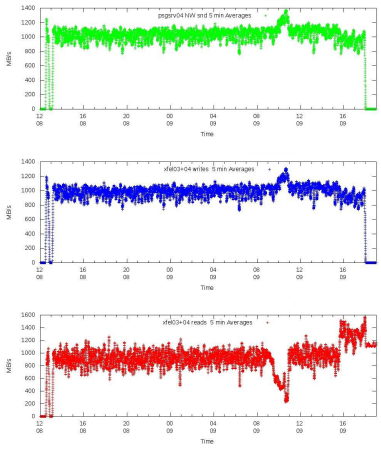


Figure 6. Outgoing network traffic from the sending host and aggregated write and read rates on the frontends.

3.6. Network Infrastructure
Beside the 'standard' IP networking based on Ethernet, we have setuped a dedicated long-haul InfiniBand link between the DESY datacenter and the EuXFEL Experimental Hall in Schenefeld

(approx. 4km fibre length), thus connecting the two InfiniBand fabrics on both sites resulting in a single RDMA capable fabric. This network is the core infrastructure to move experimental data from the 'Online Storage' systems to the central 'Offline Storage'.

4. Summary and Outlook

The pre-production system is setup and ready for the 2017 initial user operations. A small test instance will be used to further investigate QOS, LWE and all-flash configurations. Due to uncertainties in user interaction and final data processing schemes, we expect a re-design of certain components once the initial user operations completed and more precise planing is available. Initial performance measurements shows enough headroom to support even the largest detectors in preparation.

References

[1] GPFS Native RAID - Declustered Array, IBM Knowledge Center,http://www-01.ibm.com/support/knowledgecenter/SSFKCN_4.1.0/com.ibm.cluster.gpfs.v4r1.gpfs200.doc/bl1adv_introdeclustered.htm
[2] The European XFEL, http://www.xfel.eu/overview/in_brief/
[3] PETRA III, http://photon-science.desy.de/facilities/petra_iii/index_eng.html
[4] The LHC computing Grid, http://wlcg.web.cern.ch
[5] GPFS, General Parallel File System, http://www-01.ibm.com/support/knowledgecenter/SSFKCN/gpfs_welcome.html
[6] dCache, http://www.dcache.org
[7] DESY, German Electron Synchrotron, http://www.desy.de/index_eng.html