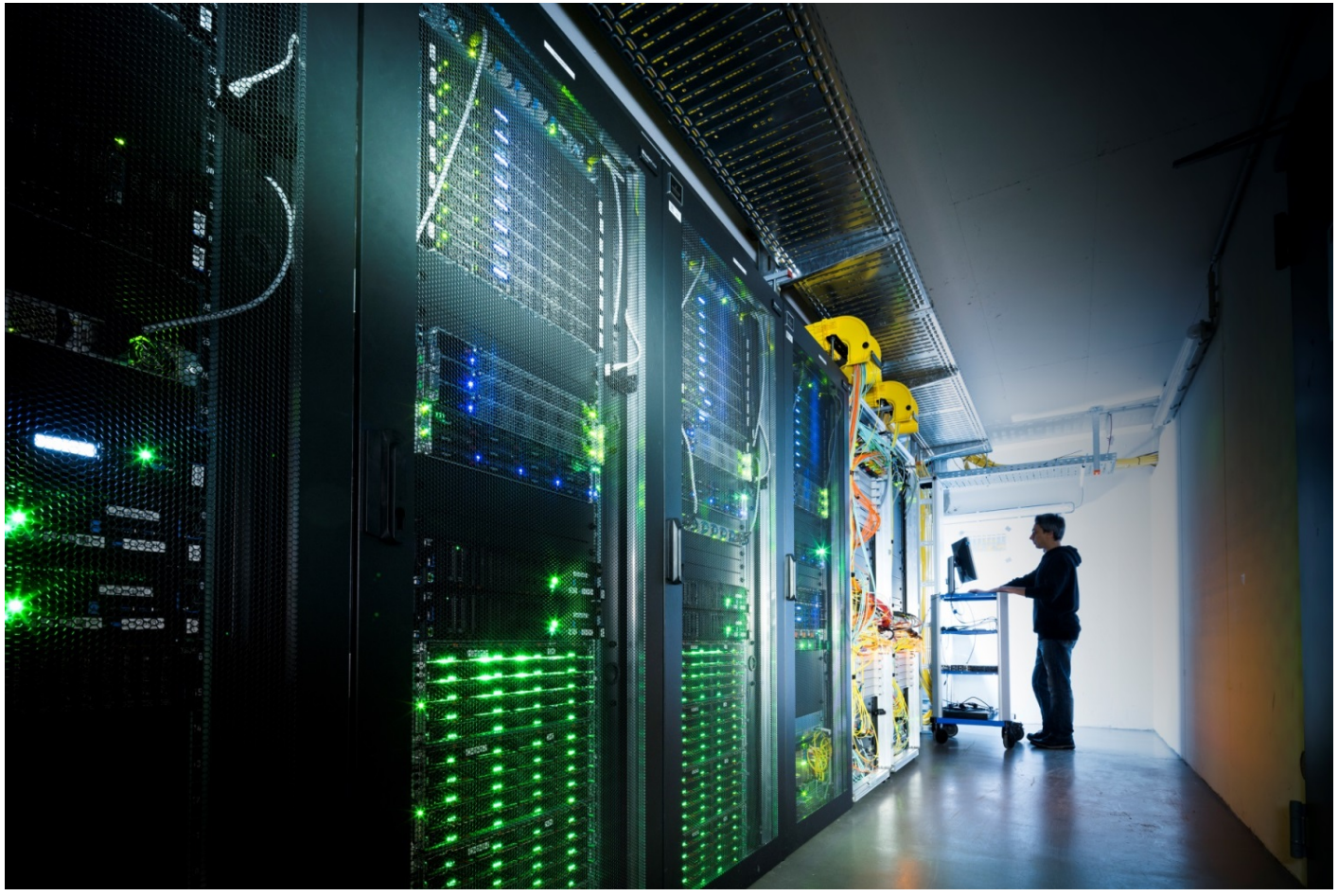
Data Management at European XFEL

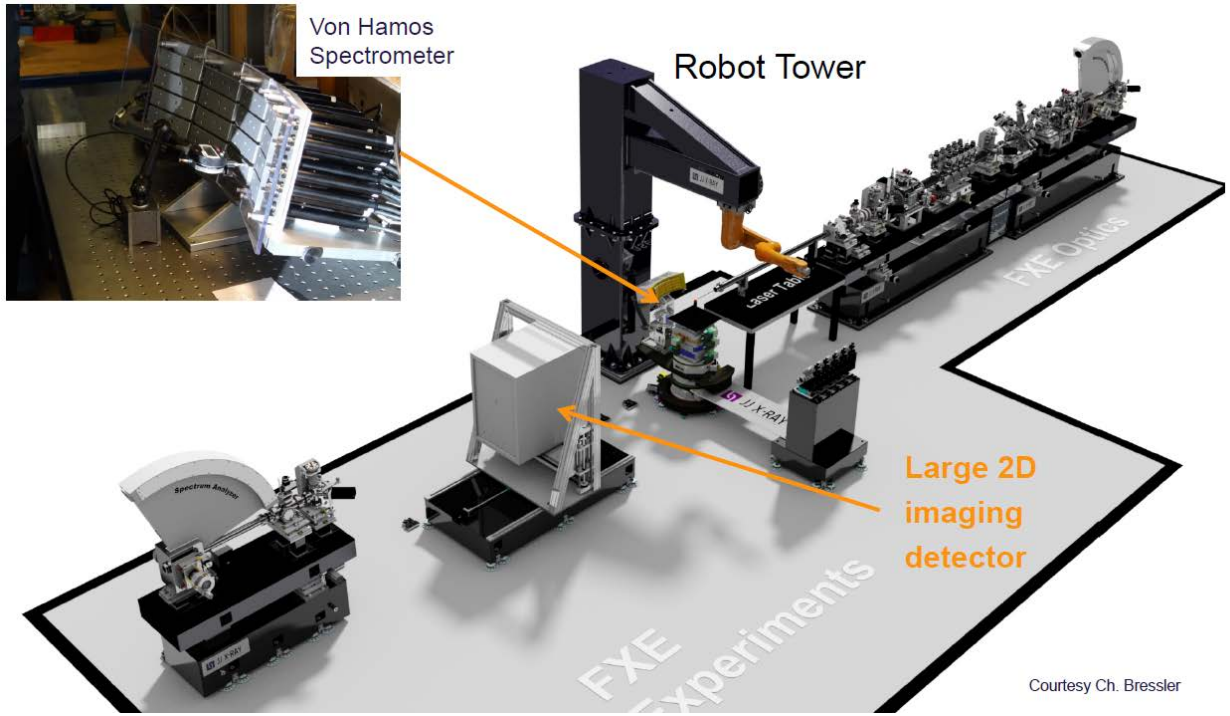
Krzysztof Wrona
IT and Data Management Group

Hamburg, 24.01.2019



Scientific Data Management

- Organize and manage data in a coherent way
- Aggregate and record large volumes of data
- Provide possibility of inspecting data during experiment
- Enable users efficiently analyze experiment data from anywhere
- Prepare for open access era



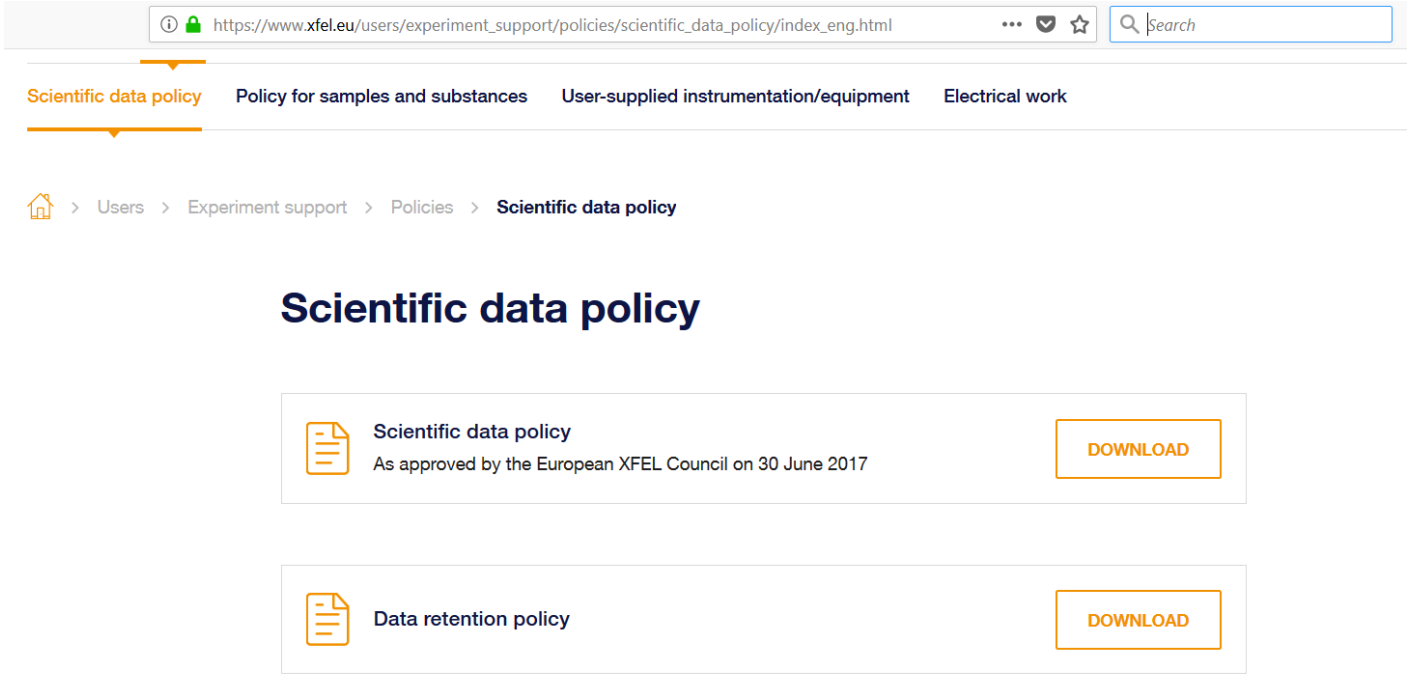
Detector type	Sampling	Data/pulse	Data/train	Data rates
1 slow camera	10Hz	-	~8MB	~80MB/s
8 channels digitizer	5 GS/s	-	~16MB	~160 MB/s
1 Mpxl 2D camera	4.5 MHz	~2 MB	~1 GB	~10 GB/s
4 Mpxl 2D camera	4.5 MHz	~8 MB	~3 GB	~30 GB/s

Strategic decisions, partners and agreements

- Present baseline strategy
  - Scientific Data Policy
    - ► Defines the rules for scientific data usage
  - Close collaboration with DESY
    - ► Some of data management services provided through DESY IT
  - Close collaboration with IBM
    - ► High performance storage for data recording and analysis
- Concept towards extending computing model outside of XFEL
  - Pilot project for Remote Data Analysis with NCBJ
  - European Open Science Cloud

Scientific Data Policy

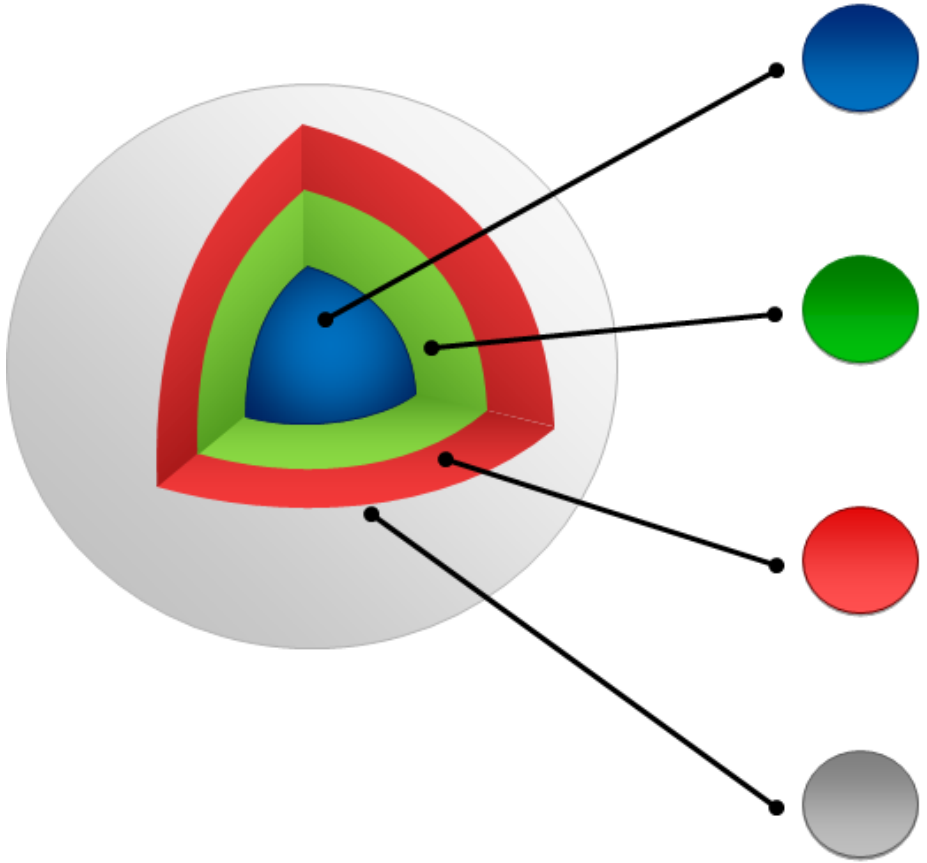
- Based on PaNdata recommendations
- General principles
  - Long term data curation
  - Open access
  - Embargo period with restricted access



- Regulates obligations and rights of European XFEL and Users with respect to data generated at European XFEL facility
- Acceptance of the scientific data policy is a prerequisite for every user to obtain beamtime – implemented through the European XFEL User Portal

Our vision – FAIR data

Findable Accessible Interoperable Reusable



DATA

The core bits

At its most basic level, data is a bitstream or binary sequence. For data to have meaning and to be FAIR, it needs to be represented in standard formats and be accompanied by Persistent Identifiers (PIDs), metadata and code. These layers of meaning enrich the data and enable reuse.

IDENTIFIERS

Persistent and unique (PIDs)

Data should be assigned a unique and persistent identifier such as a DOI or URN. This enables stable links to the object and supports citation and reuse to be tracked. Identifiers should also be applied to other related concepts such as the data authors (ORCIDs), projects (RAIDs), funders and associated research resources (RRIDs).

STANDARDS & CODE

Open, documented formats

Data should be represented in common and ideally open file formats. This enables others to reuse the data as the format is in widespread use and software is available to read the files. Open and well-documented formats are easier to preserve . Data also need to be accompanied by the code use to process and analyse the data.

METADATA

Contextual documentation

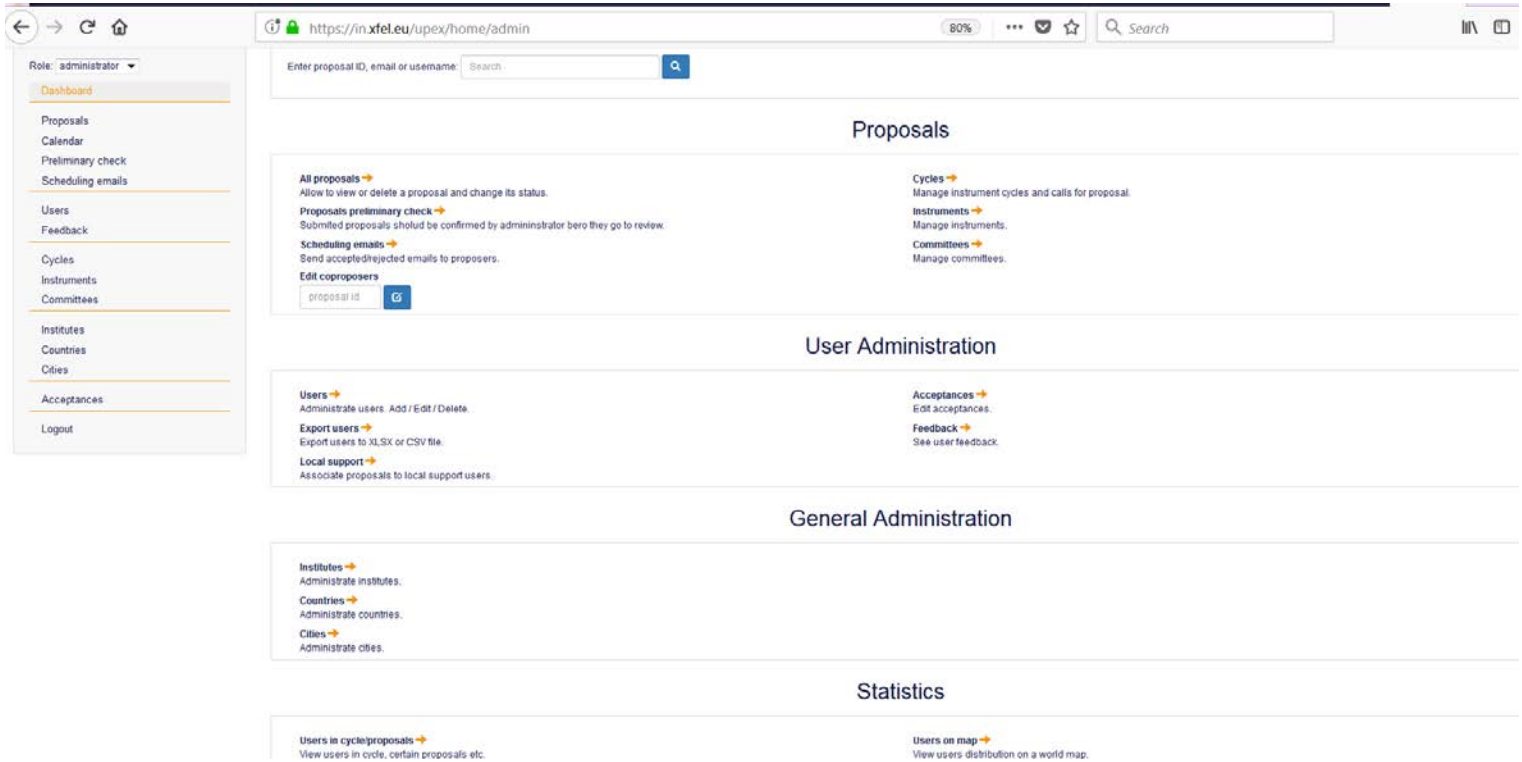
In order for data to be assessable and reusable, it should be accompanied by sufficient metadata and documentation. Basic metadata will enable data discovery, but much richer information and provenance is required to understand how, why, when and by whom the data were created. To enable the broadest reuse, data should be accompanied by a 'plurality of relevant attributes' and a clear and accessible data usage license.

PaNOSC

Initial detailed data retention scheme (currently under review)

Storage	Quota	Safety	Lifetime	Comment
dcache.raw	None	Tape Archive	6 months	Raw data
raw	None	None	2 months	Fast accessible raw data, lifetime not guaranteed
usr	5TB	Snapshots + Tape Backup	24 months	user data, results
proc	None	None	6 months	processed data e.g. calibrated
scratch	None	None	6 months	Temporary data (lifetime not guaranteed)
dcache.cal	None	Tape Archive	10 years	Calibration constants
cal	None	Tape archive	6 months	
user home	20GB	Snapshots + Tape Backup	Lifetime of the account	
archive.raw	None	-	Long-term	Long term means 5 years and XFEL will strive for 10 years
archive.cal	None	-	10 years	

UPEX – entry point for users



- User registration
- Acceptance of data policy, terms & conditions, …
- Call for proposal
- Proposal submission
- Support Review Process
- Scheduling
- Arrival forms
- Statistics
- Feedback

- As of 24.01.2018 in UPEX:
  - 2004 users registered
  - 4 calls for proposals
  - 16 instrument cycles

Experiment stages and data handling

Call for proposals submission

- Definition of instruments and instrument cycles
- Definition of resources to be used
- Definition of support teams

Submission, approval and scheduling

- Definition of experiment, PI and team members
- Definition of beamtimes, samples

Preparation phase

- Migration of metadata from User portal to the catalogue
- Creation of filesystem structure
- Upgrade of user accounts to allow access
- Initial implementation of access control lists based on experiment definition

Beamtime

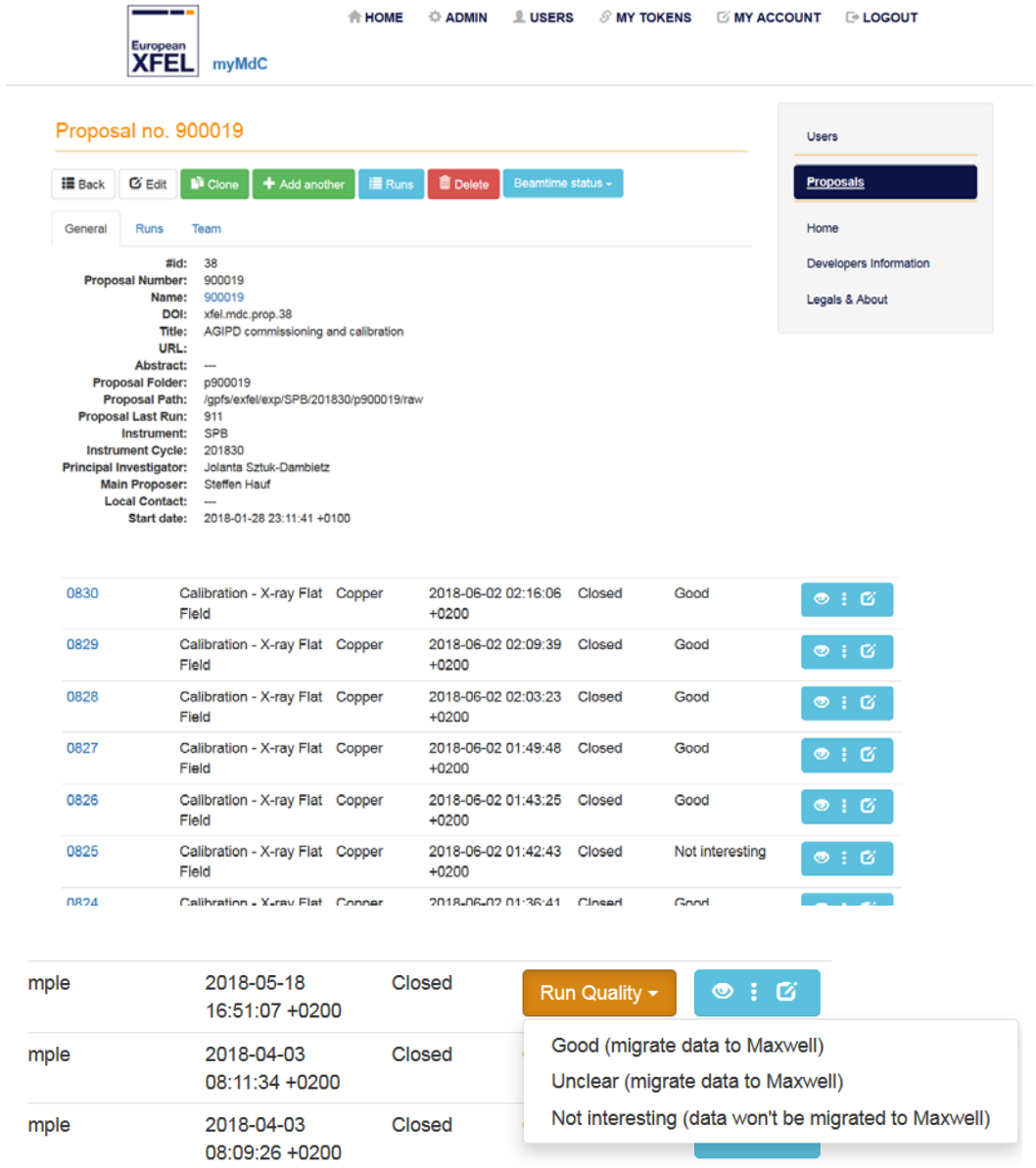
- Generation of experiment data
- Annotation of data in catalogues
- Dedicated online resources

Data analysis

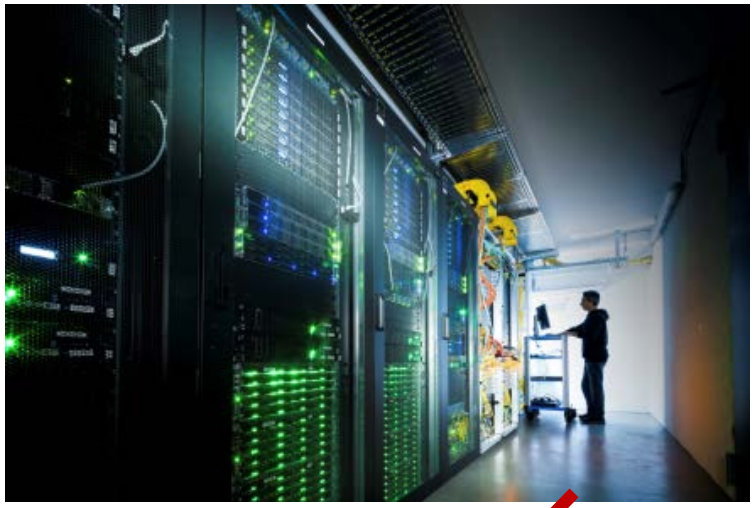
- Data processing
- Access to catalogues
- Access to shared offline computing resources

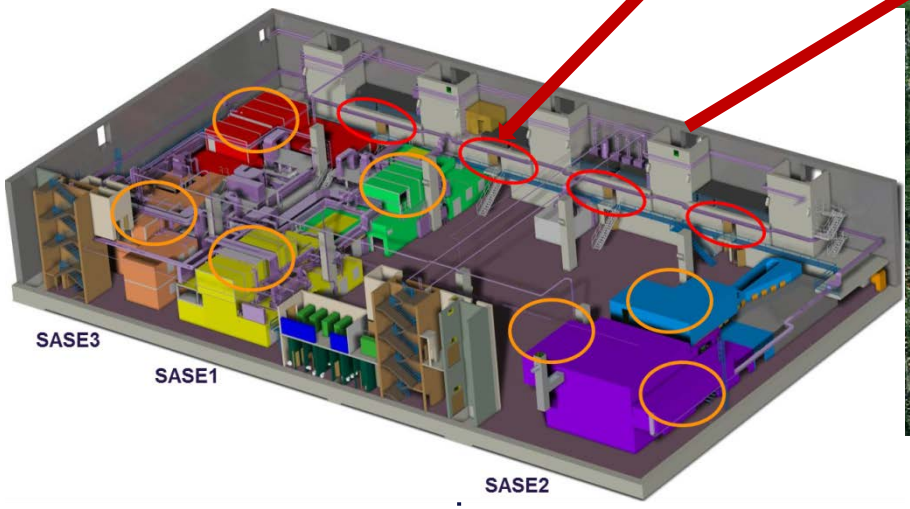
Metadata catalogue - myMdC

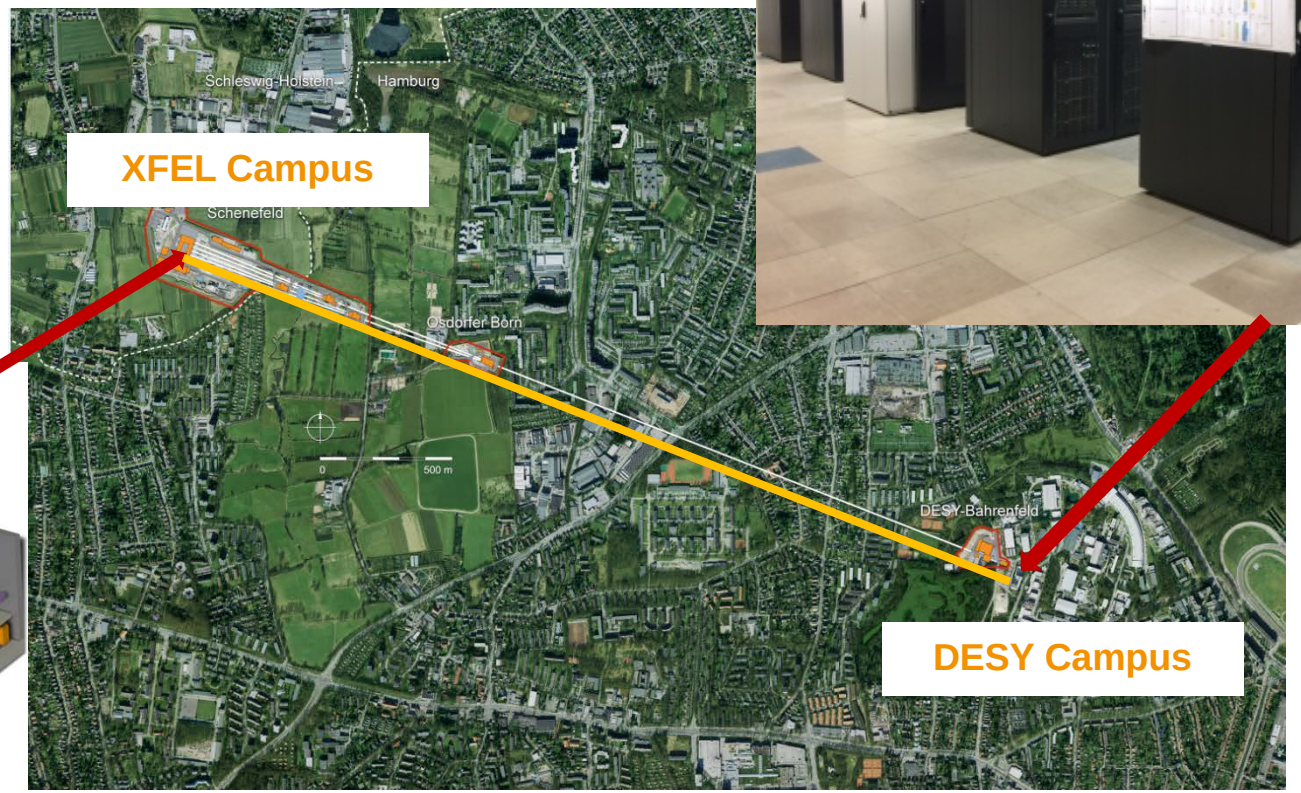
- **Provide means of storing, retrieving and query raw and run based data in an organized way**
- Glue together different metadata and data services
- Provides information to the “Run Control”
  - Generates unique run number
  - Data files for each run are registered
  - Runs are annotated with experiment an sample types
- Web interface for the users and instrument scientists
  - Users can inspect and annotate collected runs and files
  - Users can assess data quality and trigger data migrations
  - Catalogue keeps track of data files locations
  - PI manages team membership



IT infrastructure at XFEL and DESY

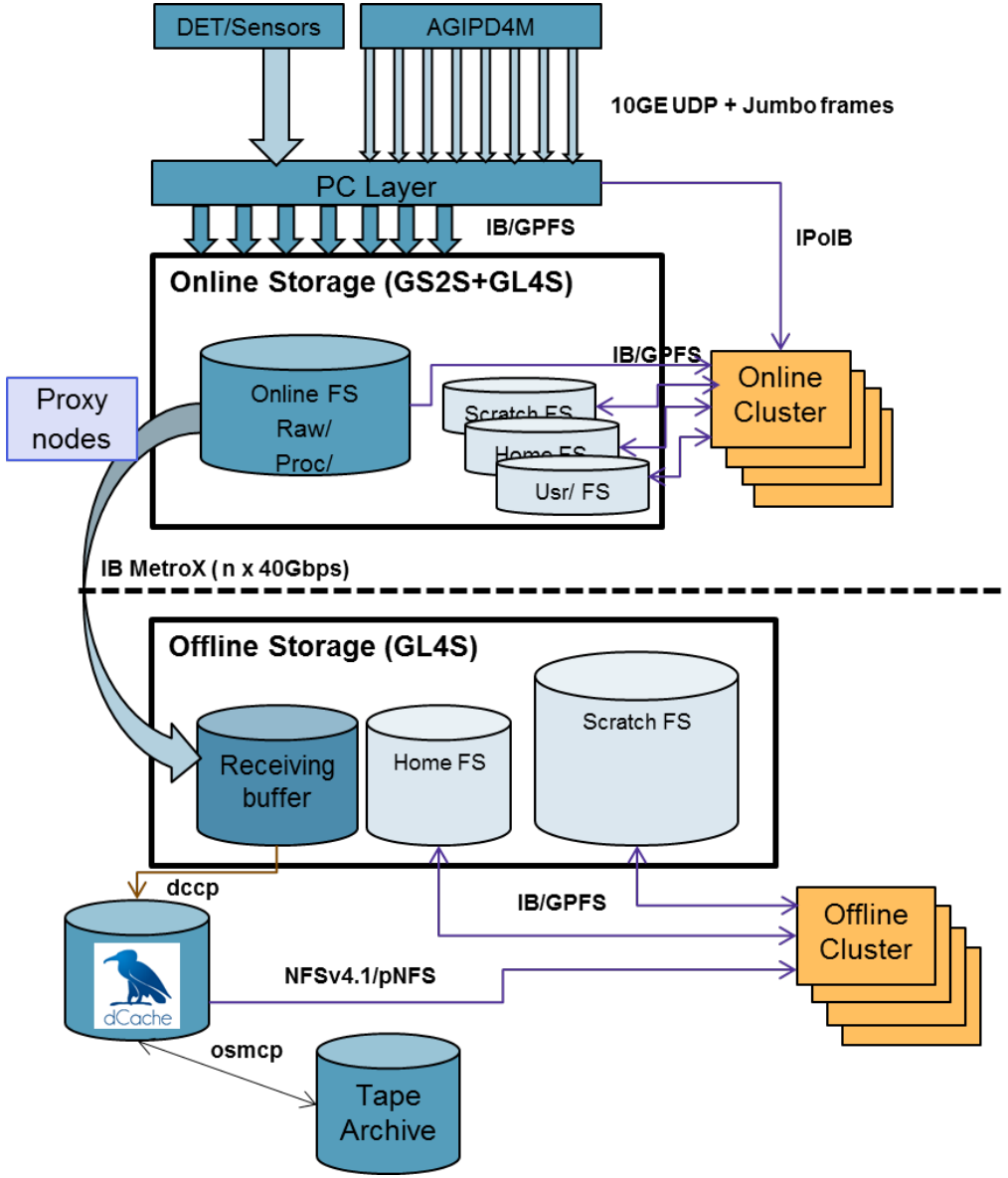






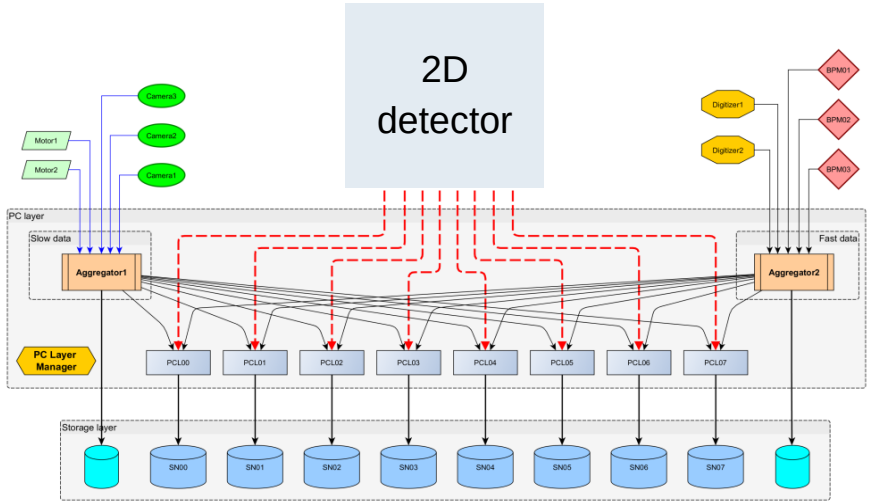
Main DAQ/DM installations

- Online systems
  - Located in Schenefeld
  - Close proximity to instruments
  - Temporary storage, caching
  - Fast feedback using dedicated computing resources
  - Dedicated installation for each SASE
- Offline system
  - Located in DESY data center
  - Common storage for all instruments and labs
  - Mid and long term storage with various data access patterns
  - Long term data archive with very rare access
  - Shared resources among all instruments (and more)



DAQ

- Devices (data sources) continuously and independently push data to DAQ system, aka. PC layer
  - Image data: 2D detectors → big & fast
    ► 2D detector sends data over 16 links
  - Pulse data: digitizers, cameras, XGM, … → fast data
  - Control data: sensors, motors → slow data

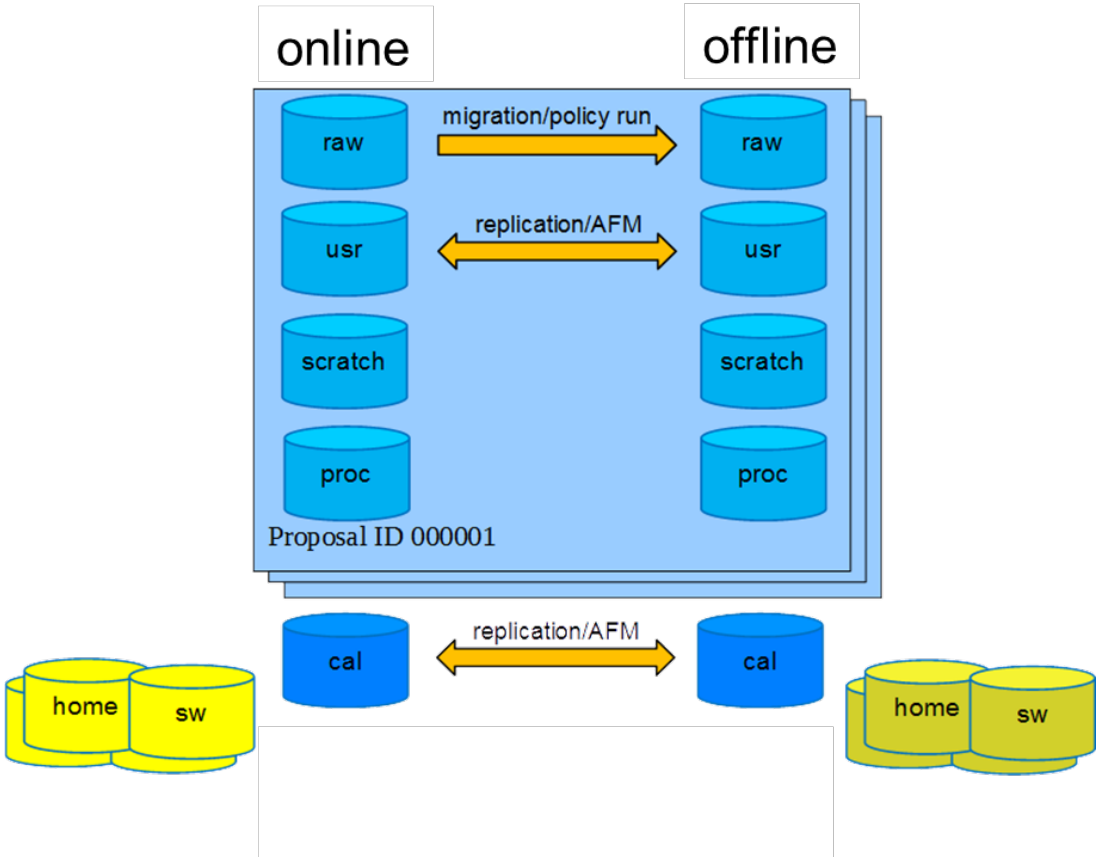


- Data handling and processing requires consolidating multiple fast/slow, small/large data streams based on train-id
  - Data is received by software devices at 10Hz (train rate) at most
  - All data are tagged with train-id (globally unique 64bits id)

- Data is formatted and stored in HDF5 files

Filesystem organization

- High level online and offline filesystem paths are identical
  - /gpfs/exfel/exp/
- Standard filesystem structure
  - <instrument>/<instrument-cycle>/p<proposal-id>
  - proc/ raw/ scratch/ usr/
  - raw/r<*XXXX*> proc/r<*XXXX*>
- Standard filenames
  - ***type*-r*XXXX*-*infix*-s*XXXXX*.h5**
    - ► **r** – run, **s** – sequence number, ***X*** – digit, **h5** – hdf5 extension
    - ► type – can be **RAW**, **CORR**
    - ► Infix – based on the aggregator/PC layer node
  - Use karabo_data python module to get abstract view and detailed information on the content and data organization



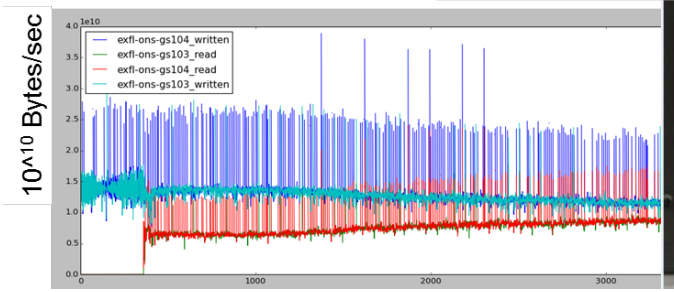
High performance storage hardware

- High performance storage for data recording and data analysis
  - Capable of handling data rates of the order of tens of GB/s
- IBM Spectrum Scale a.k.a. IBM General Parallel File System (**GPFS**)
  - GPFS native RAID (Erasure code) – short rebuild time
  - GPFS end-to-end checksum calculation for data integrity
- Online storage – per beamline
  - Original setup with disks only
  - Disk based data access was restricted only for beamline scientists
    - ► Due to observed high interference of typical data analysis access patterns with the DAQ and data migration processes
  - User data access through the streaming service from DAQ, and further calibration pipeline and karabo bridge service

Original setup at SASE1 (spinning disks only)

GPFS storage

- New storage building block
  - 2 x IBM **E**lastic **S**torage **S**ystems
    - ► GS4S (SSD, performance)
    - ► GL4S (spinning disks, capacity)
  - Up to 30GB/sec write performance
  - Transparent data movement between SSD and disk based hardware

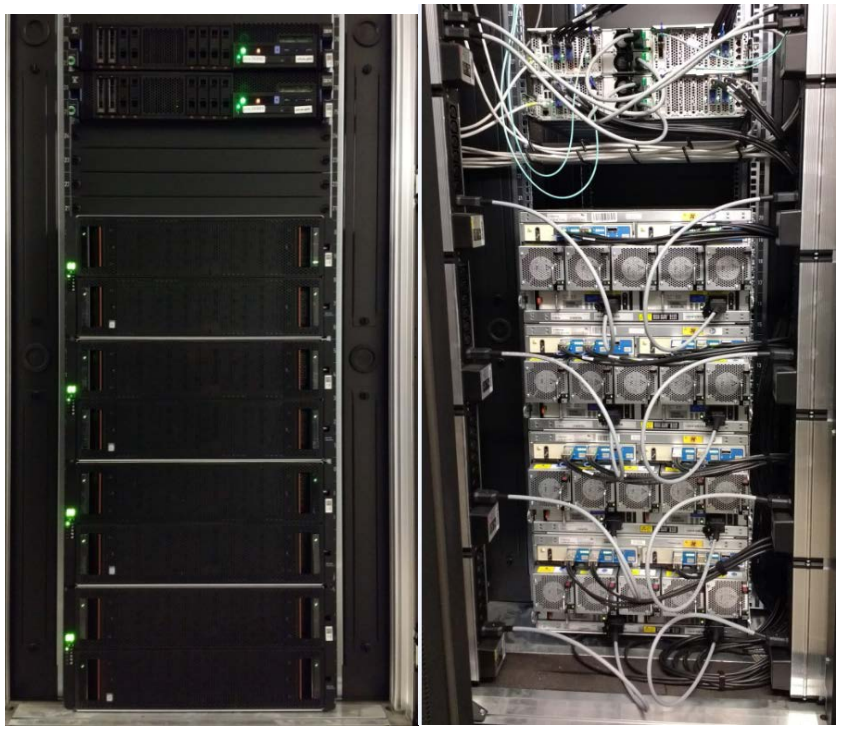


exfl-ons-gs103

exfl-ons-gs104

- First tests show that data writing performance is sufficient when data migration and random access to data is performed in parallel

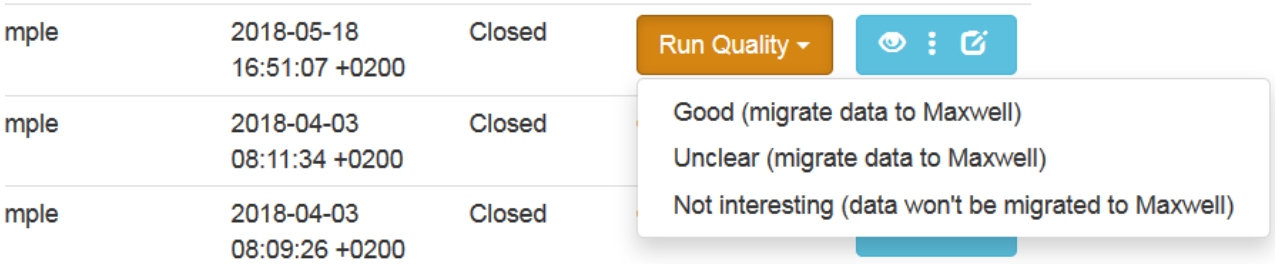
- We will open access to raw data on the online storage to the data managers (role defined in the myMdC for each proposal)
  - But we will carefully monitor if the data recording is not affected
  - SASE2 and SASE3 are ready, SASE1 in preparation



Online resources

- For each SASE
  - 8 GPU based servers for 2D detector calibration
  - 8 HPC servers for instrument specific processing
  - 4 HPC servers for users data processing and viewing
  - 1 extra GPU server will be added soon for other than 2D detector calibration purpose
- User access to the online cluster via ssh or FastX
  - from the control room PCs
  - from the dedicated PCs in the XHQ
- For support teams in addition from the gateway server

Data Migration Service (and beyond)

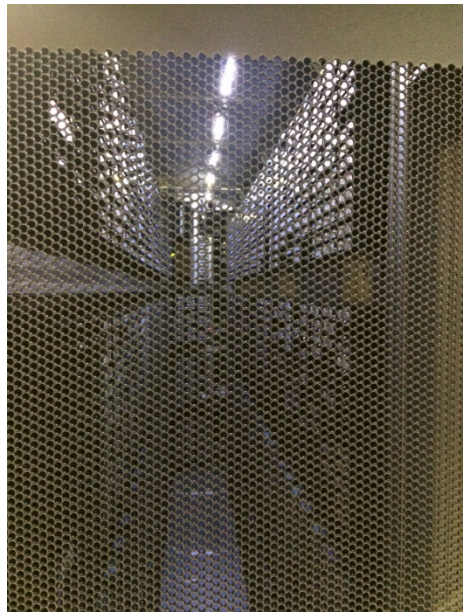


- Currently implemented
  - Simple data assessment service good, not clear, not interested
  - Allows us to remove junk data from tests or unsuccessful runs
    - ► This procedure should benefit from the user access to the online storage

- In the future extend the service to:
  - Trigger specific instrument/experiment pipeline processing
  - Introduce late assessment steps e.g. after a week, or after basic analysis was performed
  - Support event lists – only identified events are calibrated
  - Use external resources for computation -> pilot project for remote data analysis

Offline data storage

- Offline data storage in DESY data centre
  - Fast access
    - ► Hardware: IBM Elastic Storage System
    - ► Software: GPFS
    - ► Capacity: 3PB (+4PB in preparation)
  - Large capacity
    - ► Hardware: DELL Systems
    - ► Software: dCache/NFS4.1(parallel NFS)
    - ► Capacity: 10PB (+7PB soon)
  - Archive
    - ► Tape based, LTO8 (12TB/tape)
    - ► Only raw data and calibration constants are archived



Detailed initial data retention scheme – under review

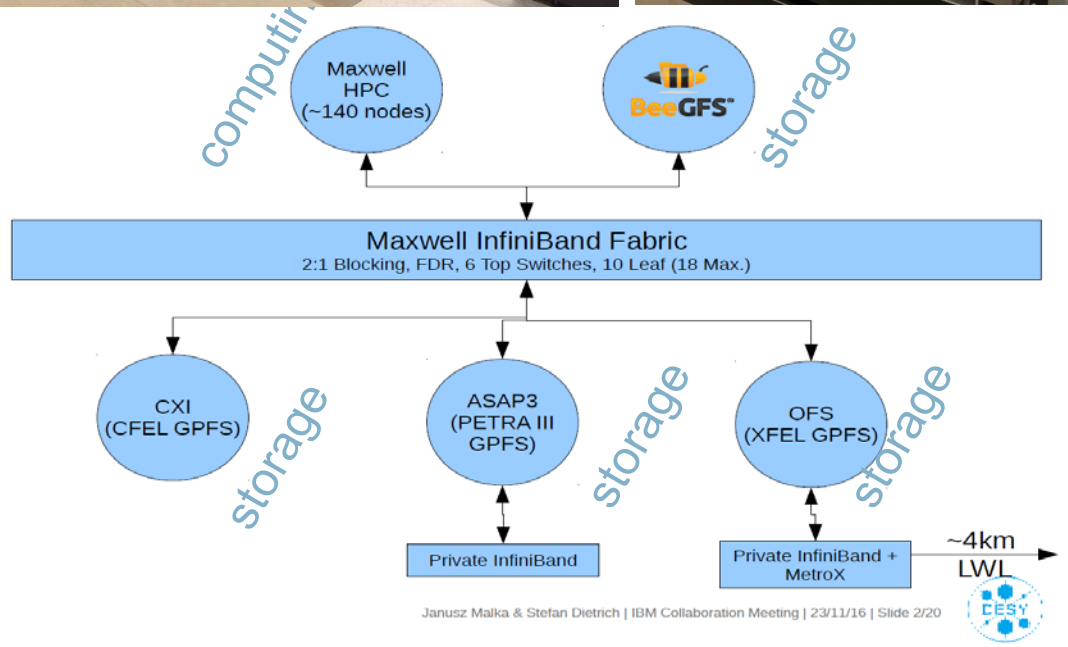
Storage	Quota	Safety	Lifetime	Comment
dcache.raw	None	Tape Archive	6 months (too short)	Raw data
raw	None	None	2 months (too long)	Fast accessible raw data, lifetime not guaranteed
usr	5TB	Snapshots + Tape Backup	24 months (too short)	user data, results
proc	None	None	6 months (too long)	processed data e.g. calibrated
scratch	None	None	6 months	Temporary data (lifetime not guaranteed)
dcache.cal	None	Tape Archive	10 years	Calibration constants
cal	None	Tape archive	6 months	
user home	20GB	Snapshots + Tape Backup	Lifetime of the account	
archive.raw	None	-	Long-term	Long term means 5 years and XFEL will strive for 10 years
archive.cal	None	-	10 years	

Data retention – new approach to be implemented

- GPFS storage needs to be used more efficiently
  - Shorter storage time
    - ► Removal of raw data from fast storage system (GPFS) right after experiment
      - Raw data accessible from dCache
      - Users are encouraged to analyze corrected (pre-processed) data
    - ► Further reduction of corrected datasets
      - Event list approach – big potential for some type of experiments (1-20% useful data)
      - More advanced processing as a service for users – e.g. azimuthal signal integration can reduce data to 0.1%
      - Smaller dataset volume can be stored for longer time
  - Processing (correction) on demand
    - ► Store for shorter time but make it available when required
  - usr/ folder retention synchronized with raw data

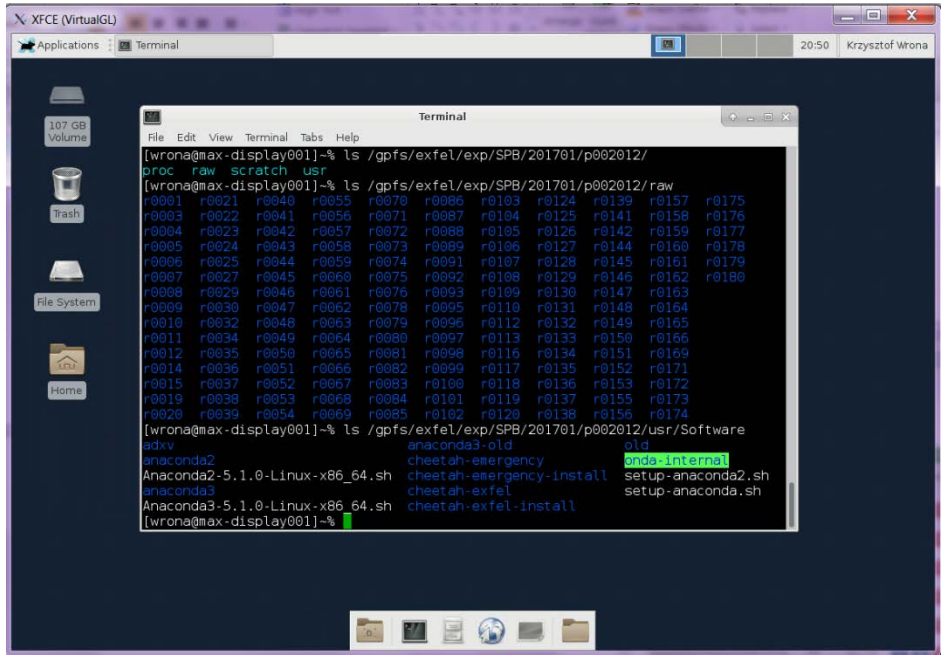
Status of Computing clusters

- XFEL resources in DESY data Center:
  - 180 nodes/6000 cores
    - ~200 TFlops
    - 512GB RAM
  - 24 nodes with GPU cards
  - Upgrade to 250 CPU nodes in 2019
- XFEL resources are integrated with the DESY Maxwell cluster
- Similar hardware in online clusters



User access to XFEL analysis facility

- Detailed Documentation for DESY Maxwell cluster
  - https://confluence.desy.de/display/IS/Maxwell
- Satellite workshop took place on 22.01.2019
  - “Computing for Photon Science made easy”
- Data export
  - FTP server
  - Globus service (in preparation)
- Standard account
  - Lifetime of 1 year after last successful experiment
  - Automatically extended: 2 x 1 year if present in the proposal



https://max-display.desy.de:3443

Acknowledgements

N. Al-Qudami[1], L.Boege[1], D. Boukhelef[1], I.Derevianko[1], S. Dietrich[2], U. Ensslin[2], K.Fillipakopoulos[1], M. Gasthuber[2], H.Giemza[3], J. Hannappel[2], Y.Kemp[2], B. Lewendel[2], L. Maia[1], J. Malka[1], M. Manetti[1], K.Ohrenberg[2], A.Paade[3], C. Patzke[2], G. Previtali[1], P.v.d.Reest[2], J. Szuba[1], P.Wasiuk[3], S. Yakubov[2]

1 - European XFEL GmbH, Hamburg, Germany

2 - DESY, Hamburg, Germany

3 – NCBJ, Swierk, Poland