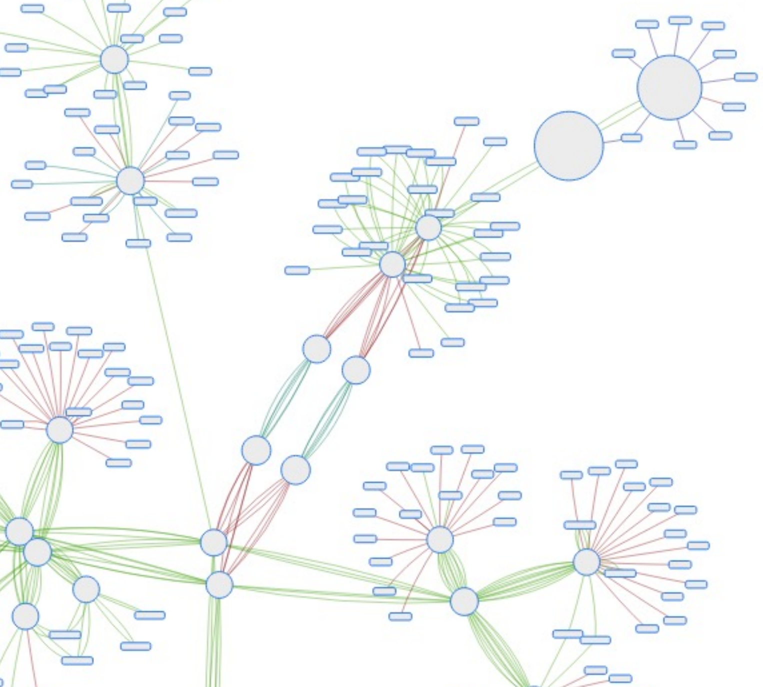
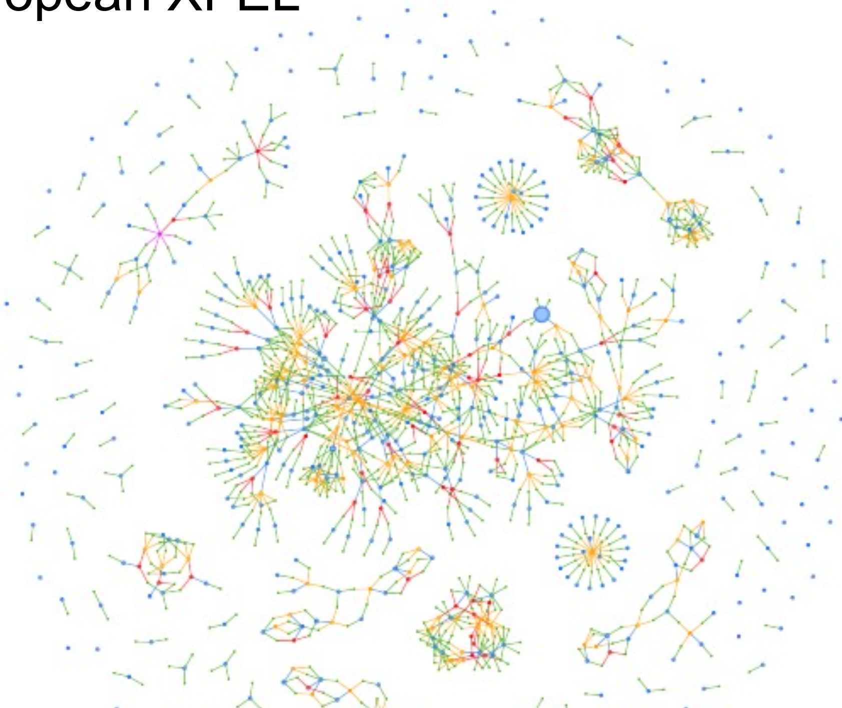
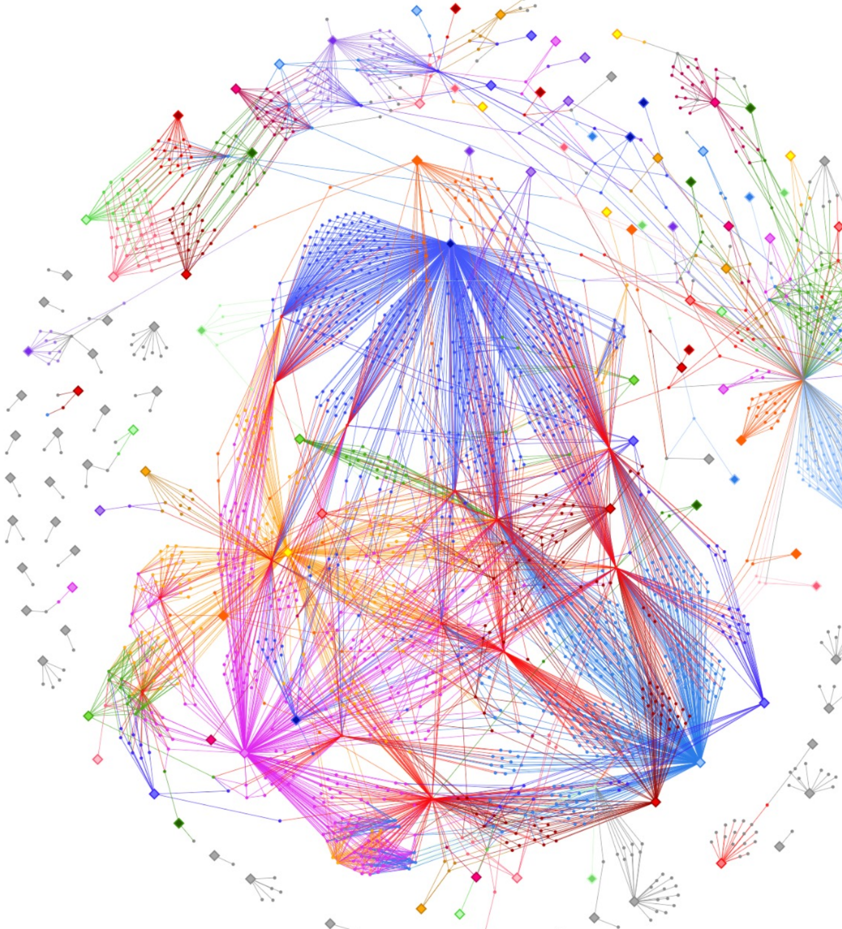
The European XFEL Data Stack

Steffen Hauf with contributions from across the EuXFEL Data Dept.

Head of Controls Group, European XFEL



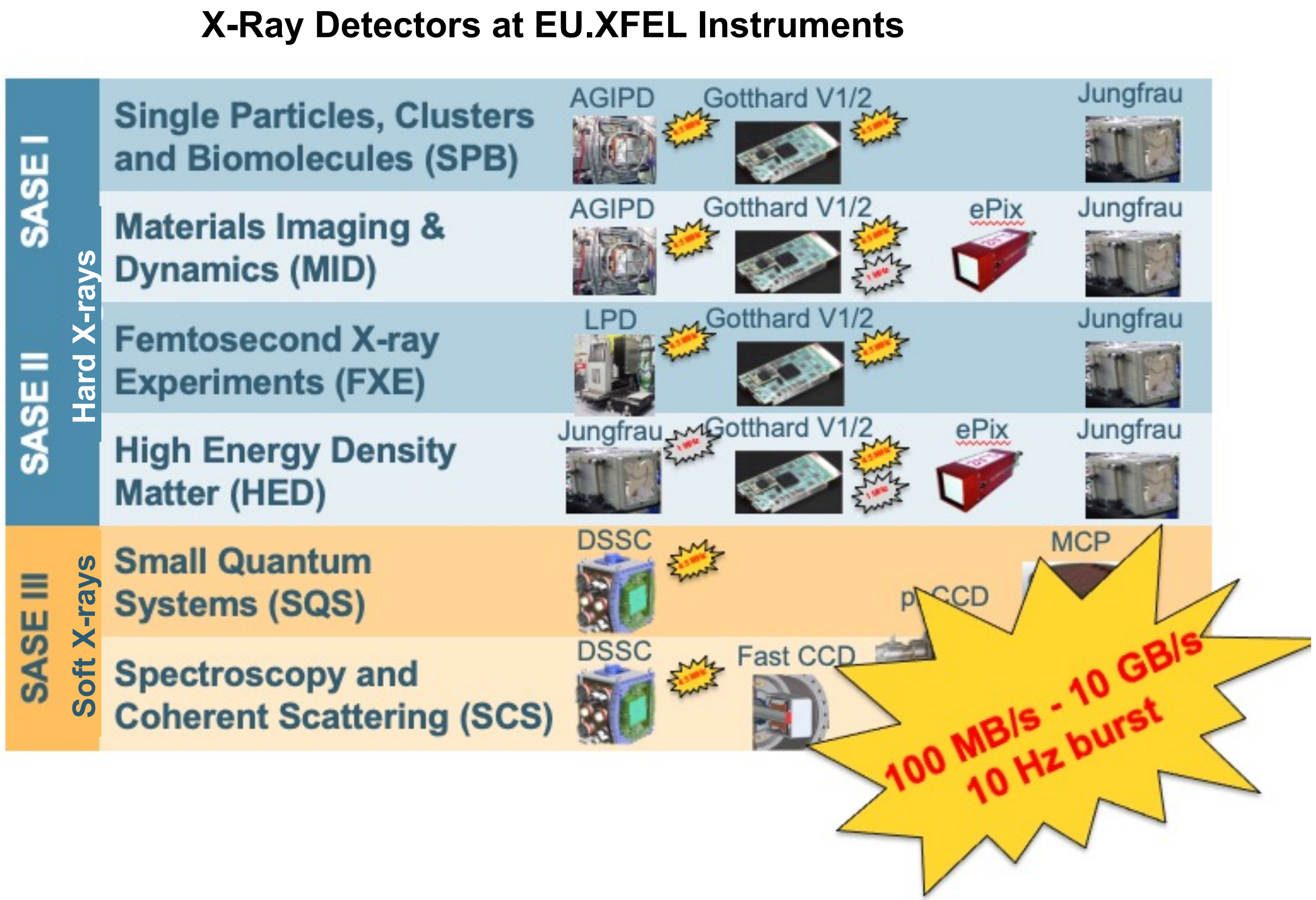




Outline

* The European XFEL – a data perspective
* The EU.XFEL Data Operation Center
* Experiences from running the Data Operation Center

The European XFEL – a Data Perspective



The European XFEL – a Data Perspective

Custom FPGA-based Data Producers at EU.XFEL

MicroTCA Crates
Large 12 slot 9U and small 6 slot 2U (including MCH, Power Supply and CPU)

X2Timer
XFEL Timing System module for synchronization (clocks and triggers) and pulse parameters from NAT

DAMC2
Required for Clock & Control system for fast 2D detectors, VETO System, Machine Protection System and photon beam loss monitors from DESY

SIS8300
Fast 125MSPS ADC with 10 channels and 16bit resolution for diagnostics and detectors from Struck Innovative Syst…e

ADQ412/ADQ14/ADQ7
High-speed digitizers from 1.8GSPS to 10GSPS with 12 to 14 bit resolution from Te…yne SP Devices

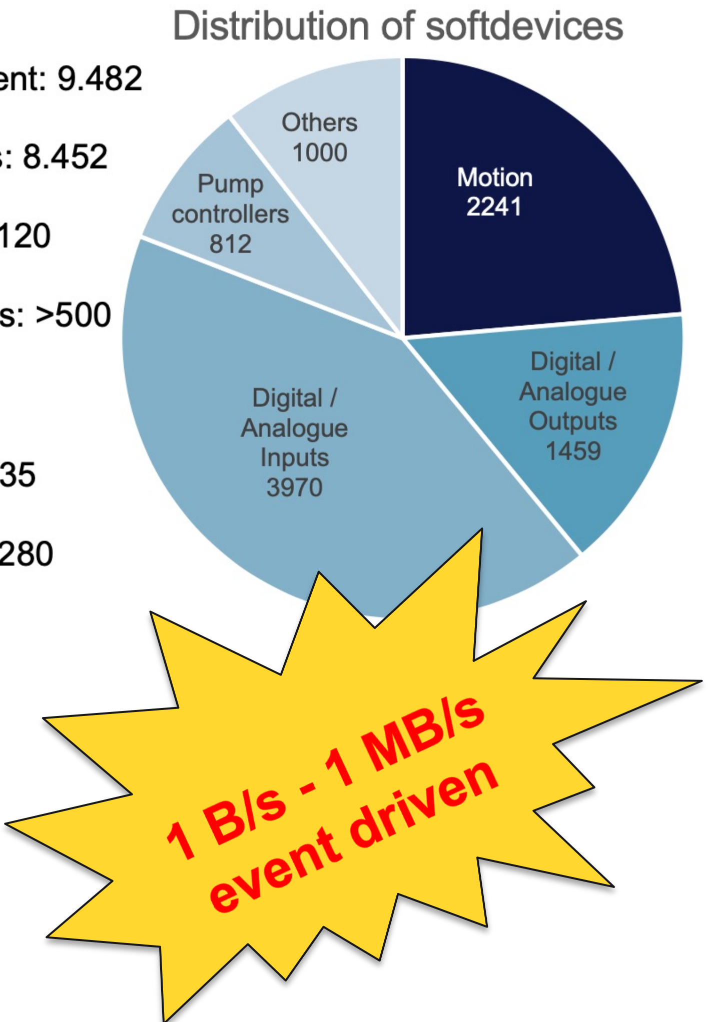
1 MB/s - 1 GB/s
10 Hz burst

The European XFEL – a Data Perspective

PLC Systems at EU.XFEL

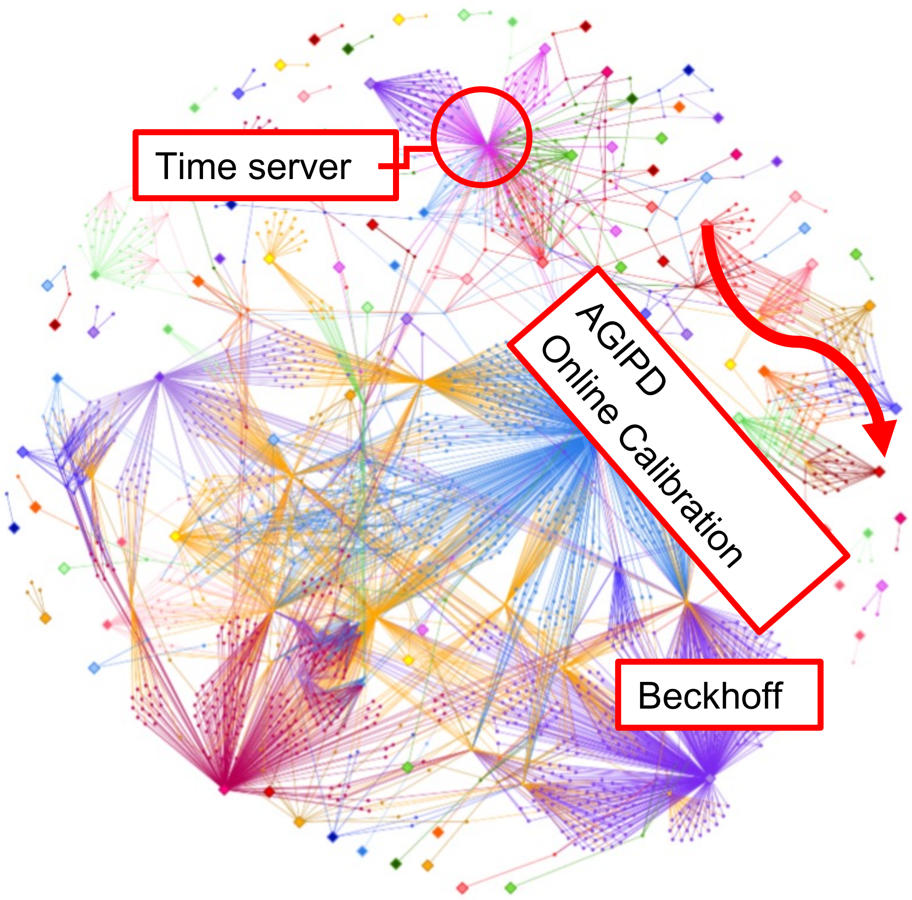
- Softdevices/Equipment: 9.482
- Number of Terminals: 8.452
- PLC Loops / CPUs: 120
- PLC Modules / crates: >500

- MicroTCA Systems: 35
- Digitizer channels: >280



The European XFEL – a Data Perspective

The Karabo Control System for the MID instrument

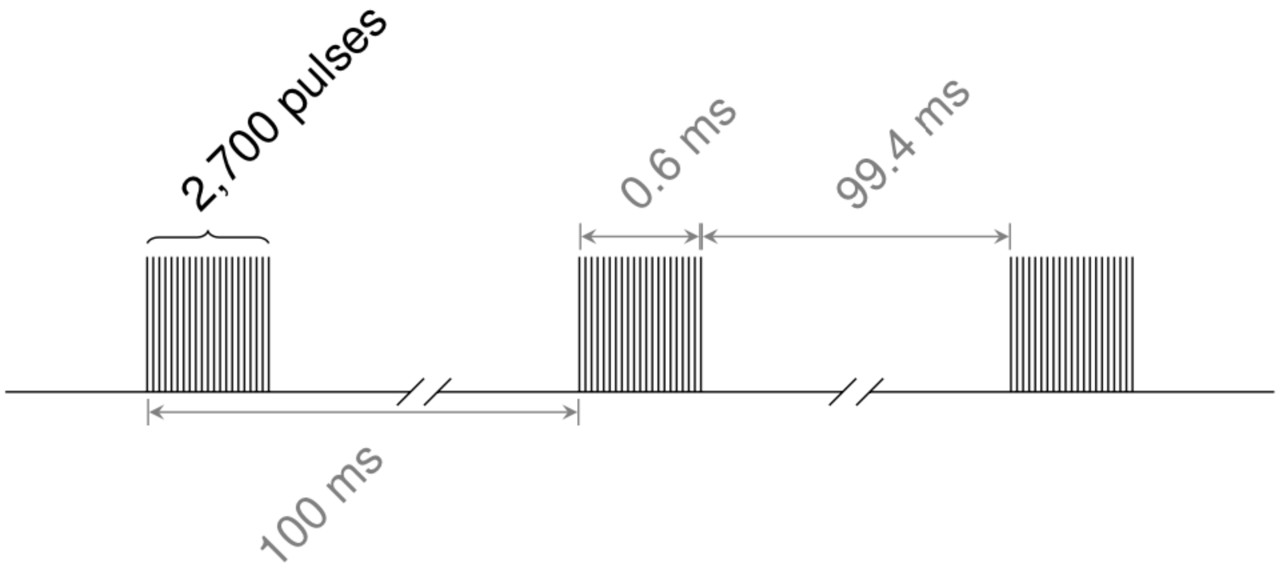


topic	# properties	# devices
LA1	26914.0	257.0
LA2	28103.0	250.0
LA3	32013.0	298.0
SA1	192656.0	1588.0
SA2	274426.0	2260.0
SA3	233216.0	1886.0
FXE	121905.0	905.0
HED	268688.0	1892.0
MID	258094.0	1789.0
SCS	205557.0	1444.0
SPB	267424.0	1872.0
SQS	306817.0	1954.0
Total	2215813.0	16395.0

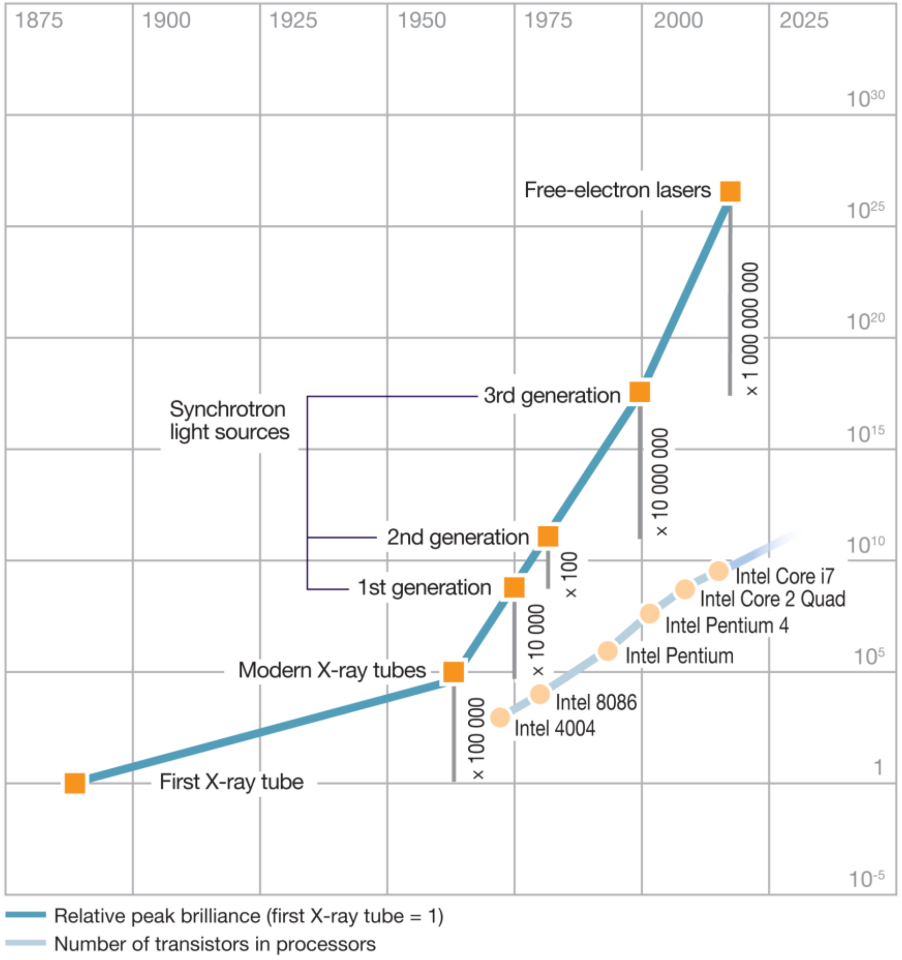
Number of control properties at EU.XFEL

European XFEL – Time Structure

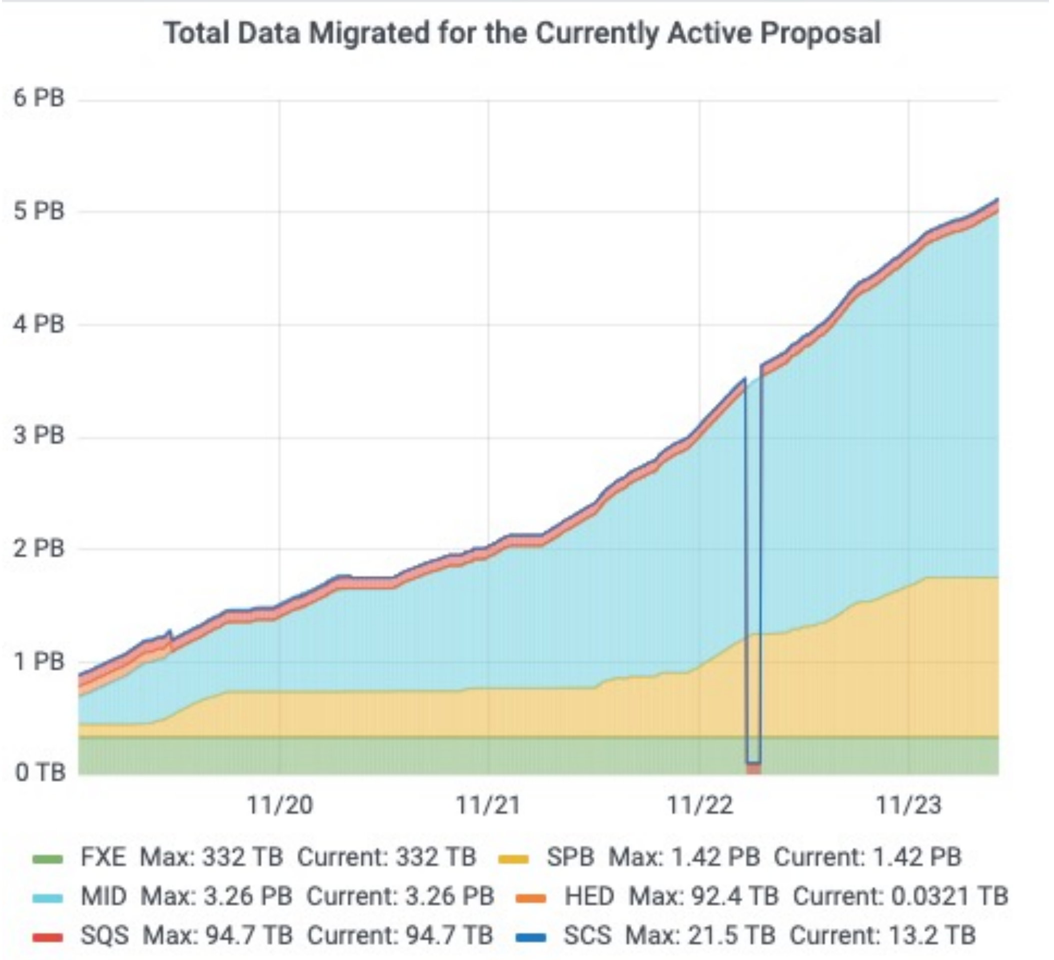
* 10 Hz base structure
* Up to 4.5 MHz burst
* Readout of pulse resolving detectors in the 99.4 ms gap
* → Bursts of up 20 Gb/s data for MHz detectors
* →10 Hz or event driven updates for everything else



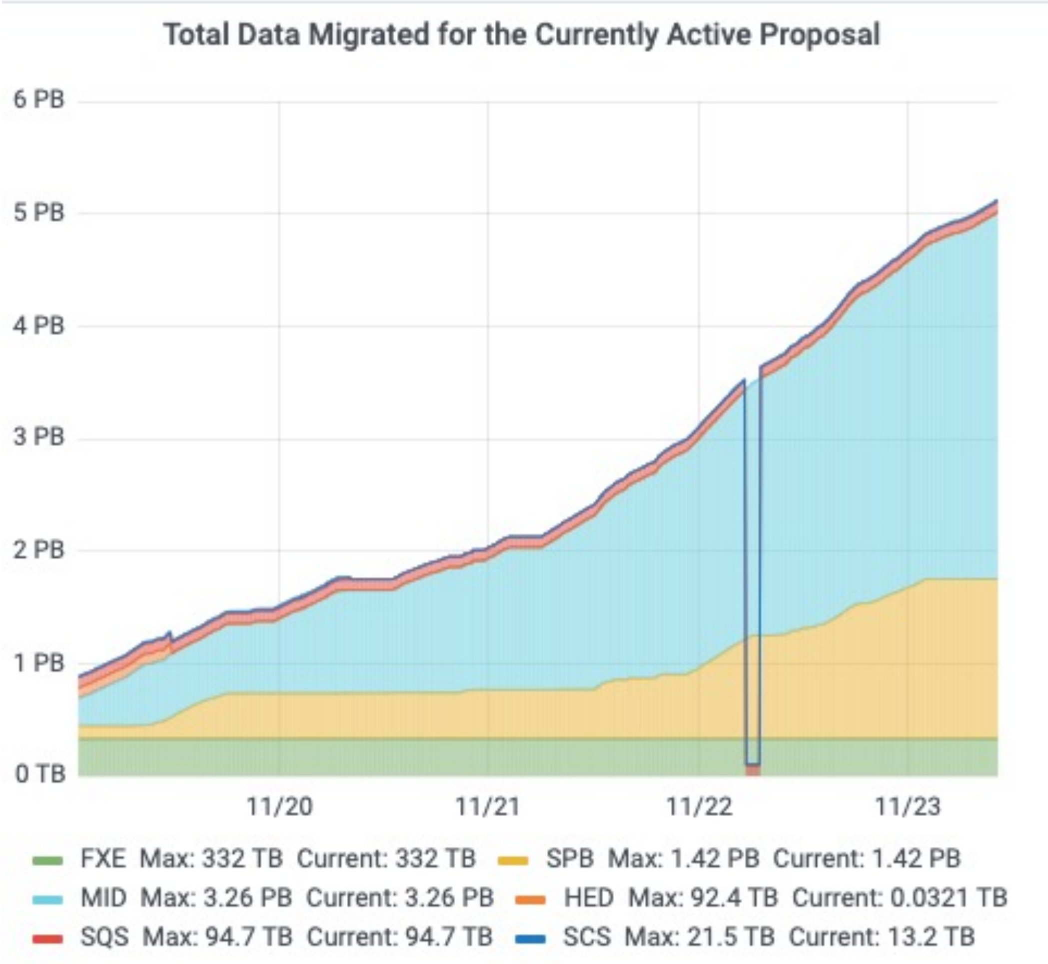
The European XFEL – a Data Perspective



Data Produced in CW46 2021 – All Data Systems Working



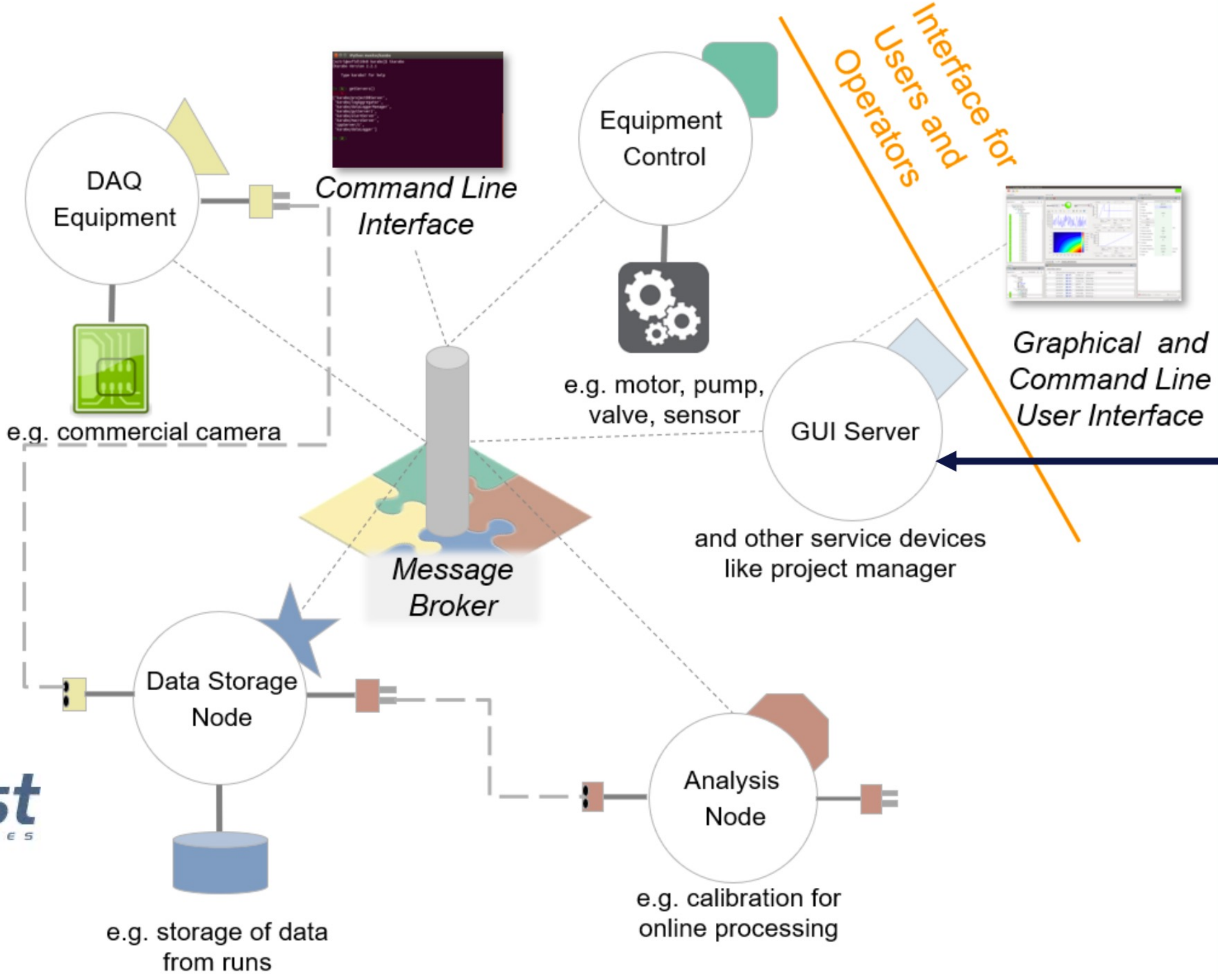
The European XFEL – a Data Perspective



If data services are working, facilities like EU.XFEL easily produce scientific data in the Peta-Byte range in a few days

Karabo - Architecture

* Central Message Broker (Control and slow data)
  * Currently: OpenMQ
  * Soon RabbitMQ.
* Event driven:
  * Data propagates through the system when values change – push not poll
* Message driven:
  * Signal – Slot paradigm
  * Asynchronous core, synchronous convenience in middleware

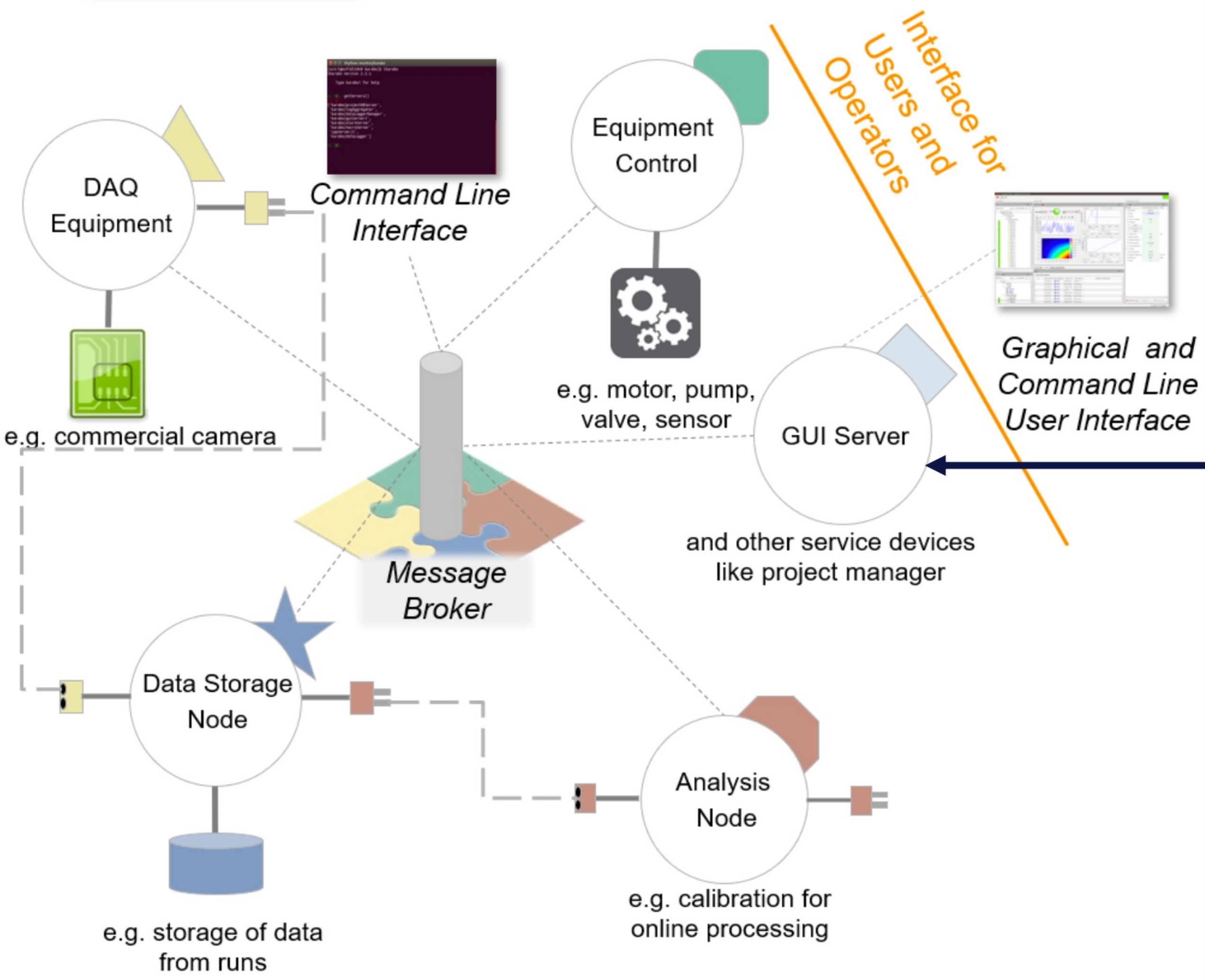


Karabo - Architecture

* pipeline (p2p) connections (scientific/large) data
  * Scatter/Gather/Copy/Distribute
  * Block/Drop on congestion
  * TCP
  * Also GUI Server – GUI client
  * Capable of saturating a 10G line

* Dynamic, discoverable topology
  * No central database instance

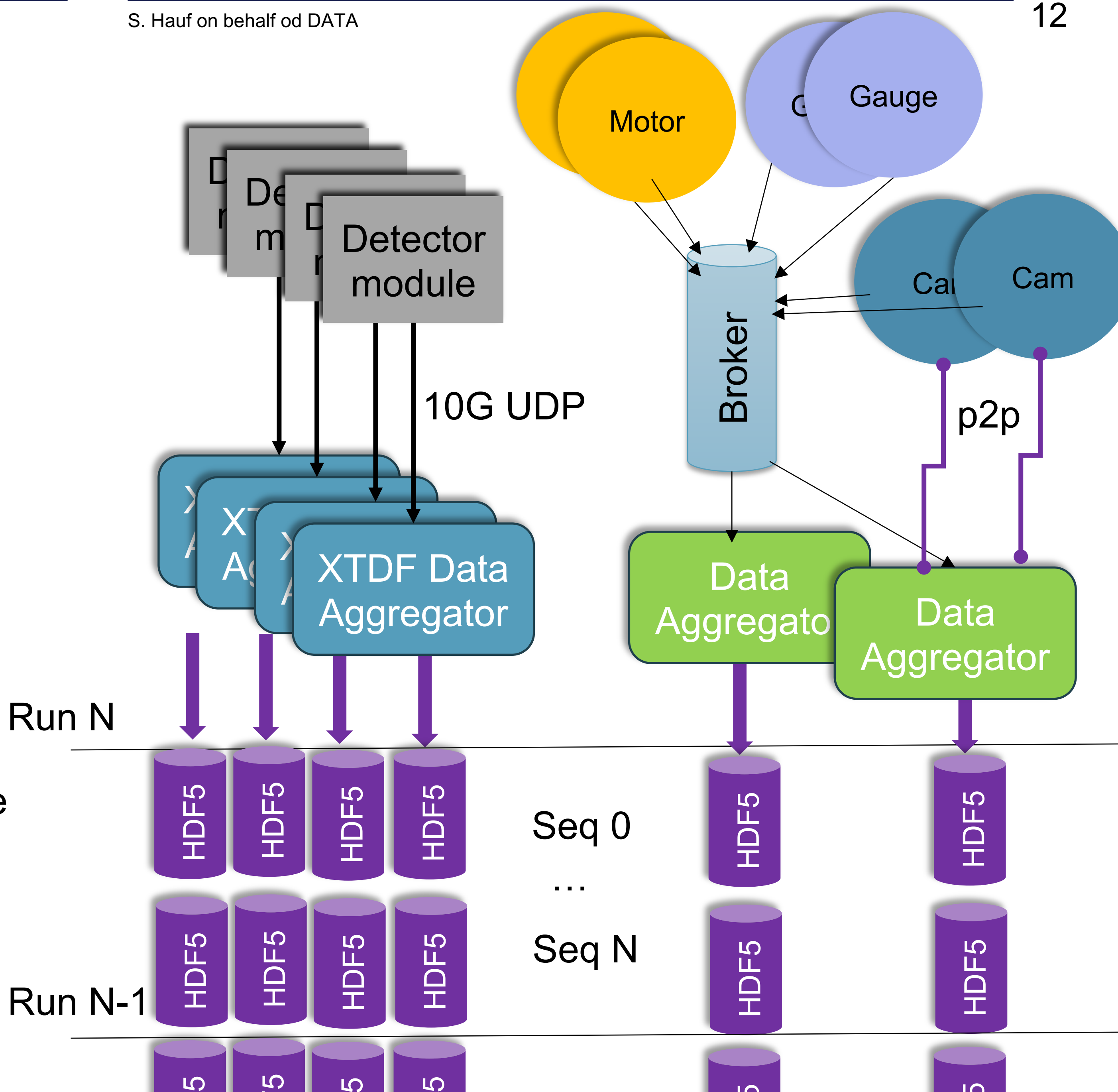
* GUI Server:
  * Gateway to the Control system



Centralized Data Acquisition

* Central DAQ instance per instrument/installation
* Distributed "Data Aggregators"
  * 1 per MHz detector module (e.g. 16 total)
  * ~5-10 for remaining data
    - ► Fast pipeline data: Imagers, digitizers
    - ► Slow broker data
* Data acquisition is run-based. A run is a period of (semi-) static experimental conditions
  * Importantly, data shapes* and Karabo device schema do not change

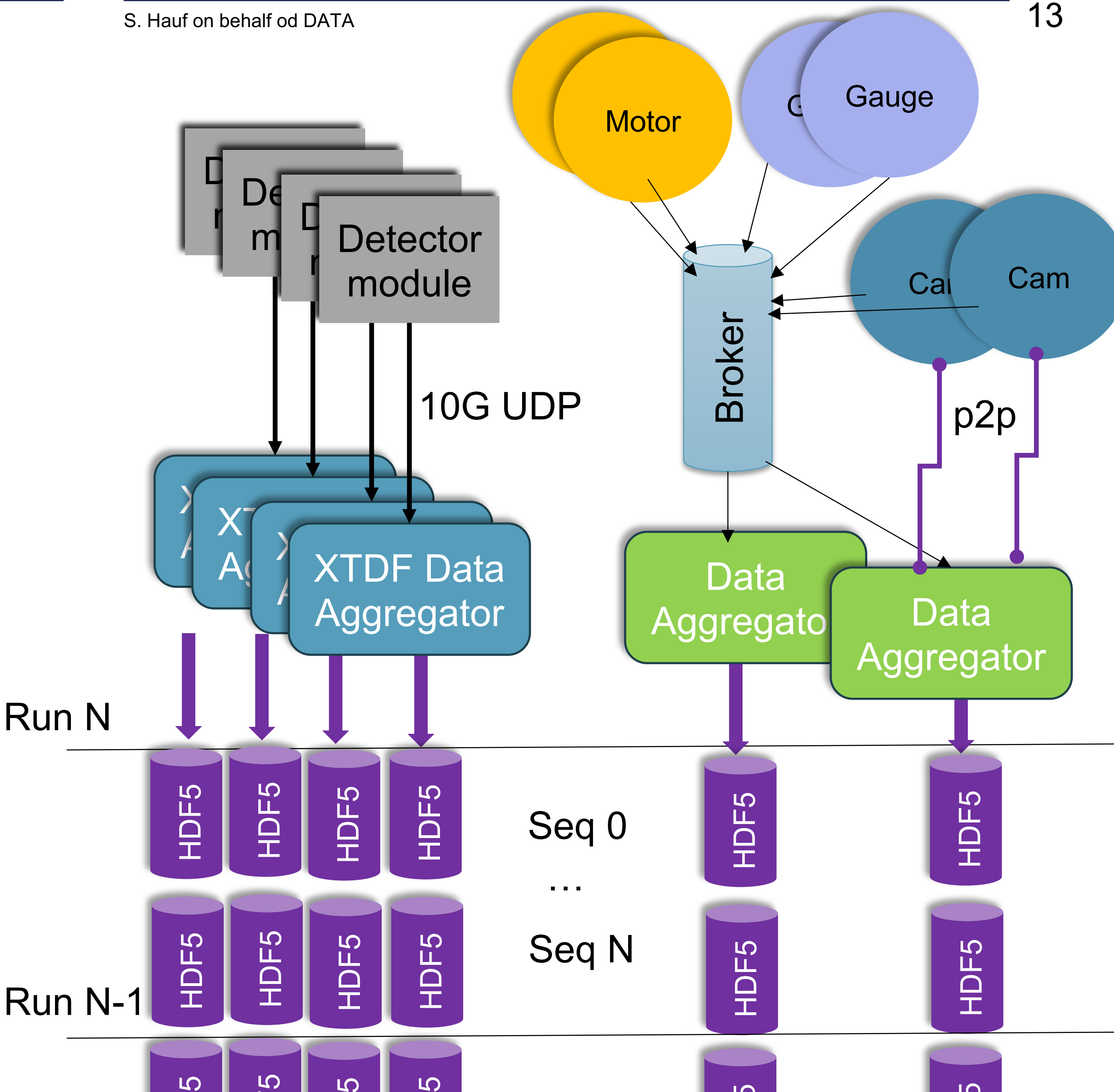
* Except along pulse axis



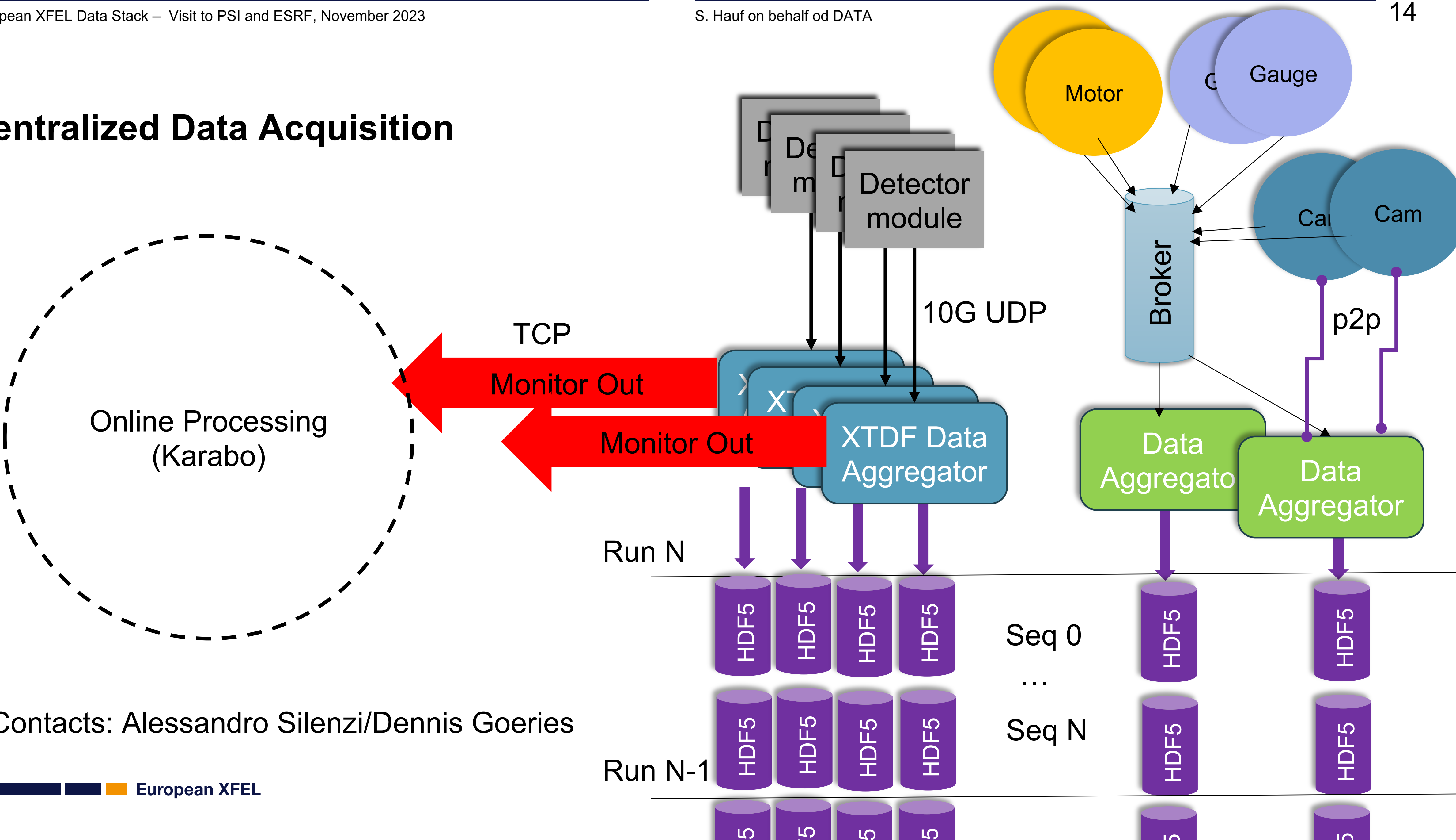
Centralized Data Acquisition

* Implemented as Karabo Devices (C++)
* Up to ˜30 GB/s for a full installation
* Single aggregator can handle a few 1000 slow control properties
* Within one sequence file all sources are aligned by train id in the rows of the HDF5 tables
* Current work: refactoring to facilitate integration of online data reduction using RDMA offloading of processing.

Contacts: Alessandro Silenzi/Dennis Goeries

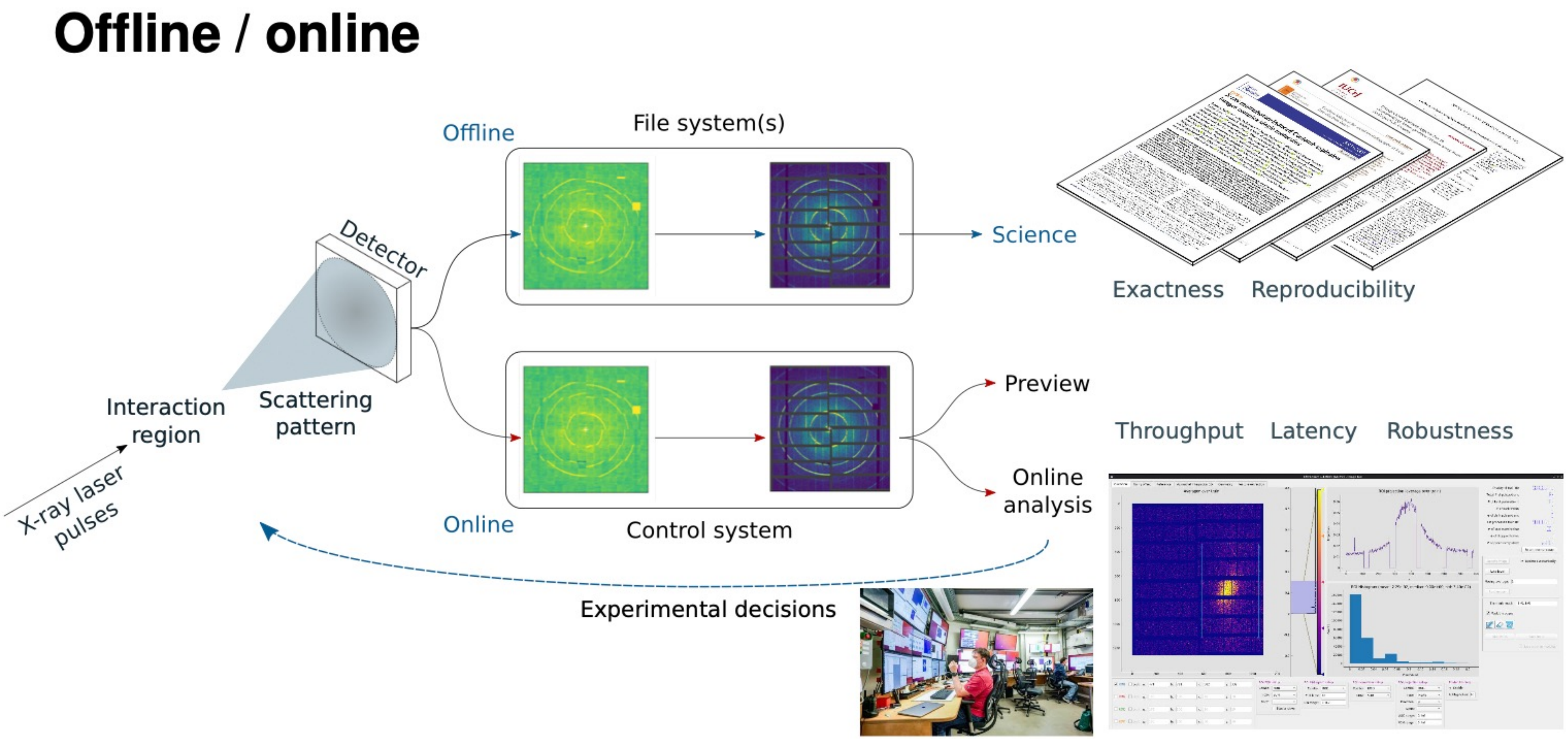


Centralized Data Acquisition



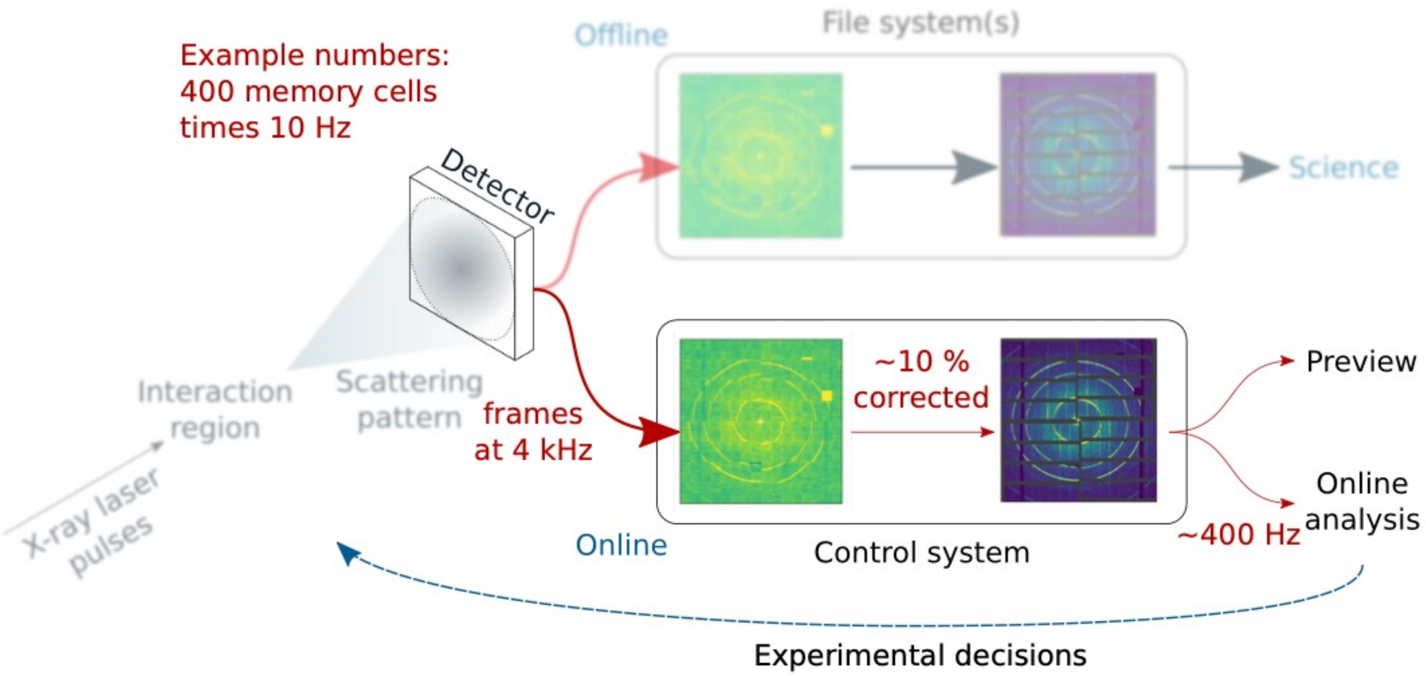
Contacts: Alessandro Silenzi/Dennis Goeries

Importance of Facility-Side Processing: Calibration



Importance of Facility-Side Processing: Online Calibration

Bottleneck: initial calibration pipeline



Approximate bandwidth needed for image data + gain information only:

detector	cells	bandwidth	per module
JUNGFRAU 4M	16	16 Gbit/s	2 Gbit/s
DSSC 1M	400	67 Gbit/s	4.2 Gbit/s
LPD 1M	512	86 Gbit/s	5.4 Gbit/s
AGIPD 1M	352	118 Gbit/s	7.4 Gbit/s
DSSC 1M	800	134 Gbit/s	8.4 Gbit/s

Can correction be made to keep up?

Throughput
Latency
Robustness

Contact: David Hammer (Data Analysis)

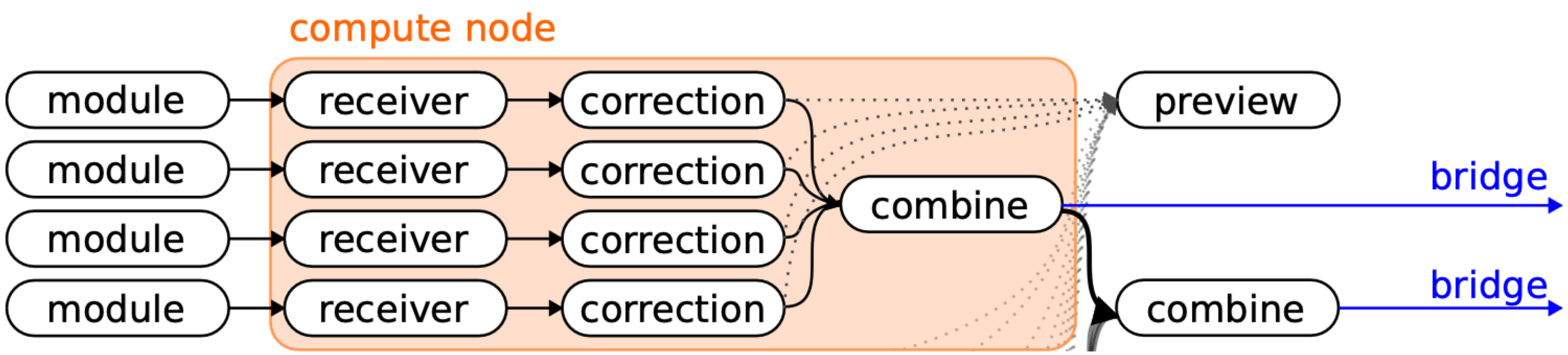
Importance of Facility-Side Processing: Online Calibration

- Sufficient I/O bandwidth to pass single module data into calibration host
- Once data is on the host, use shared memory and GPU memory to link different correction steps
- Sufficient I/O bandwidth to pass single module data from calibration host to online processing code.

Approximate bandwidth needed for image data + gain information only:

detector	cells	bandwidth	per module
JUNGFRAU 4M	16	16 Gbit/s	2 Gbit/s
DSSC 1M	400	67 Gbit/s	4.2 Gbit/s
LPD 1M	512	86 Gbit/s	5.4 Gbit/s
AGIPD 1M	352	118 Gbit/s	7.4 Gbit/s
DSSC 1M	800	134 Gbit/s	8.4 Gbit/s

Can correction be made to keep up?

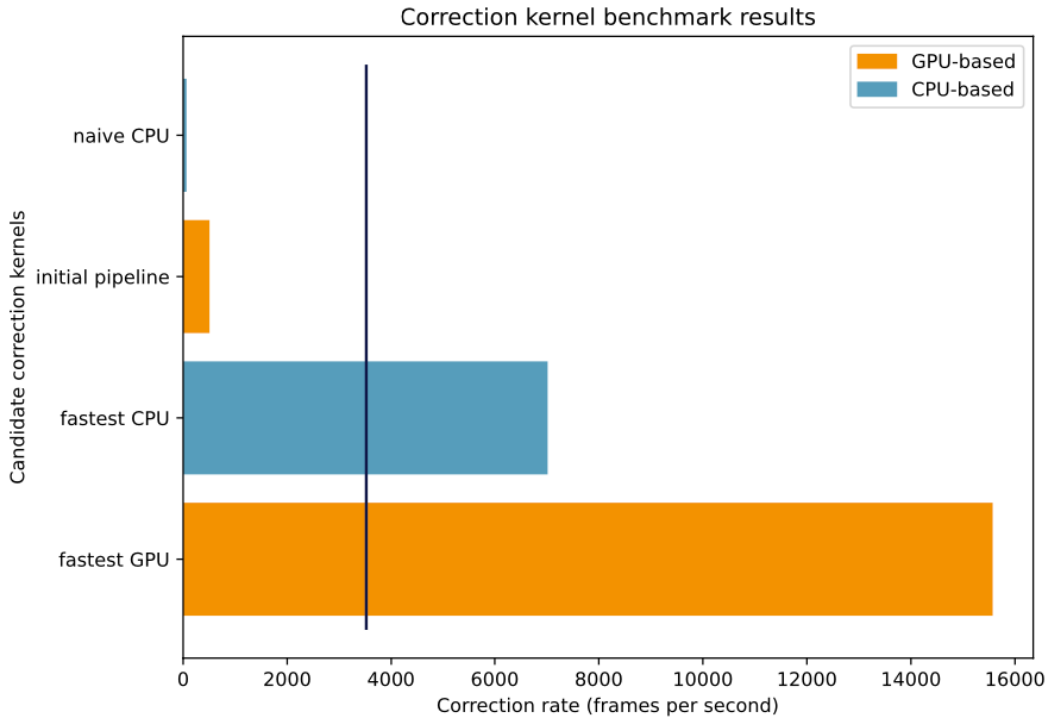


Importance of Facility-Side Processing: Online Calibration

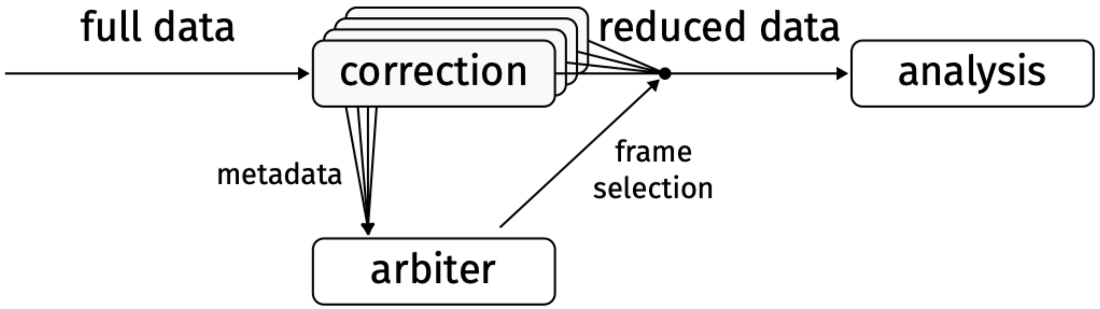
Benchmarks

Initial findings:

- Correction at full speed feasible
  - Test setting: AGIPD with 352 memory cells
- GPUs useful for big detectors
- Can even process more modules per node



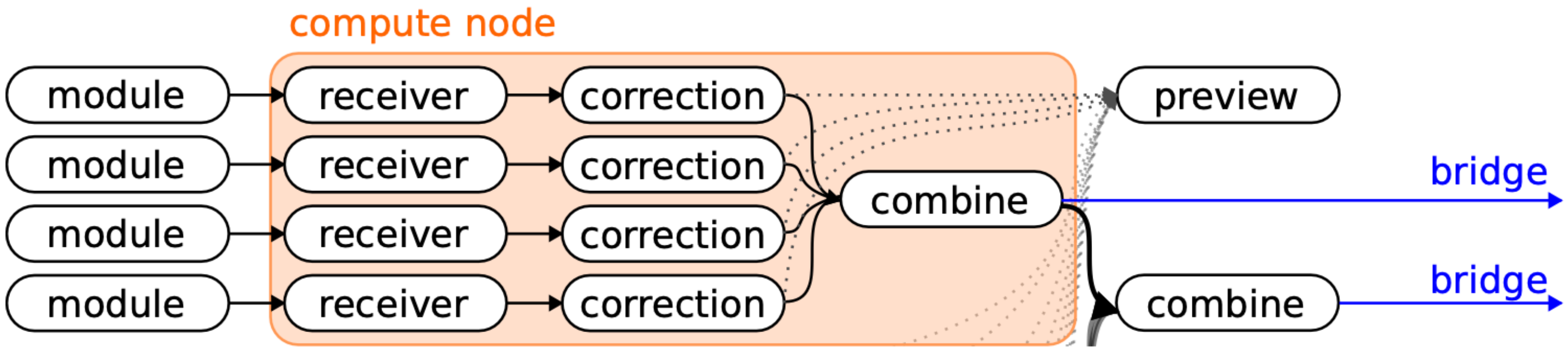
Importance of Facility-Side Processing: Online Calibration and Data Reduction



Approximate bandwidth needed for image data + gain information only:

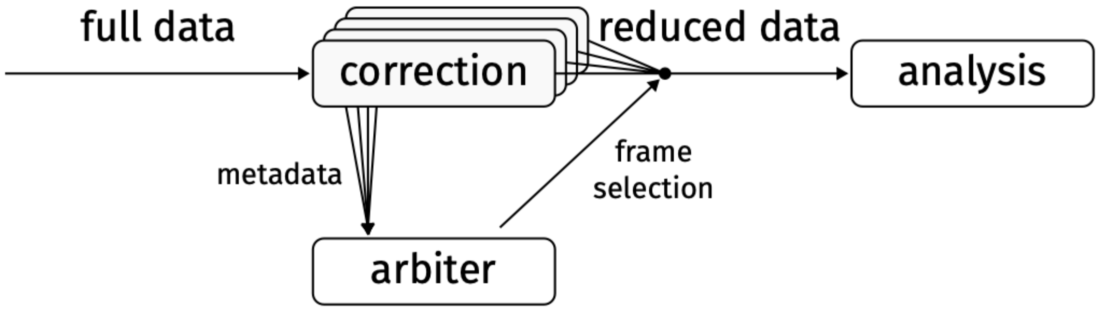
detector	cells	bandwidth	per module
JUNGFRAU 4M	16	16 Gbit/s	2 Gbit/s
DSSC 1M	400	67 Gbit/s	4.2 Gbit/s
LPD 1M	512	86 Gbit/s	5.4 Gbit/s
AGIPD 1M	352	118 Gbit/s	7.4 Gbit/s
DSSC 1M	800	134 Gbit/s	8.4 Gbit/s

Can correction be made to keep up?



Contact: Egor Sobolev (Data Analysis)

Importance of Facility-Side Processing: Online Calibration and Data Reduction



Approximate bandwidth needed for image data + gain information only:

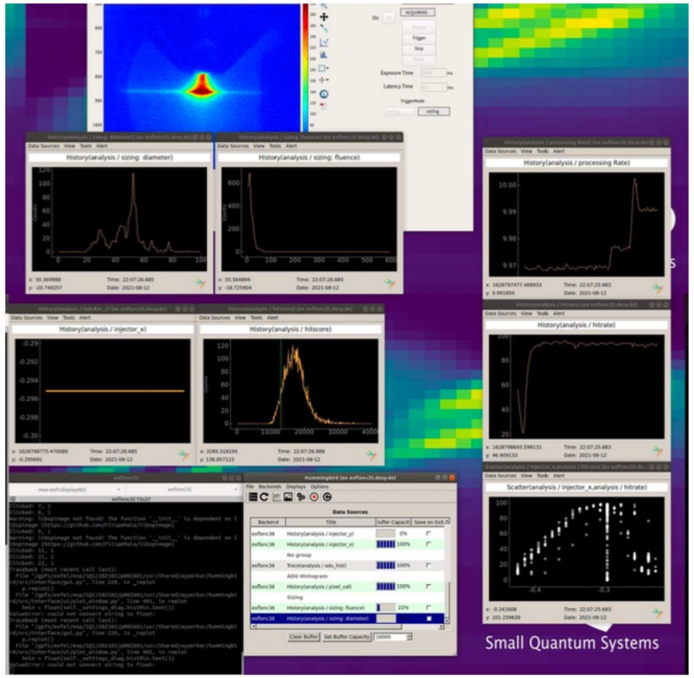
detector	cells	bandwidth	per module
JUNGFRAU 4M	16	16 Gbit/s	2 Gbit/s
DSSC 1M	400	67 Gbit/s	4.2 Gbit/s
LPD 1M	512	86 Gbit/s	5.4 Gbit/s
AGIPD 1M	352	118 Gbit/s	7.4 Gbit/s
DSSC 1M	800	134 Gbit/s	8.4 Gbit/s

Can correction be made to keep up?

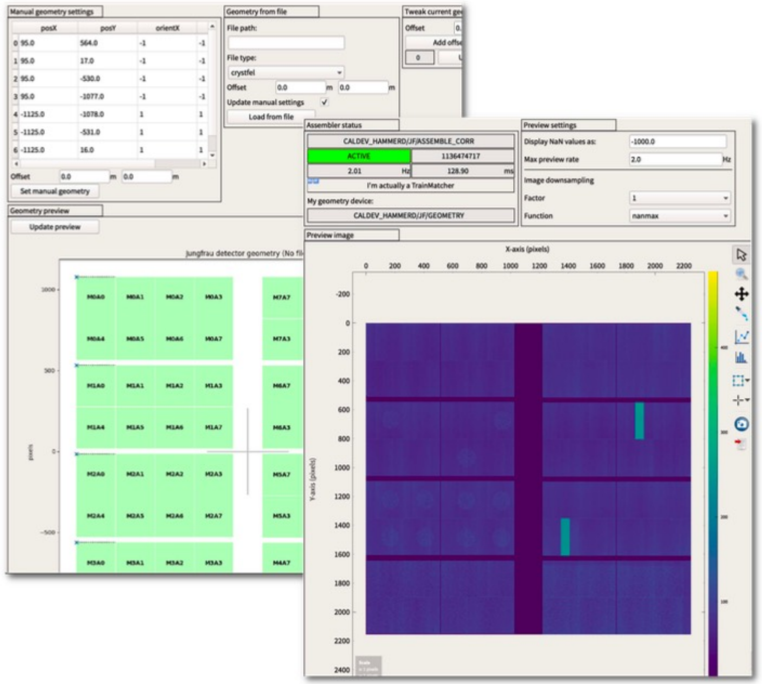
- Sufficient I/O bandwidth to pass single module data into calibration host
- However, data serialization to transfer via TCP takes too long to scale to all modules of a Mpixel detector
- Solution (work in progress): RDMA via Infiniband: 80Gbit/s on 100 Gbit/s link from Karabo device to Karabo device for ~100MB images stacks.

Importance of Facility-Side Processing: Online Processing Tools (Gallery)

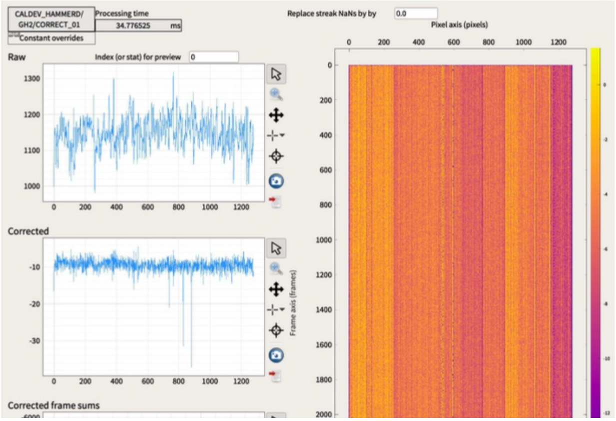
DSSC: 4000 frames/s into Hummingbird

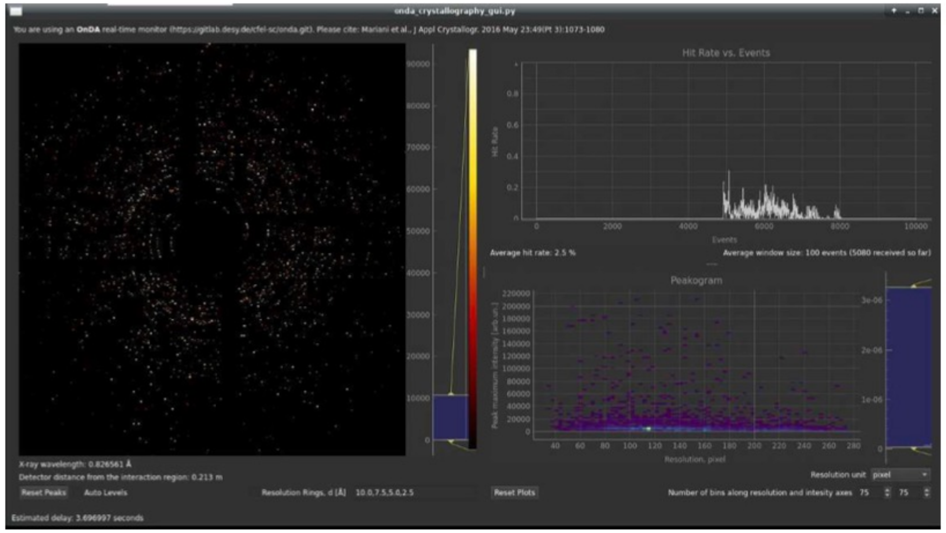


Jungfrau 4M: geometry assembly



Gotthard II: Commissioning GUI

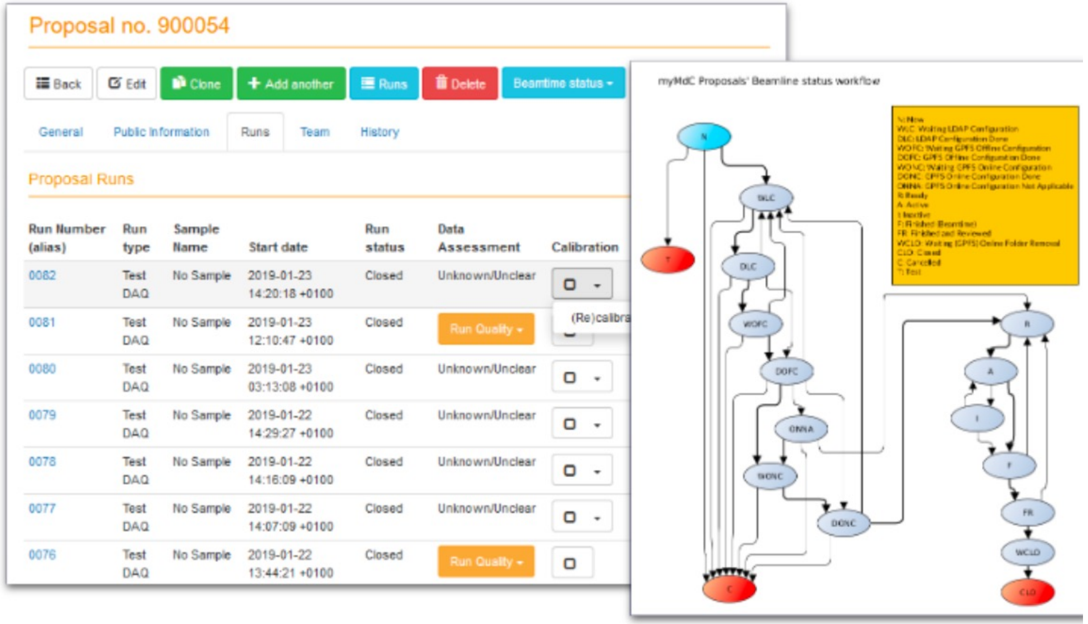




AGIPD and LPD into OnDA

Data Management Activities using myMdC

- Preparation for beamtime (storage folders, LDAP,…)
- Catalog and document the data generated
- Data migration services
- Keep track of data location
- Allow data publishing and export
- Manage permissions and accesses to the data in an uniform manner and guarantee that only authorized people have access to the data

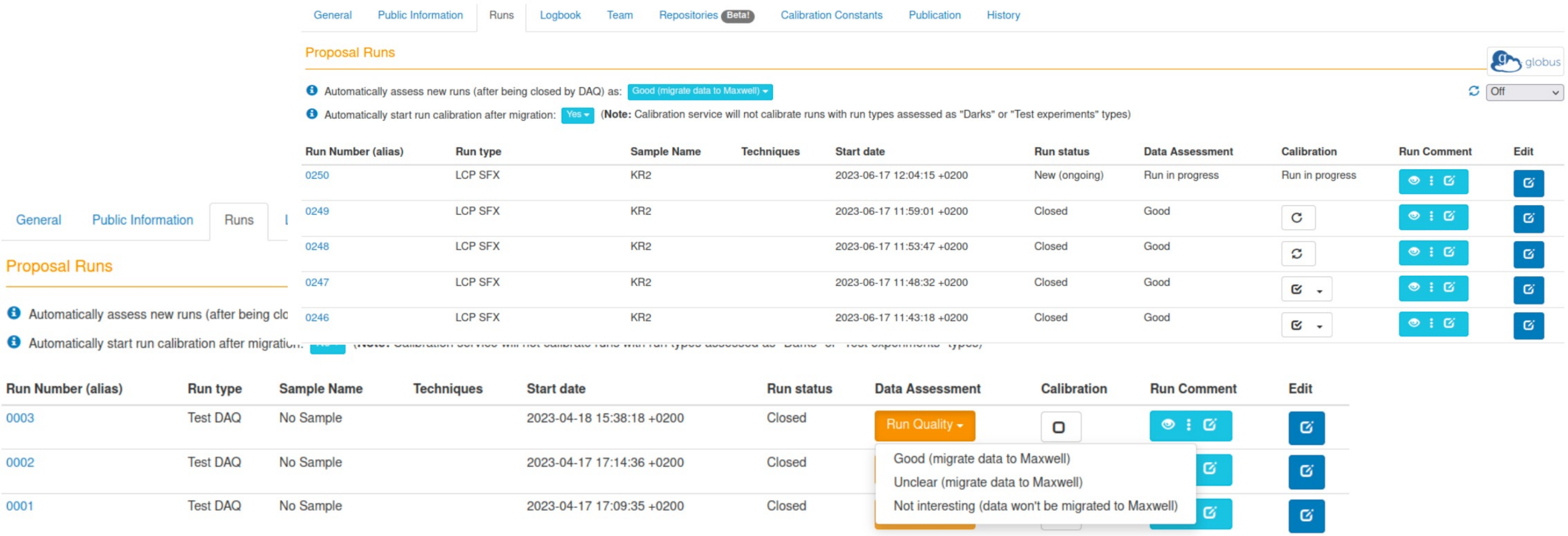


- Data Retention Policies control and execution
  - Data Quality Assessment
  - Data Calibration
  - Data Archiving
  - Data Cleanup

Contact: Luis Maia (ITDM)

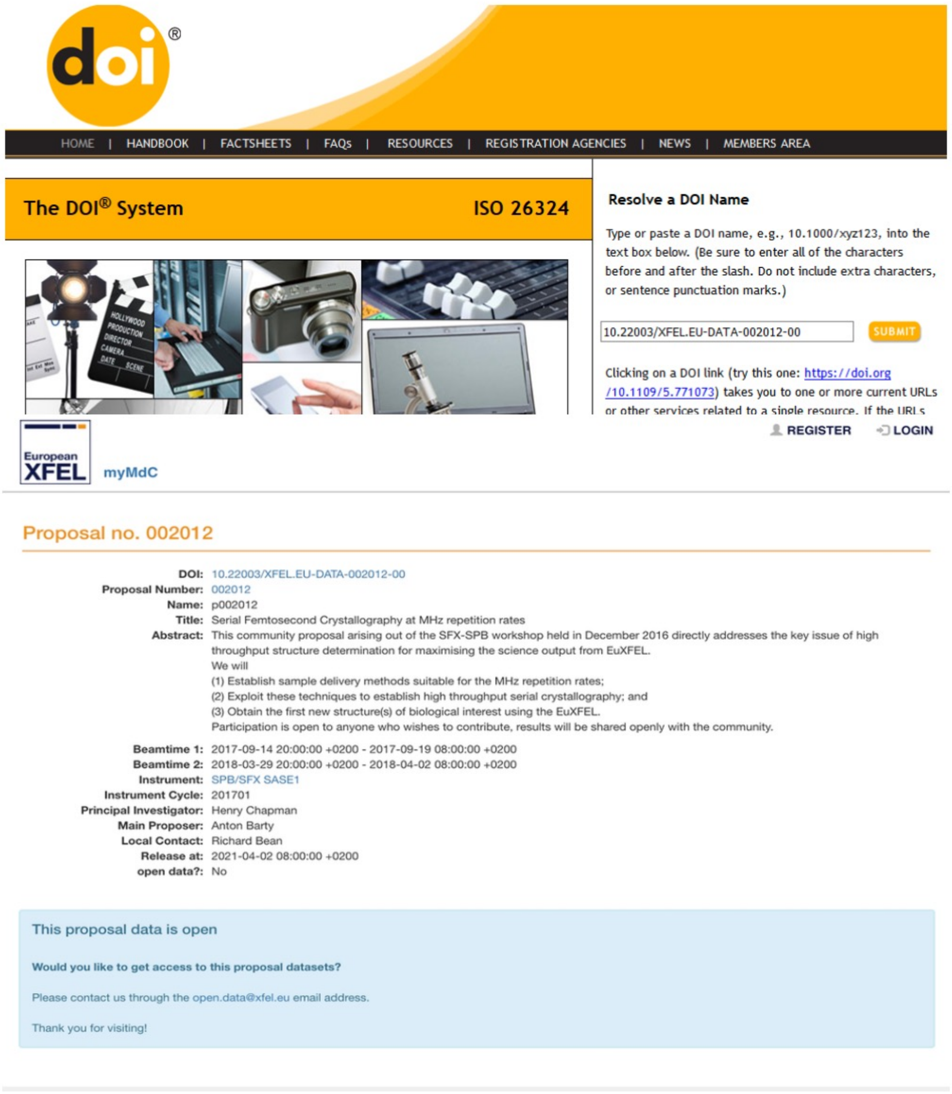
Data Management and Data Migration

- Runs which are assessed as Good or Unclear are preserved long term and accessible from Maxwell cluster and by export services.

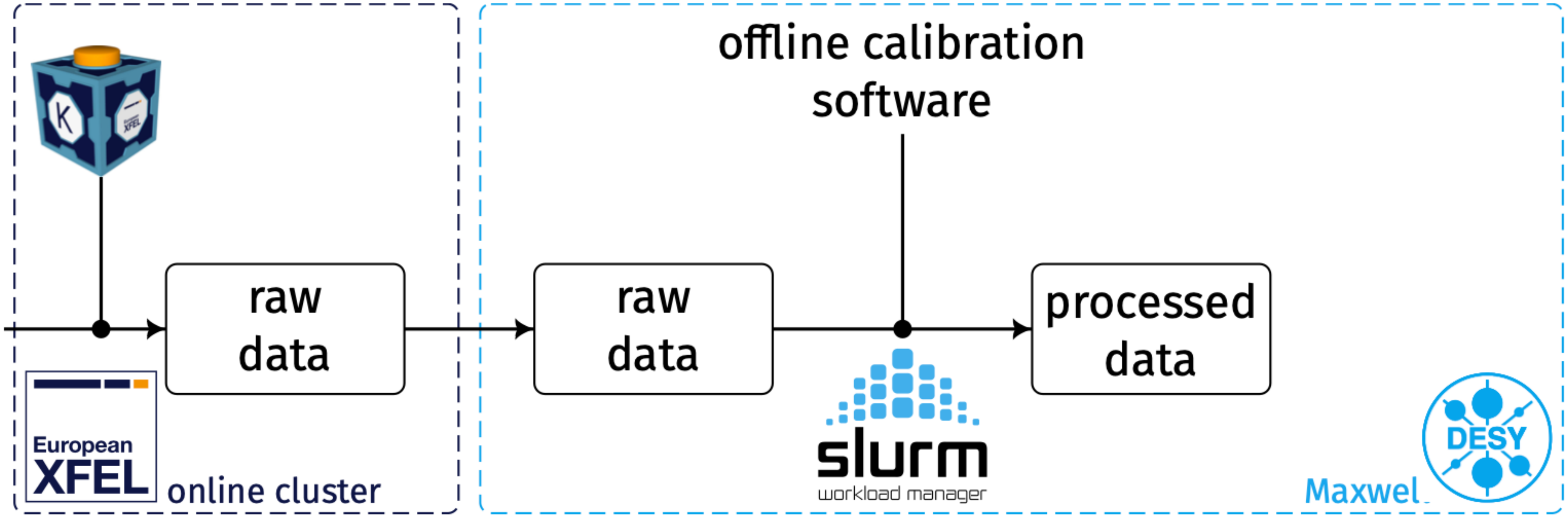


DOI workflow

- DOI automatically registered for proposal data after proposal reached status Finished and Reviewed

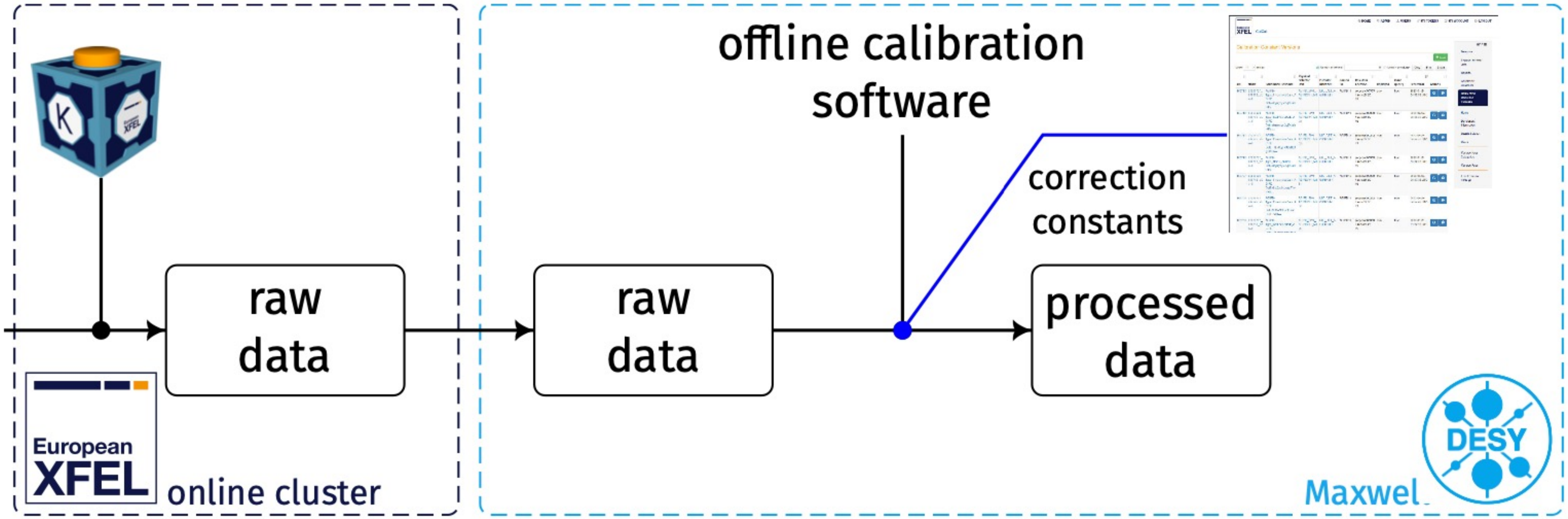


Importance of Facility-Side Processing: Offline Calibration



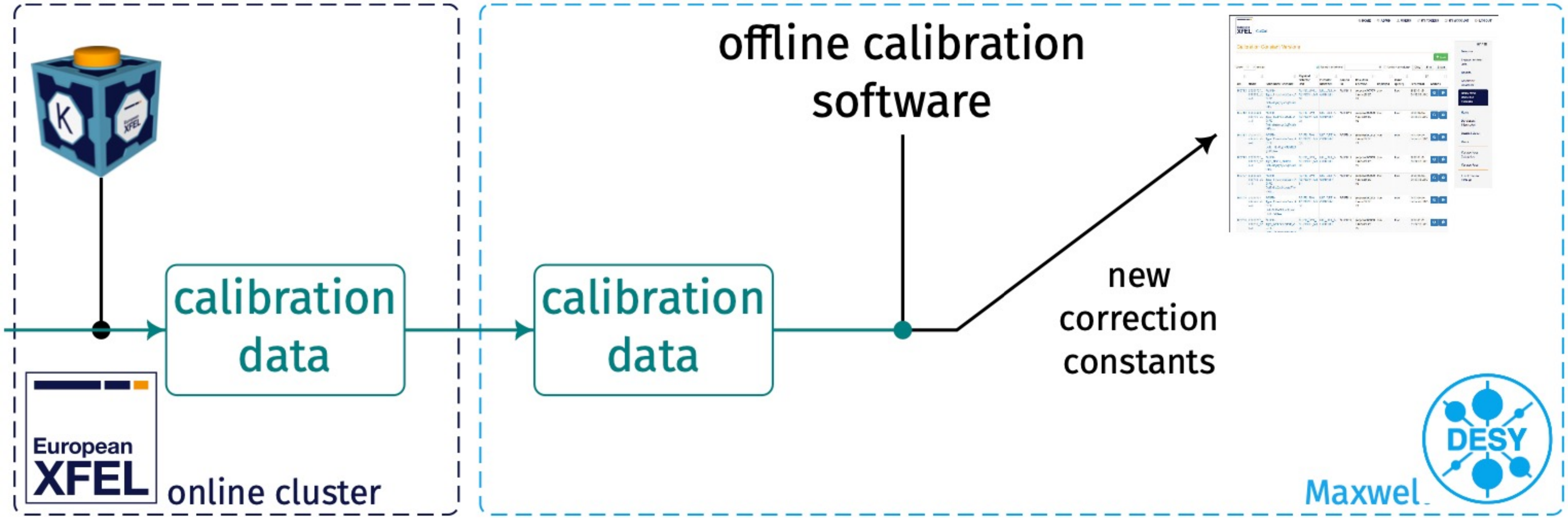
Contact: Philipp Schmidt (Data Analysis)

Importance of Facility-Side Processing: Offline Calibration



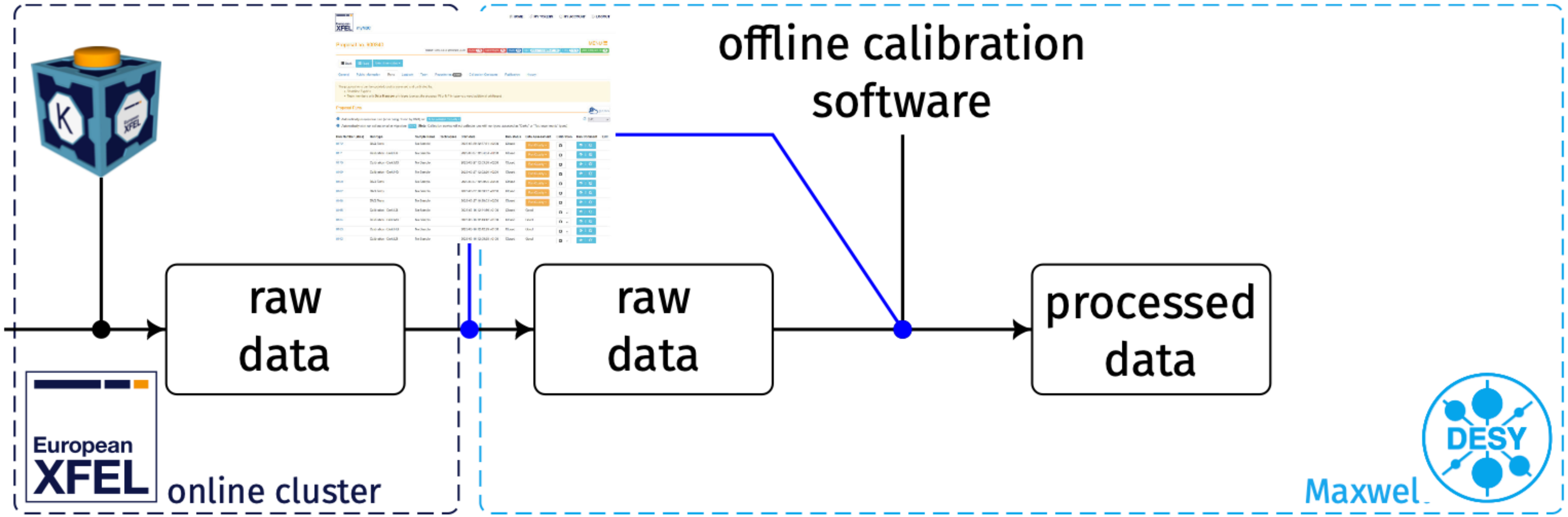
Automated processing: raw to processed data

Importance of Facility-Side Processing: Offline Calibration



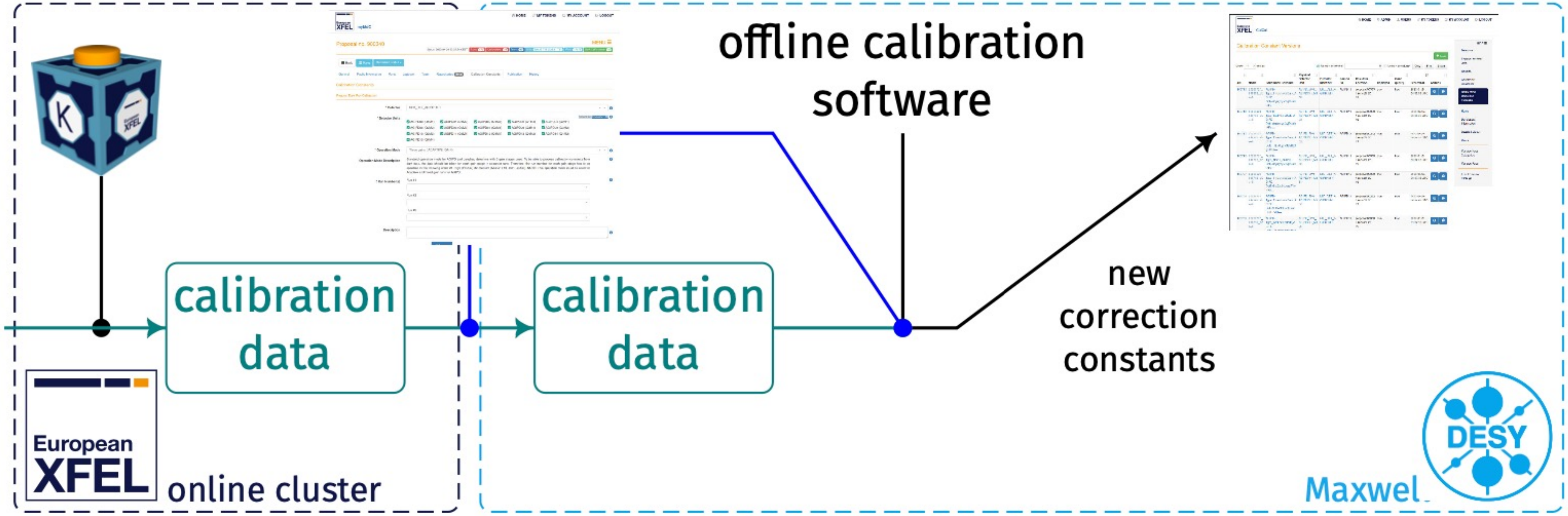
Automated processing: new correction constants

Importance of Facility-Side Processing: Offline Calibration



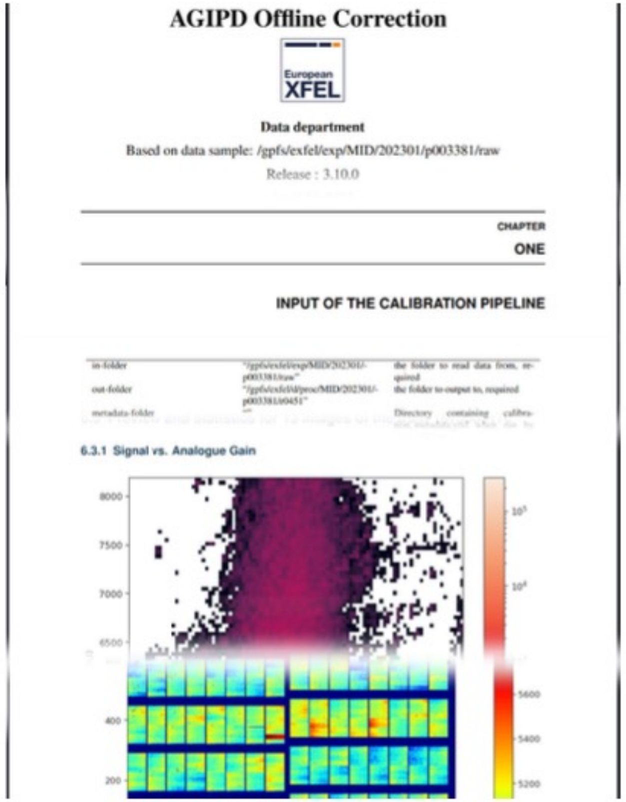
On-request processing: raw to processed data

Importance of Facility-Side Processing: Offline Calibration



On-request processing: new calibration constants

Importance of Facility-Side Processing: Offline Calibration - Tech Stack



Fully reproducable

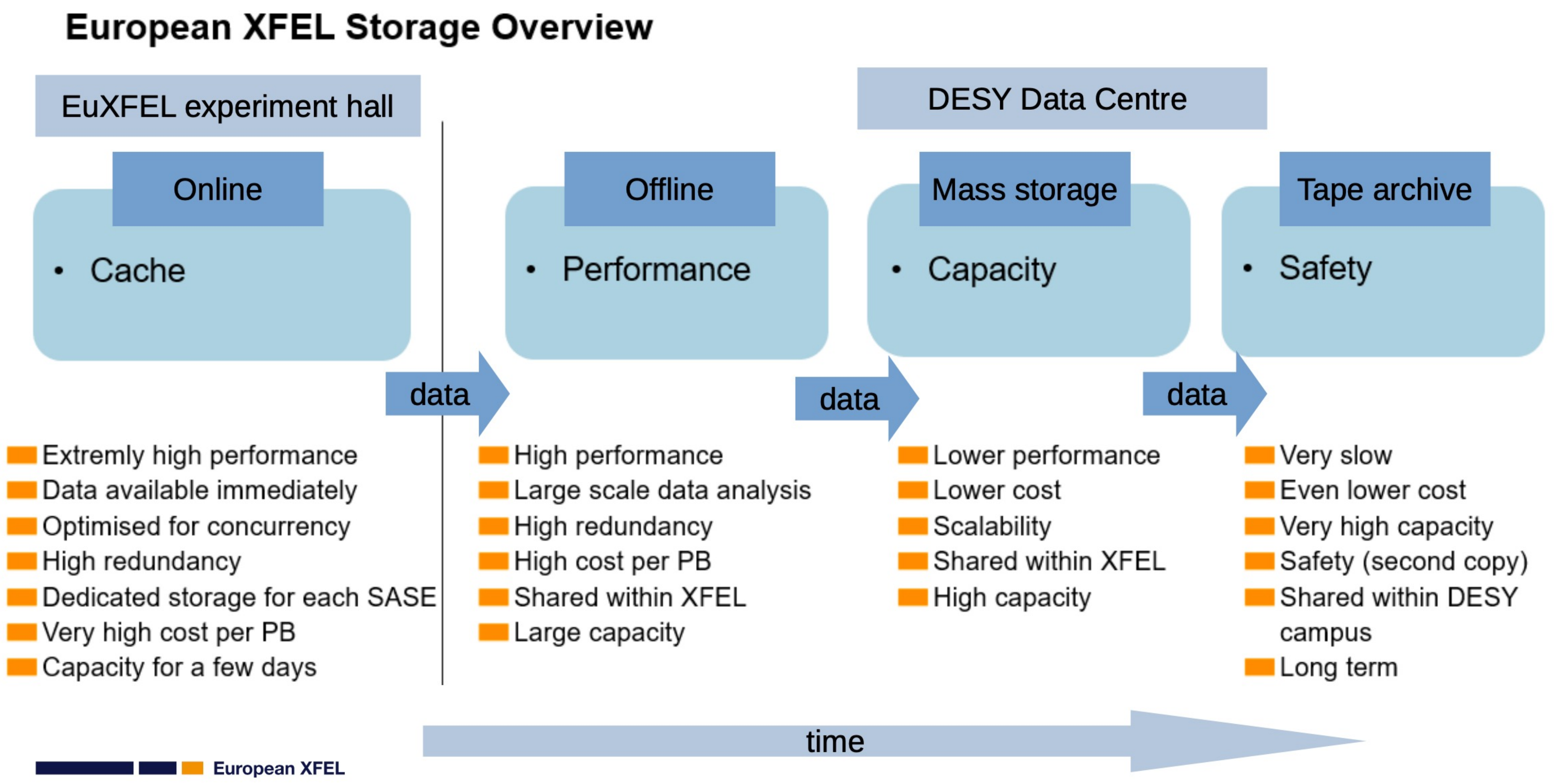
Infrastructure locations - overview

3 computer rooms (red) in the experiment hall at EuXFEL campus, Schenefeld

EuXFEL hardware in DESY Data Centre, Hamburg

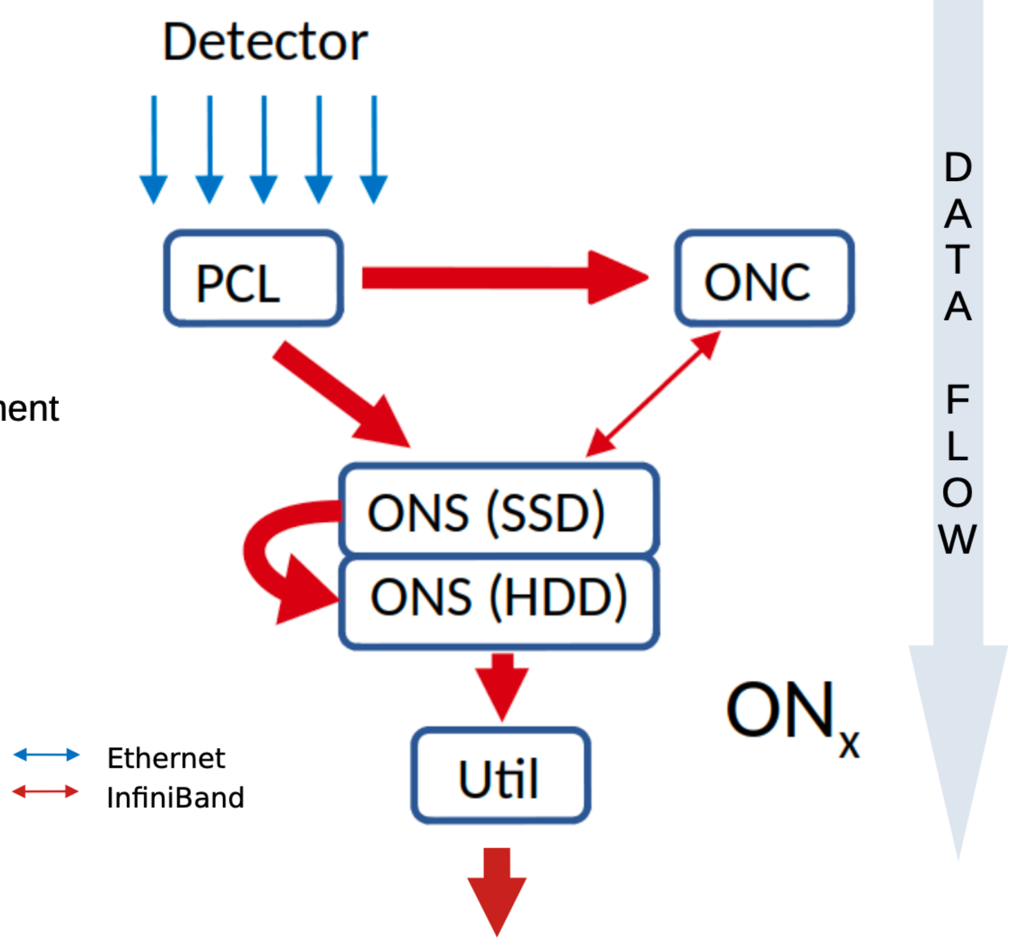
~4430m of optical fiber

Contact: Janusz Malka (ITDM)


European XFEL Storage Overview
EuXFEL experiment hall
DESY Data Centre
Online
• Cache
Offline
• Performance
Mass storage
• Capacity
Tape archive
• Safety
data
data
data
Extremly high performance
Data available immediately
Optimised for concurrency
High redundancy
Dedicated storage for each SASE
Very high cost per PB
Capacity for a few days
High performance
Large scale data analysis
High redundancy
High cost per PB
Shared within XFEL
Large capacity
Lower performance
Lower cost
Scalability
Shared within XFEL
High capacity
Very slow
Even lower cost
Very high capacity
Safety (second copy)
Shared within DESY campus
Long term
time
European XFEL

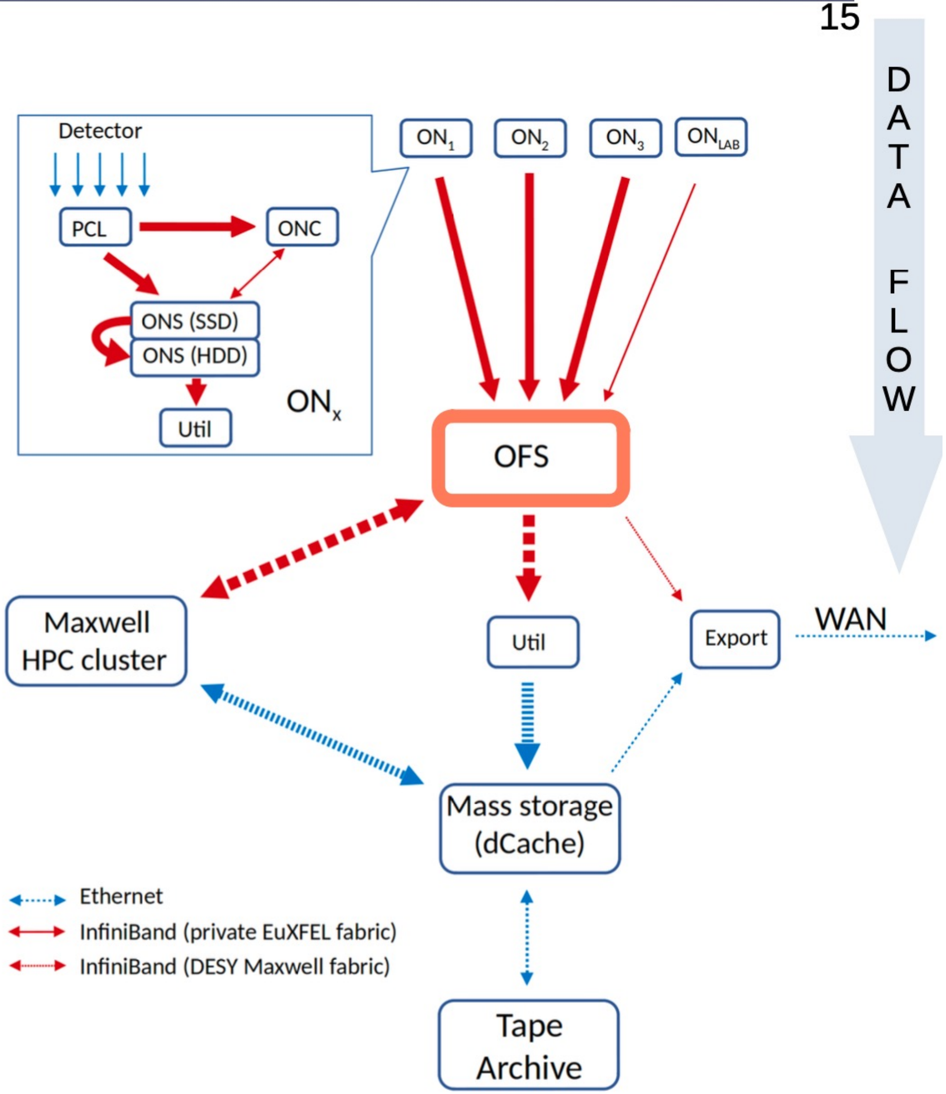
Online infrastructure (per SASE)

- PC-Layer (**DAQ** ~30nodes)
  - data aggregation from different sources
  - HDF5 file creation (up to 15GB/s)
  - data source for the on-line data analysis
- ONC – Online Computing Cluster (**CPU**, **GPU** ~20nodes)
  - captures raw data stream from DAQ
  - initial online calibration of raw data
  - online data preview and analysis provide feedback to experiment
  - first data quality assessment
  - exclusive for experiment
- ONS – Online Storage Cluster
  - storage for the raw data from DAQ
  - two layers **400**TB **SSD/NVMe** and **2.5**PB **HDD**
  - source for raw data copy to the offline storage
  - based on IBM **Storage Scale System** (f.k.a GPFS)
  - write performance (**50-60**GB/s **write**)
  - 5 file systems



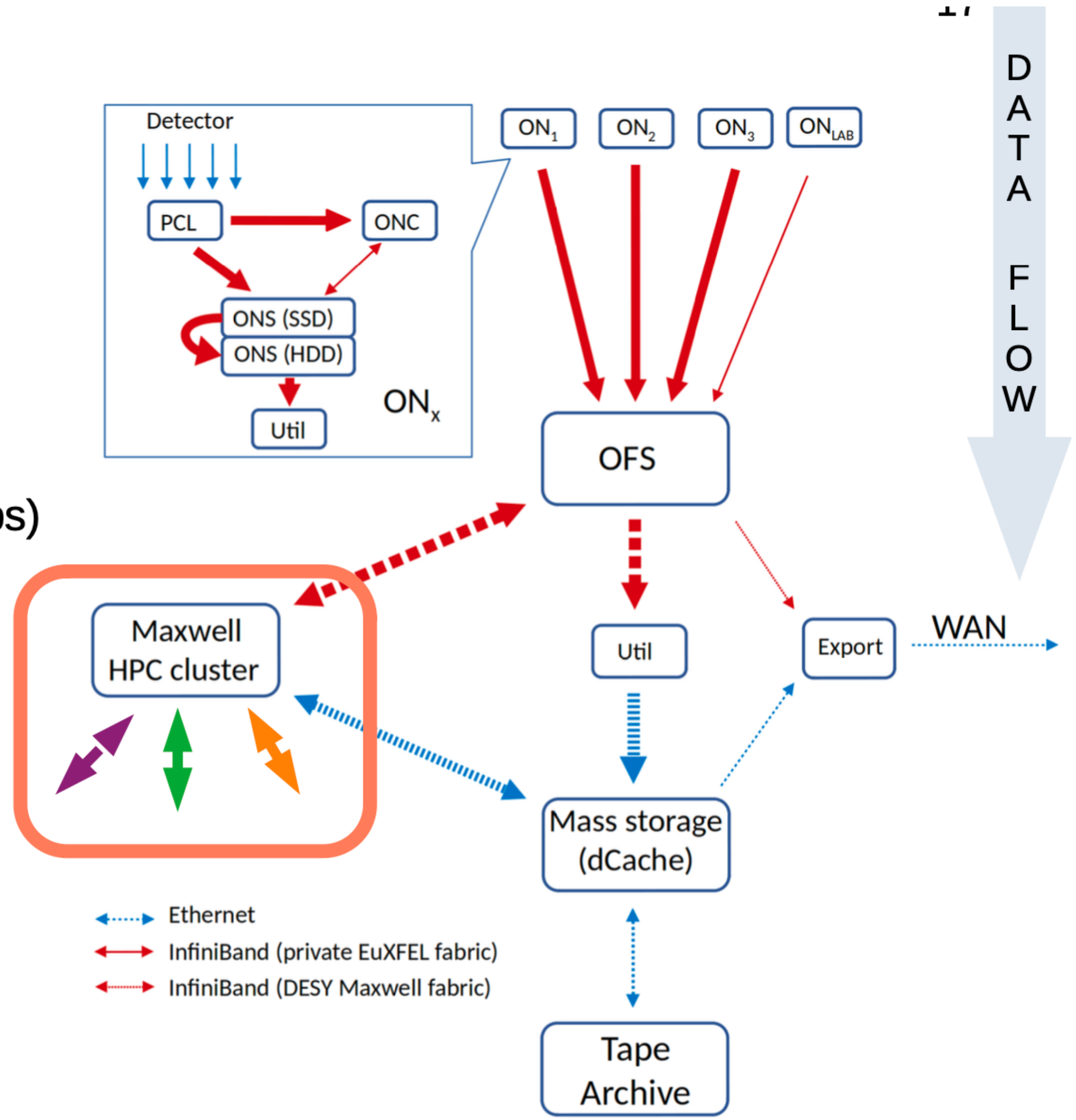
Offline infrastructure (common)

- High-performance storage - (Offline)
  - shared data storage for all SASEs
  - connects private EuXFEL IB and DESY Maxwell IB
  - data source for the offline data analysis
  - short term storage for raw data
  - storage for processed data and intermediate results
  - source for data transfer to the dCache system
  - based on **IBM Storage Scale System** (f.k.a GPFS)
  - size of **50PB** in 7 file systems
  - **175GB/s** observed read processes (during data taking)
- Maxwell – HPC computing cluster
- Mass storage
- Tape Archive
- Export services



Offline infrastructure (common)

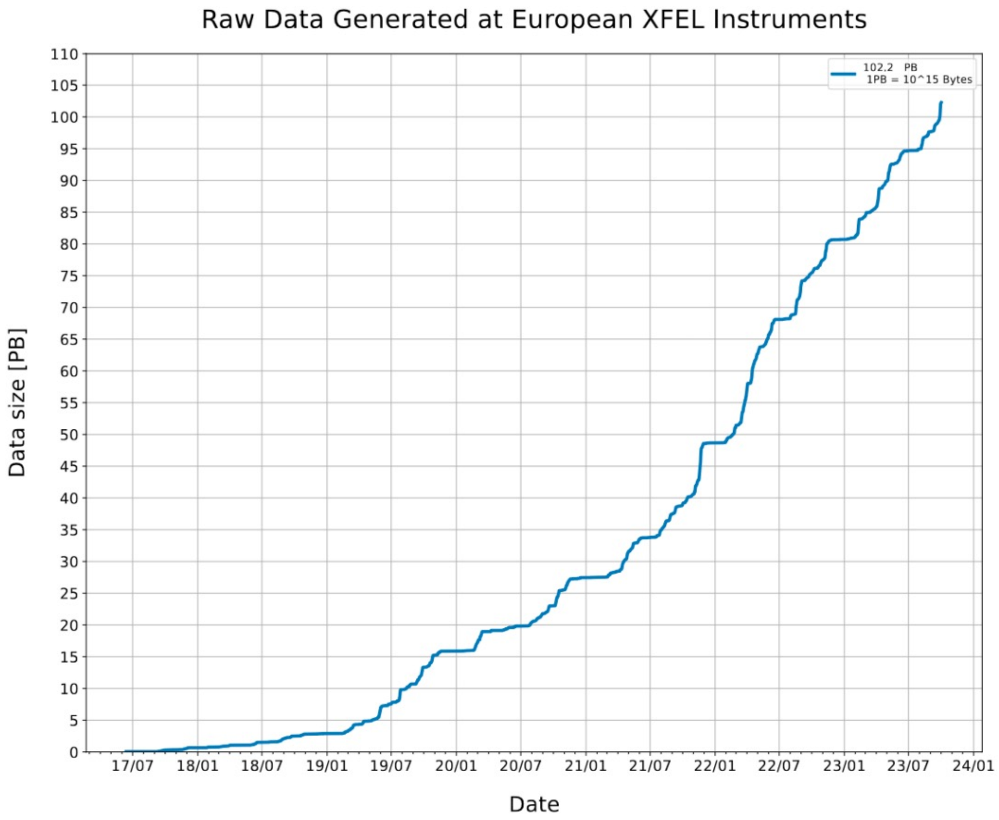
- High-performance storage - (Offline)
- Maxwell – HPC computing cluster
  - shared computing cluster among user communities
  - extremely heterogeneous in type and age
  - ~940nodes in total (4000TFlops)
  - 439 CPU, 30 GPU EuXFEL contribution (1200TFlops)
  - SLURM managed
  - post experiment data analysis and reprocessing
  - reservations for active experiments for data processing during experiment and analysis
- Mass storage
- Tape Archive
- Export services



European XFEL

Infrastructure Summary

- The infrastructure has capacity to accept and process up to 2PB of data per day (peak performance), which demonstrates the remarkable capabilities of all sub-services involved in this process.
  - 102PB of RAW data as of now
  - 50PB of processed data
- Simple approach of storing all generated data for long-term is not sustainable any longer
  - We are addressing the issue currently by modifying our scientific data policy, applying data reduction techniques and introducing data management plans
  - Looking for statistics driven data management within boundaries of data policies



Summary

* Infrastructure and Software go hand-in-hand to achieve peak storage of 2PB/day, and online streams of 4kHz mega pixel images.

* Separation between online and offline data stream
  * Online: in-memory, staged pipelines using Karabo and ZMQ (to user space). Python and C++
  * Offline: HDF5 files with most computing done on Maxwell HPC. Mainly Python.

* Data reduction will be the next challenge in terms of throughput (and is mandatory to address the storage challenge)
  * Online data reduction exceeds current stream protocol bandwidth. RDMA is a promising evolution path currently being implemented.

Expert Contacts

* Data Management via MyMDC: Luis Maia (luis.maia@xfel.eu)
* Data Infrastructure: Janusz Malka (janusz.malka@xfel.eu)
* Data Acquisition: Dennis Goeries (dennis.goeries@xfel.eu)
* Online Calibration: David Hammer (david.hammer@xfel.eu)
* Offline Calibration: Philipp Schmidt (philipp.schmidt@xfel.eu)
* Data Reduction Algorithms: Egor Sobolev (egor.sobolev@xfel.eu)
* Extradata data access library: Thomas Kluyver (thomas.kluyver@xfel.eu)
* Group Heads
  * Data Analysis (luca.gelisio@xfel.eu)
  * ITDM (krzysztof.wrona@xfel.eu)