Exploiting DMA for accelerator operation and the related data challenges

Common challenges between DMA ST1 and ST3?

Annika Eichler, MSK IPC, DESY
03.01.2025

ML for particle accelerators

ML for Accelerators

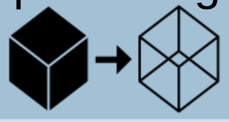
What are the most important fields?

Data analysis

- Understanding physics
- Find new correlations of parameters
- Identify relevant data channels

→ New physical insight

Estimating and predicting



Surrogate models

→ Fast models for online control and optimization, and for accelerator design

Virtual diagnostics

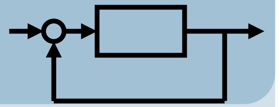
→ Additional, nondestructive, (online) information

Fault diagnosis

- Predict & prevent failures
- Protect the system
- Identify poor conditions
- Find the root cause of errors encountered

→ Improve the availability/ reliability of machine operation

Tuning and control



- Exploit data to retrieve desired machine settings
- Push the way of operation
- Optimize performance

→ Better performance for users

[1811.03172] Opportunities in Machine Learning for Particle Accelerators (arxiv.org) 2018

ML for Accelerators

What are the most important fields?

Data analysis

- Understanding physics
- Find new correlations of parameters
- Identify relevant data channels

→ New physical insight

Estimating and predicting

Surrogate models

→ Fast models for online control and optimization, and for accelerator design

Virtual diagnostics

→ Additional, nondestructive, (online) information

Fault diagnosis

- Predict & prevent failures
- Protect the system
- Identify poor conditions
- Find the root cause of errors encountered

→ Improve the availability/ reliability of machine operation

Tuning and control

- Exploit data to retrieve desired machine settings
- Push the way of operation
- Optimize performance

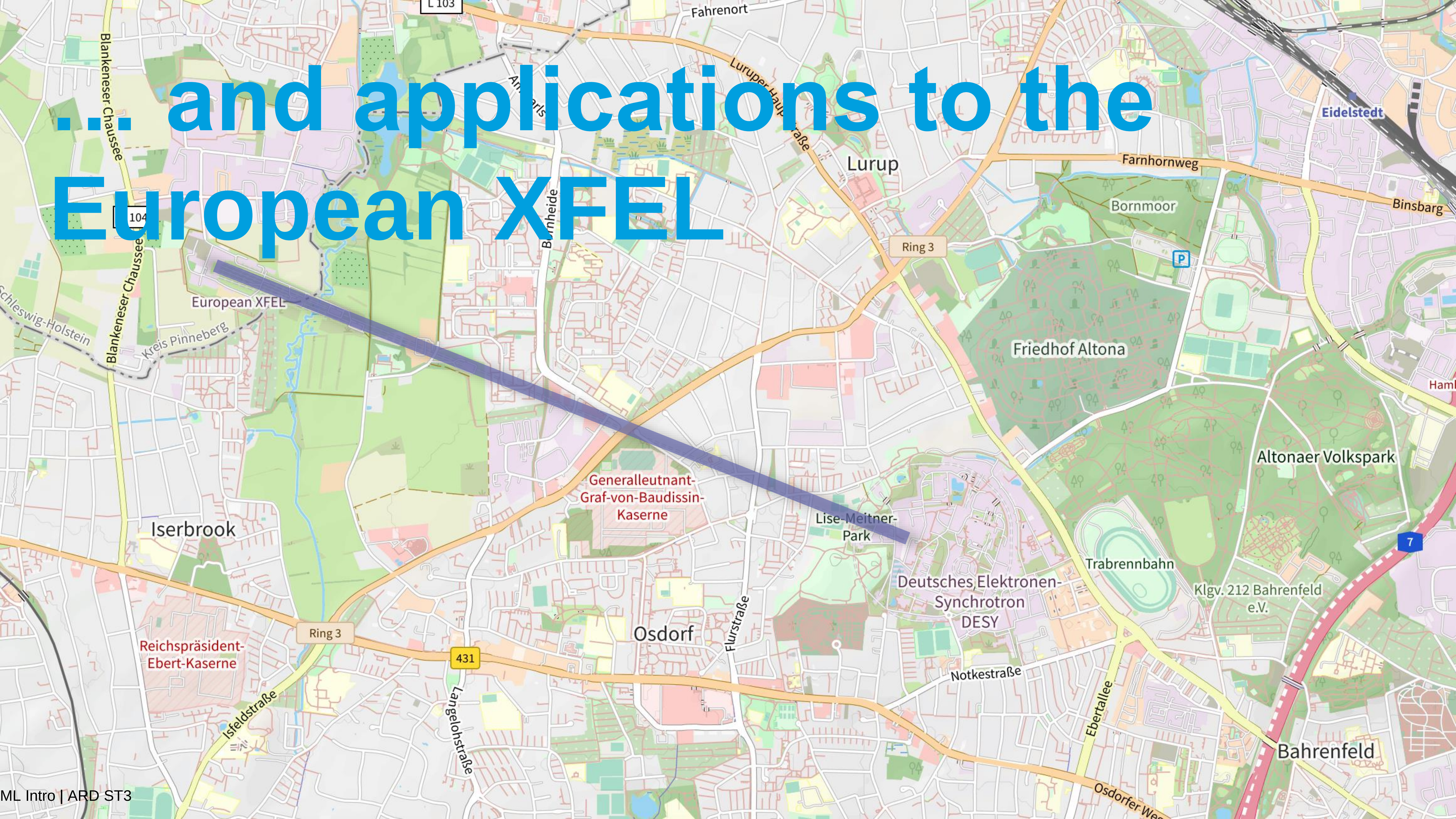
→ Better performance for users

Human Machine Interaction

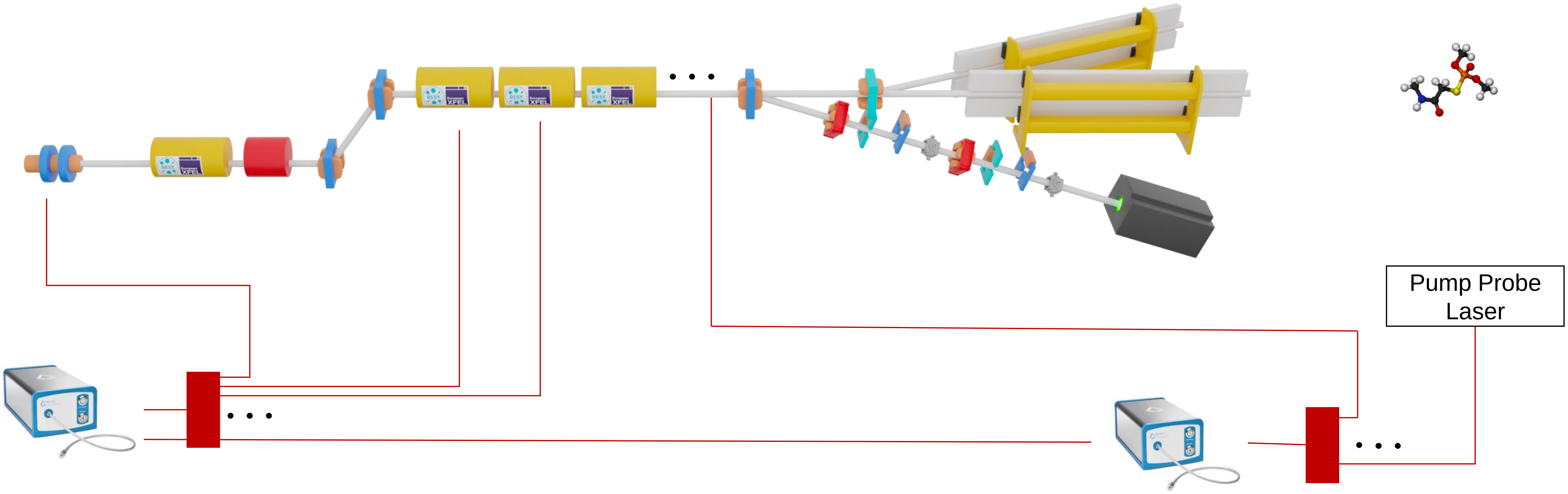
- Better represent the data: Visualization also with virtual reality
- Get information easier: Improved information retrieval from documentation and logbook
- Ability to ask: Chatbots for Q&A

→ More human understandable feedback (and action)

… and applications to the European XFEL

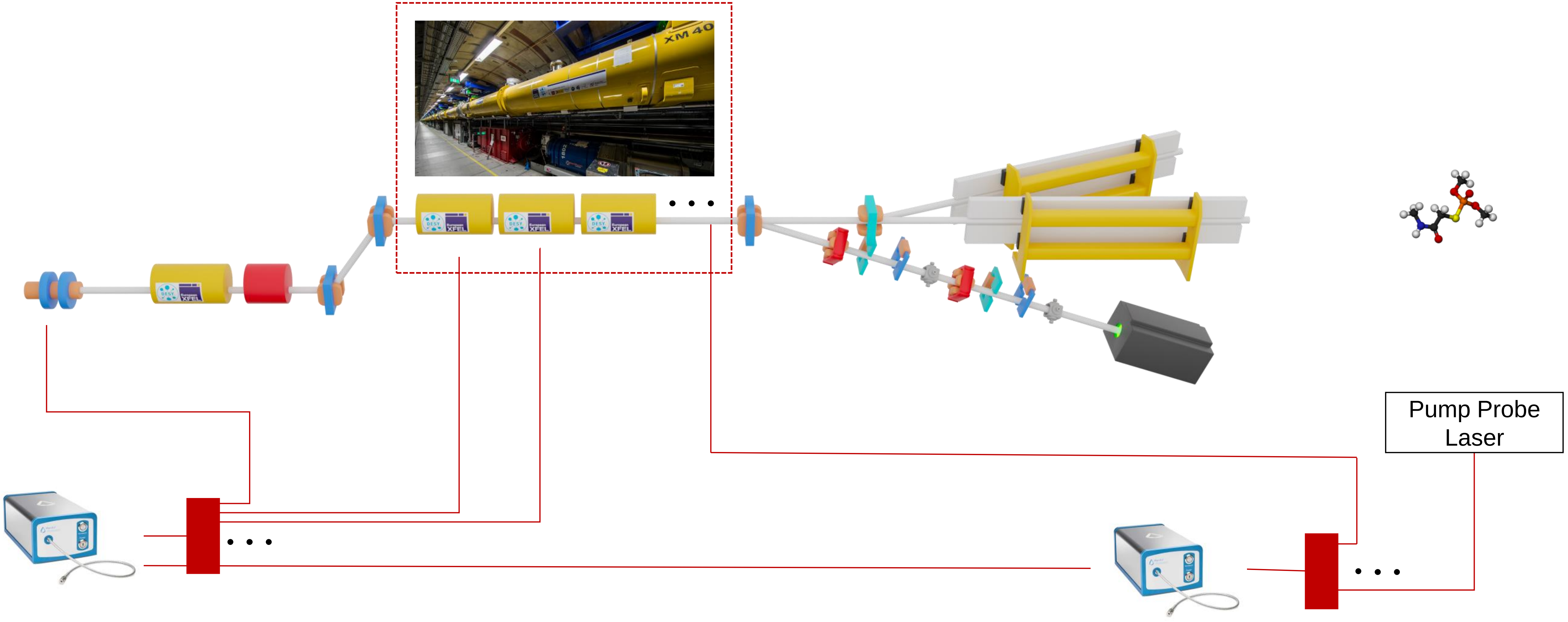


MATTER AND TECHNOLOGIES



Fault detection and classification for SRF cavities

MATTER AND TECHNOLOGIES MT



Fault detection and classification for SRF cavities

MATTER AND TECHNOLOGIES MT

Goal: Detect and classify quenches and take countermeasure

Quench: severe fault (cavity walls **lose superconductivity)**

So far:

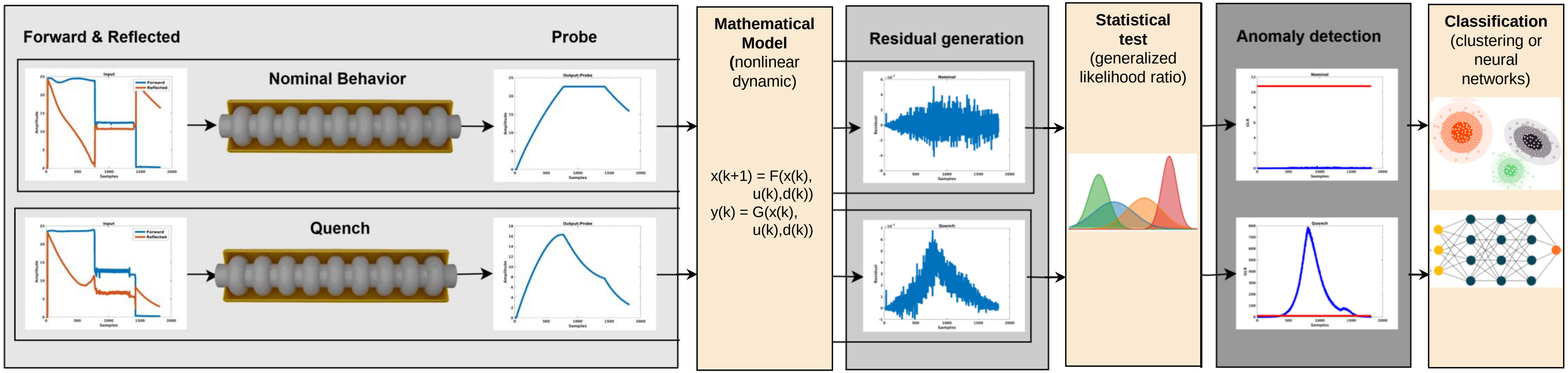
Quench detection system (in operation since 2017)

- No intra-pulse detection and compensation possible
- **20% false positive rate,** 93% true positive rate (in 2022)

New approach:

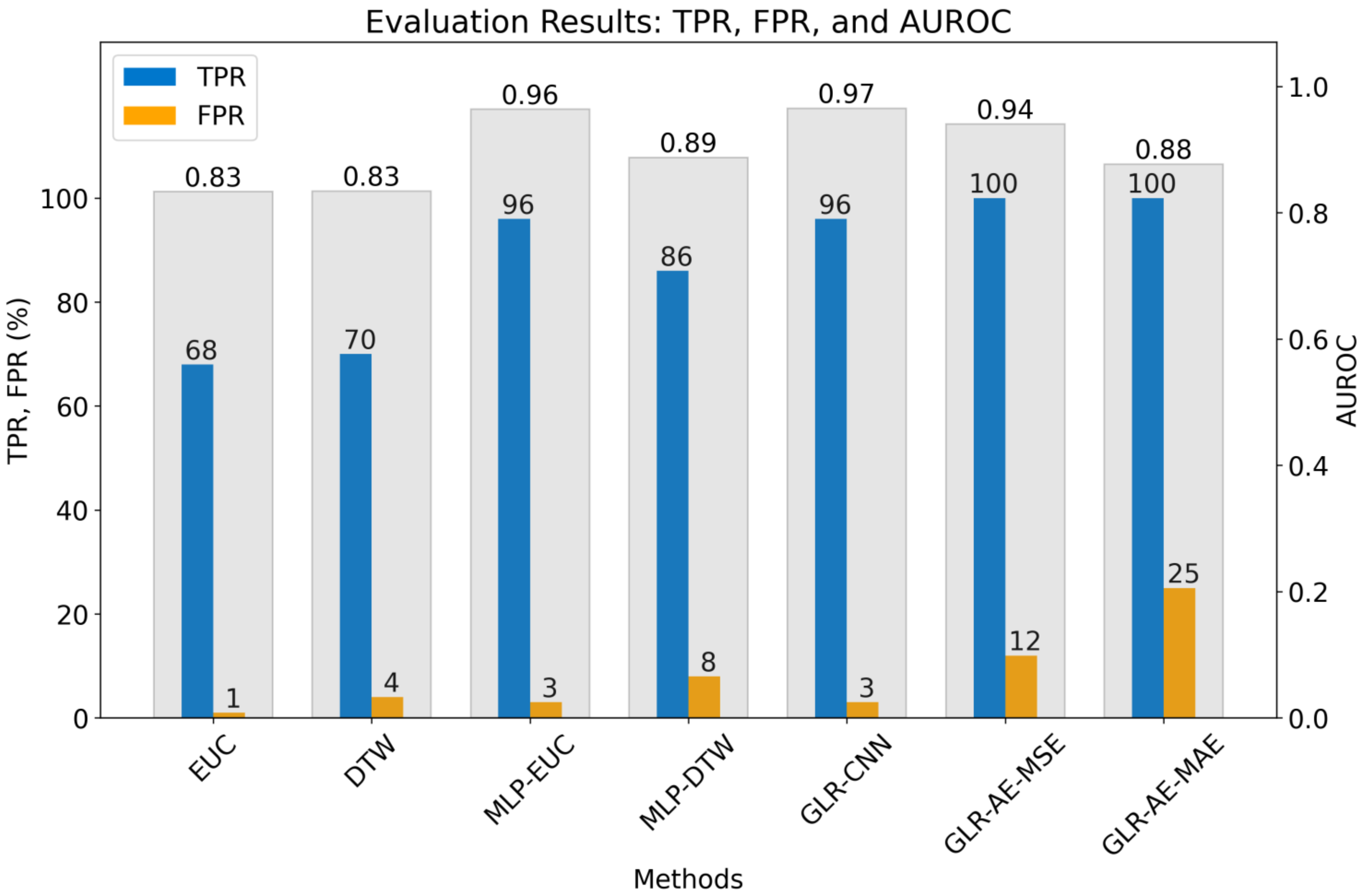
Quench detection using ML-based solution

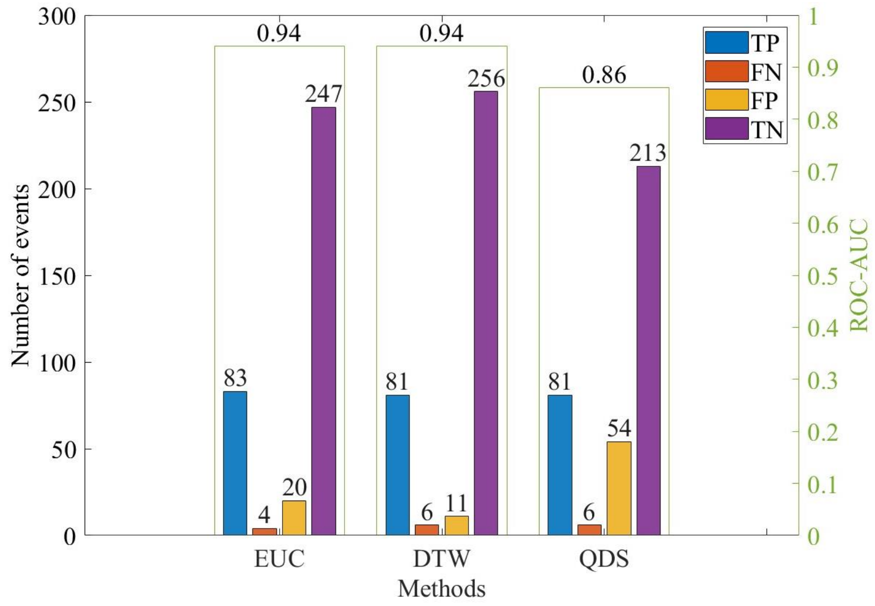
- Three-stage approach
  - **Feature generation** based on physical model (parity space)
  - **Anomaly detection** based on statistical tests
  - **Classification** based on ML



Results

MATTER AND TECHNOLOGIES





QDS : TPR = 0.93 and FPR = 0.20

EUC : TPR = 0.95 and FPR = 0.07

DTW : TPR = 0.93 and FPR = 0.04

A Eichler, J Branlard, JHK Timm. **Anomaly detection at the European X-ray Free Electron Laser using a parity-space-based method**. In *Physical Review Accelerators and Beams* 26 (1), 2023.

L Boukela, A Eichler, J Branlard, NZ Jomhari. **A Two-Stage Machine Learning-Aided Approach for Quench Identification at the European XFEL**. In *IFAC-PapersOnLine*, 58(4), 2024.

Deployment & Evaluation

Data labeling, online deployment

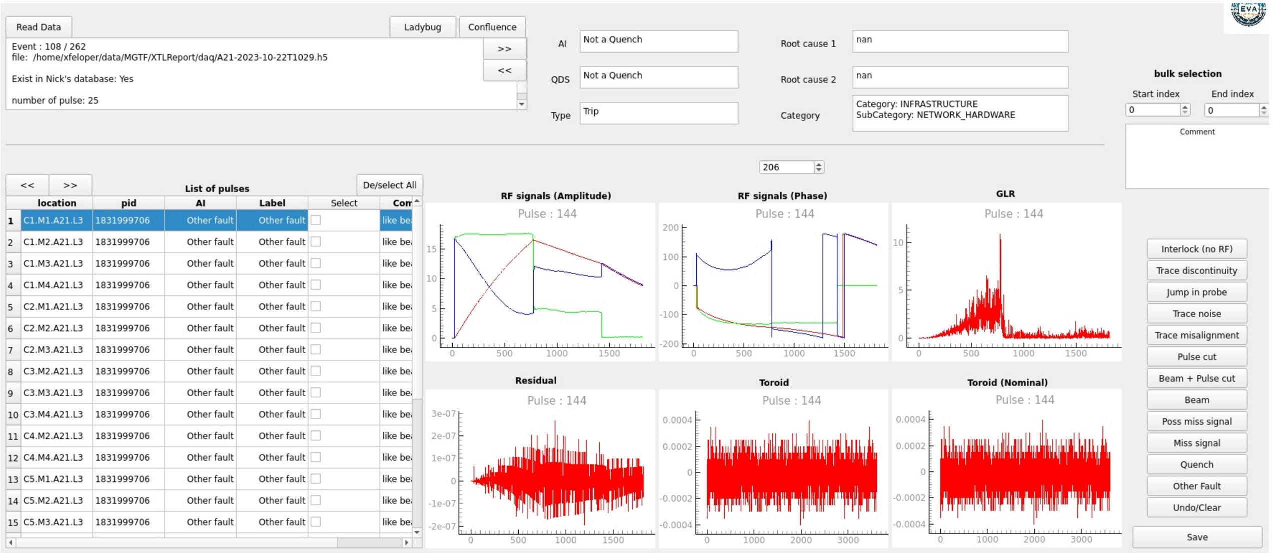
Offline: support by daily e-mails

- **Go online!**
  - **Software:** Implementation in C++ to run live on one station
  - **Firmware:** Implementation has been deployed in winter shutdown 2024/2025

Location	PID	Timestamp	Type of anomaly	maxGradient
C1.M4.A8.L3	2138819411	11-Oct-2024 12:38:27	possible quench (DTW)	18.68
C1.M4.A8.L3	2138819412	11-Oct-2024 12:38:27	quench	18.81
C4.M4.A8.L3	2138819410	11-Oct-2024 12:38:27	quench	23.01
C4.M4.A8.L3	2138819411	11-Oct-2024 12:38:27	quench	19.5
C4.M4.A8.L3	2138819412	11-Oct-2024 12:38:27	possible quench (DTW)	13.64

Human in-the-loop approach

- Data labeling for other types of faults with help from the LLRF experts
- GUI was developed to ease the labelling



Data issues

Data within one pulse:

- 6 signals per cavity, 1MHz sample rate, ~2ms pulse length, 10Hz pulse frequency, 800 cavities

  → ~96.000.000 samples per second
- Collection of all trips (unlabeled) since 2019 exists (snapshot files)
- Analysis results of Interlock system in data base
- **DAQ system** (2 weeks of data buffer)

Implemented solution: C++

- Bandwidth → not possible to get data out of the tunnel
- Implementation in C++ to run live on one station
- 2 Servers in the tunnel (do not touch running system), radiation → not possible to get data to the separate server
- Does not scale well!
- Local storage to store the flagged anomalies in ringbuffer

Firmware

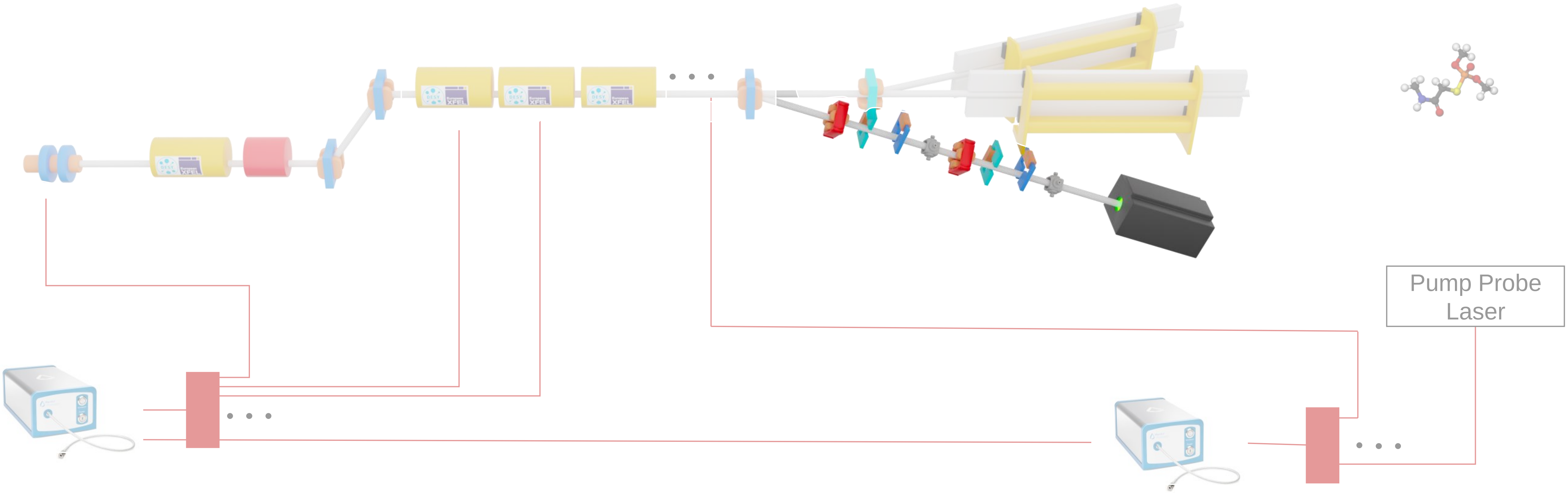
- Analyzing data directly only possible with experts

General

- Missing beam data in snapshots (at least of the one which would be online available)

Autonomous accelerator tuning

Reinforcement learning: From ARES Sinbad to the European XFEL

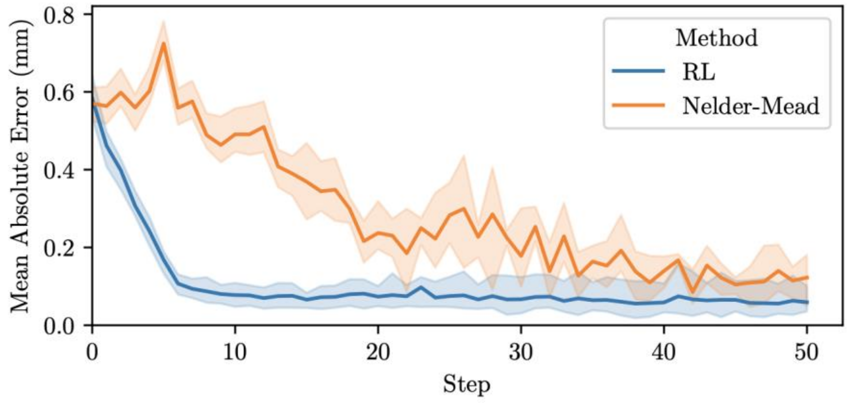


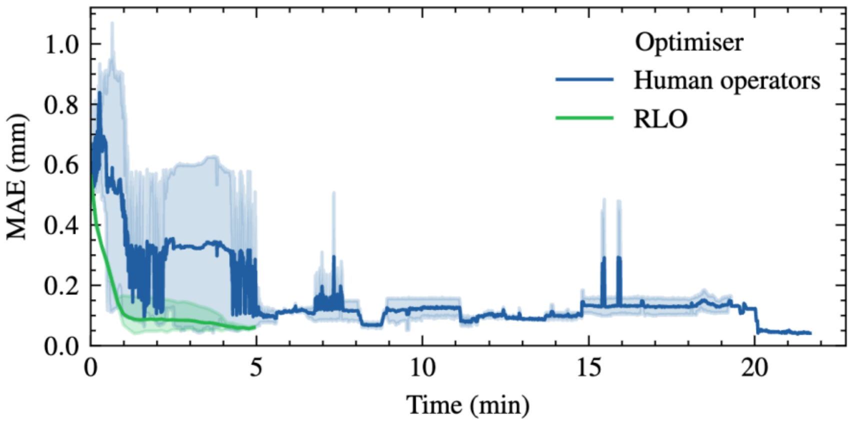
Autonomous accelerator tuning

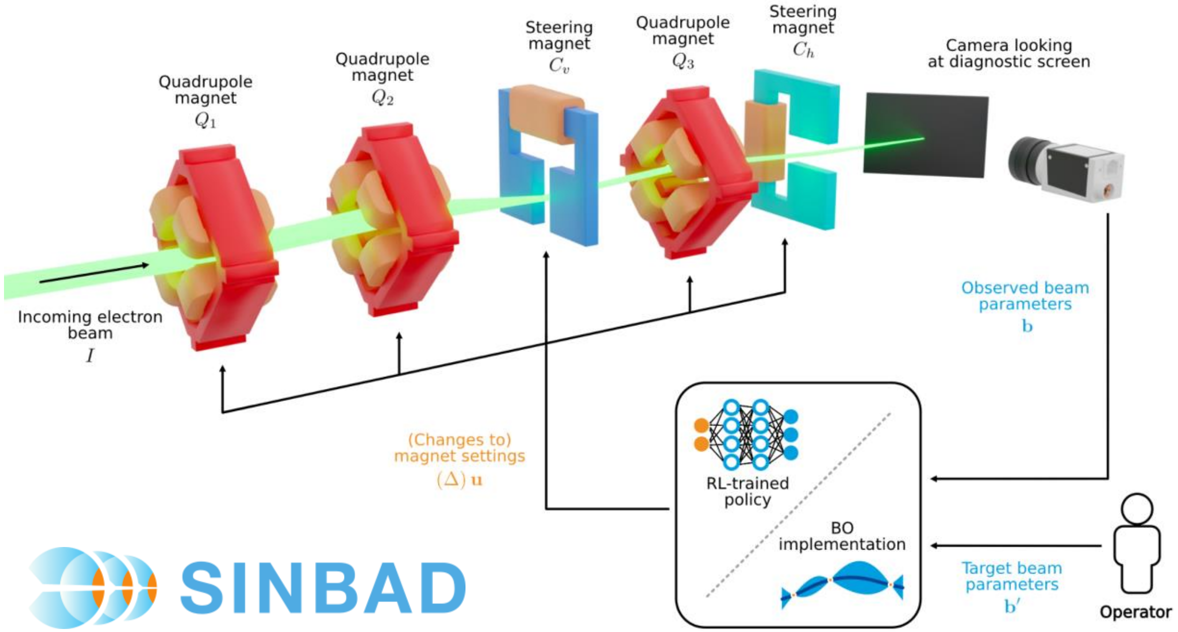
Reinforcement learning: From ARES Sinbad to the European XFEL

Reinforcement learning-trained optimization at ARES

- Deploy a RL-trained optimization algorithm trained purely in simulation to the **real-world** with **zero-shot learning** thanks to **domain randomization.**
- The trained policy **outperforms other optimization algorithms and expert human operators**.







J Kaiser, O Stein, A Eichler. **Learning-based Optimisation of Particle Accelerators Under Partial Observability Without Real-World Training**. In *International Conference on Machine Learning*, 2022.
J Kaiser, C Xu, A Eichler, et. al. **Reinforcement learning-trained optimisers and Bayesian optimisation for online particle accelerator tuning**. In *Scientific reports* 14 (1), 2024

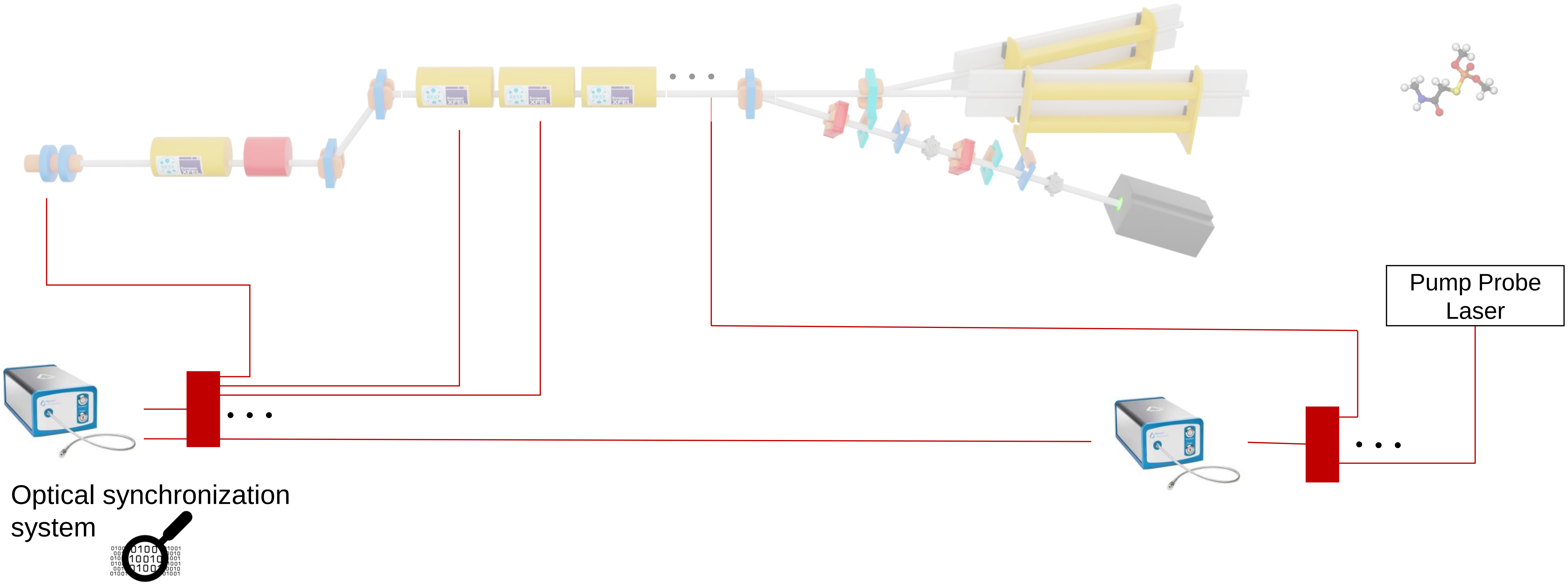
Autonomous accelerator tuning

Data issues

- Shift data often messy, manually set up, but no time for cleaning
- Backups necessary
- Mid-term archive … sometimes its not clear whether data needs to be archived longterm … back-upped storage still needed, but tape backups too slow
- Moving data around
  - Can be quite slow from some machines to others
  - Permissions … functional accounts for data archiving, functional accounts for operations, user accounts
  - Long-term archiving is necessary

The optical synchronization system

MATTER AND TECHNOLOGIES



DAQ

Current Situation at DESY / EuXFEL

Large-scale accelerators provide huge amounts of data

- And it’s getting more

> 10 Million data addresses in DOOCS control system for EuXFEL

- Configuration, measurements, extracted features..

> 20.000 high data-rate channels

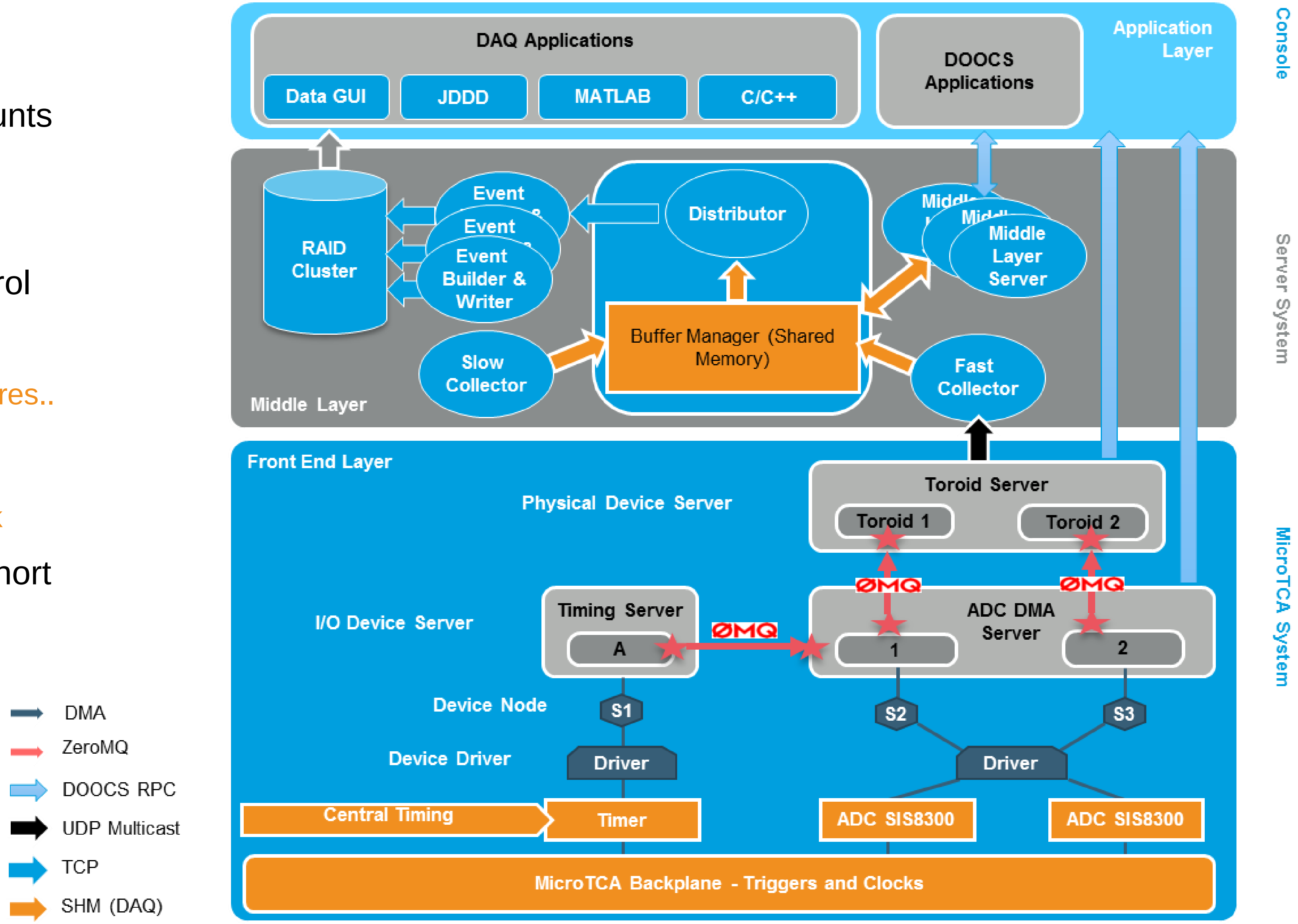
- Not feasible to completely transfer via network

30 TB/day of data recorded for EuXFEL in short term archive

- In many cases not evaluated before deletion

< 1 % of available data goes to central DAQ

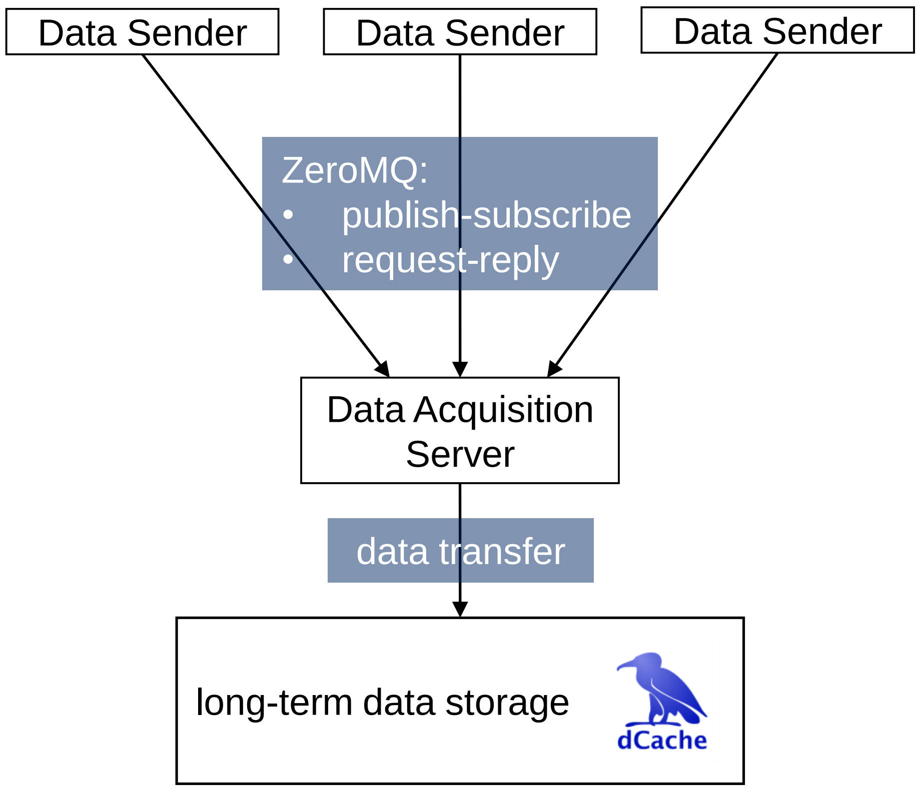
- Often channels required for certain conclusion are missing



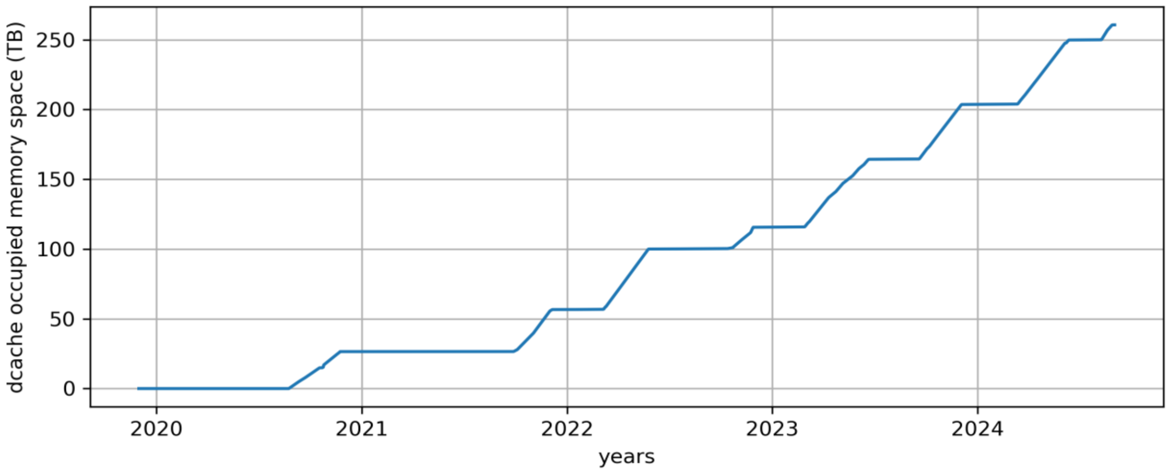
Courtesy: T. Wilksen

Making data available of the optical synchronization system

Building up a DAQ system



- Data sources **~41k control system channels**
  - Controller I/O of all feedback systems
  - Configuration
  - Environment (*T*, relative humidity, air pressure)
- dCache volume **~250 TB** since 2021
  - 10 Hz acquisition rate
  - Daily 10-second long snapshots of "fast" data
- 5-day ring buffer
  - Fast data (up to 300 MHz) of select subsystems, e.g. MLO, SLO, LSUs



A Grünhagen, M Schütte, A Eichler, M Tropmann-Frick, G Fey. **Enhancing Data Acquisition and Fault Analysis for Large-Scale Facilities: A Case Study on the Laser-Based Synchronization System at the European X-Ray Free-Electron Laser**. In *LWDA*, 2023

DAQ & Data Collection Challenges for LbSync

2 PhD students spent for this

DAQ Setup problem

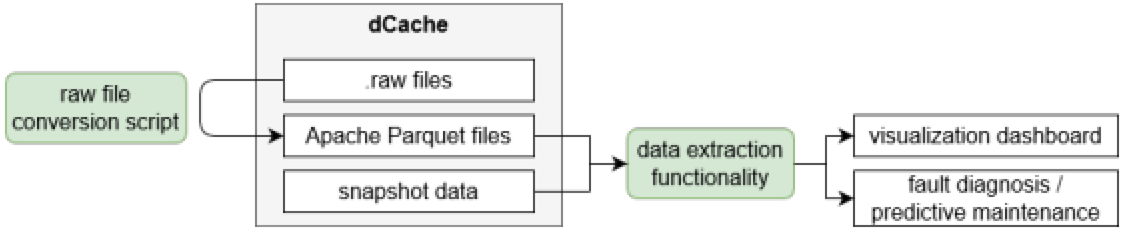
- EuXFEL DAQ configuration is “stiff” & managed by experts.
  - Needs to be updated regularly, or new data channels might be missed.
  - Flexible & quick-to-deploy python tools help to be “agile” and collect data on the spot. Useful for smaller campaigns.
- Not all data channels / types are compatible with DAQ.
- Collection & storage are inefficient in terms of network bandwidth and disk space.
- Different users have different requirements.
  - Easy to propose a solution that fits you, but doesn’t work for others.

Data and data readout issues

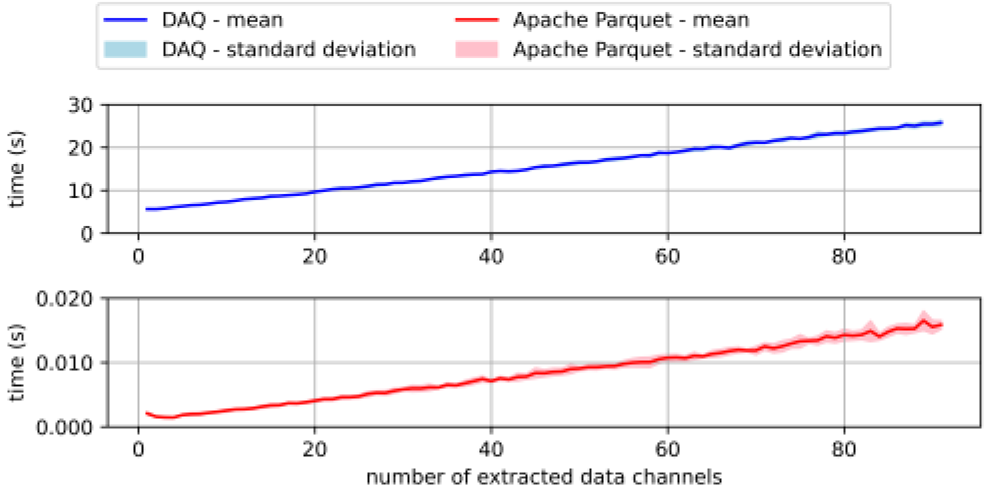
- Data is produced in non-standard format, conversion not straight forward. Not F.A.I.R.
  → Readout
- Comparing data from different times is hard, not all work on the accelerator is documented.
  → Metadata
- Even documented work is difficult to include in analysis. Metadata needs more structure or ML-based evaluation.
  → Metadata

LbSync DAQ Readout Problems

- **Current Setup**
- MCS DAQ System provides time series data
- Data is stored in a DESY in-house .raw format
- Challenges with .raw format:
  - Not F.A.I.R. compliant
  - Slow data readout (5 KB/s) → Data generation faster than retrieval
  - Automatically stored on dCache

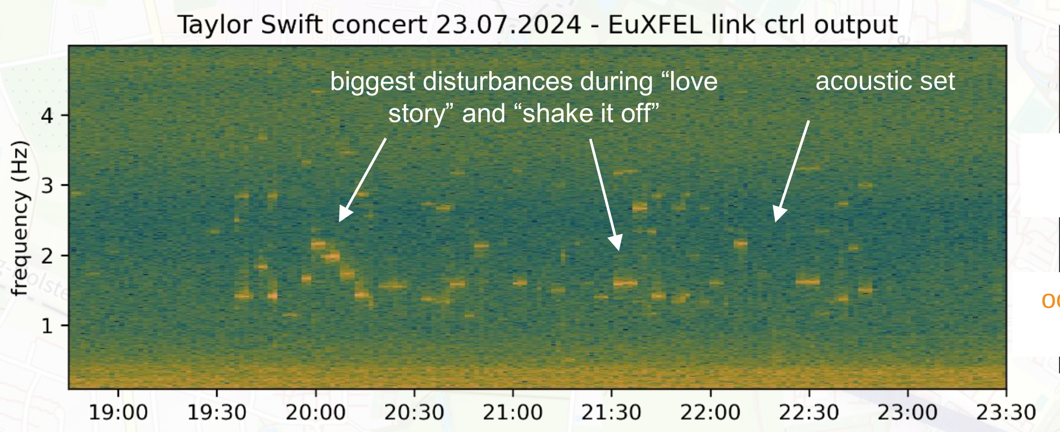


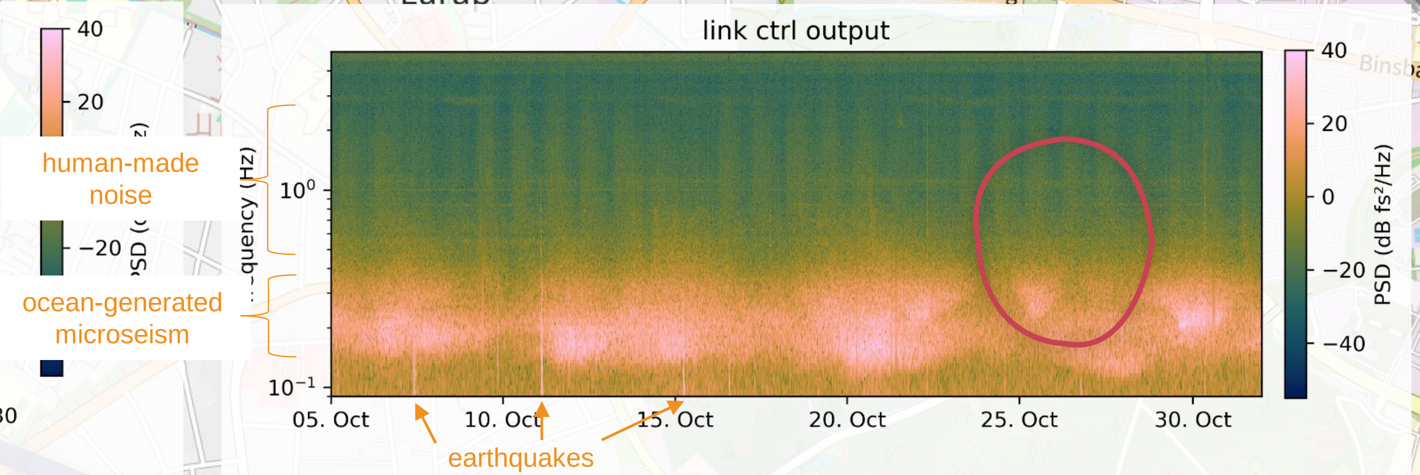
- **Current Improvements**
- Online conversion script developed
  - Based on pydaq (MCS)
  - Converts .raw to parquet format
  - Only a limited set of available data is converted
  - Not scalable for full data conversion
- Faster data readout 220 MB/s using parquet



Disturbance analysis

Of fiber link stabilization unit





- Seismic activities change the tunnel length
  - affects synchronisation **fibre length**
- Sources of disturbance
  - ocean-generated microseism
  - earthquakes
  - human-made → traffic, civil constructions, concerts


Spiegel, Foto: Gregor Fischer / TAS RM / Getty Images

The Metadata Problem

A DAQ only gets you so far

What is Metadata

- Describes the data you have collected
  - Which data is collected, for which time, at what rate, in which format, using which software…
- But also information that tells you how to interpret data.
  - SNR of an ADC, coefficients of passive components used in sensors, Device Serial Numbers, Experimental Setups in general…
  - Everything that you need to make sense of data or that influences what you are measuring.

Why is Metadata important?

- For data analytics and machine learning, it is nearly impossible to adequately work with data, that changes it’s meaning at an unknown point in the time series.
- Maintenance activities cannot be distinguished from faults / crosstalk.
- Results can be catastrophically falsified if data safety is not guaranteed.

LbSync DAQ – Potential Solutions

- dCache doesn’t like when data is continuously overwritten. WORM (Write Once, Read Multiple Times)
- **Challenges & Potential Solutions**
- InfluxDB → Optimized for time-series, but scaling to >400 TB challenging
  - InfluxDB is optimized for dynamic, frequently updated time-series data.
  - WORM storage means no modifications or deletions are possible, which conflicts with InfluxDB’s architecture, as it often overwrites or aggregates data points.
  - **Solution**: InfluxDB could be used as a short-term buffer (e.g., for live data) before storing data in Parquet/dCache.
- Apache Spark → Distributed processing for parallel data access
  - Spark mainly works with batch processing on existing data and does not require write operations on storage
  - It can directly access dCache if the file format is well-structured
  - This is not the case for .raw files, but it works well with Parquet
- Converting raw data into usable data is still the bottleneck. To make data usable, the DAQ must be optimized for databases

Other (long-term) (multi-channel) analysis

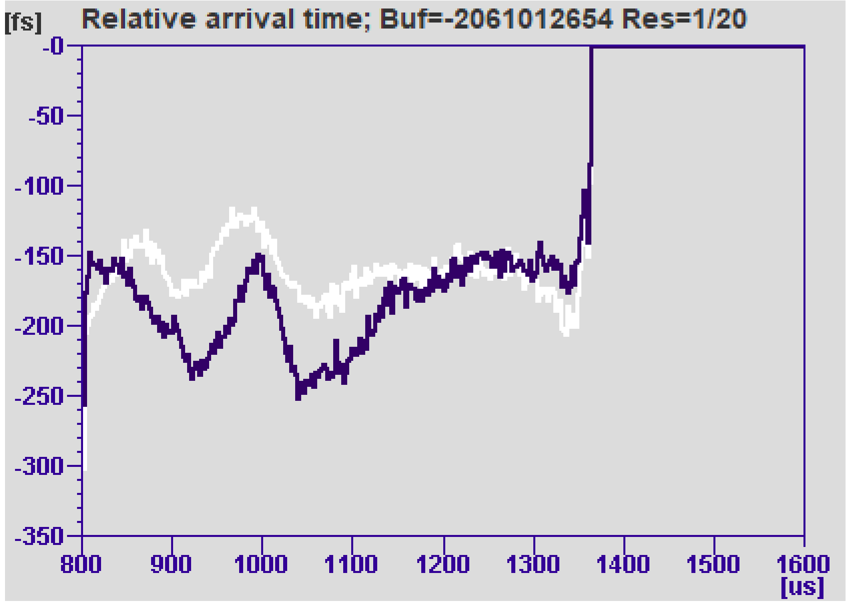
Data challenges

Data reduction and subsampling

Efficient data reduction strategies are needed to reduce storage requirements as well as to store longer-term data.

Example longitudinal bunch feedback :

- The accelerator returns data for each bunch
  - BAM, BCM, accelerating modules
- A single measurements per macropulse would be sufficient for modeling purposes
- Either filter in middlelayer or incur large storage and transmission overhead
  - Currently custom script for data collection for filtering of the bunches



Other (long-term) (multi-channel) analysis

Data challenges

Timing

Timestamps or macropulse ID mismatching makes it difficult to correlate machine parameters with experimental results.

- For example the mismatched mpid of RF pulses may lead to incorrect evaluation of the detuning and bandwidth of the cavity, and may also lead to incorrect anomaly detection.

Front end histories

- Each subsystem has different filter settings, different update rates, and data quality
- Some components have very limited historical data (such as RF pulses) or even no historical tracking data for data analysis or debugging.
- There is a large amount of data and metadata missing from historical data, making it difficult to analyze, especially long-term data analysis. In addition, historical data are noisy.

Thank you

And thanks to the MSK IPC group

Contact

DESY. Deutsches
Elektronen-Synchrotron

www.desy.de

Annika Eichler
MSK
annika.eichler@desy.de
+49 (0)40 8998 4041

TUHH
Hamburg University
of Technology
www.tuhh.de

Annika Eichler
ICS
annika.eichler@tuhh.de