LEAPS-Innov WP7.2 (M7-M18)

“Assessment of **future needs** and **development of metrics** for data compression and reduction”

Task coordinators:

- **Nicolas Soler** (ALBA)
- **Vincent Favre-Nicolin** (ESRF)

“Information on site-specific scientific needs of experiments at LEAPS facilities will be gathered through **interviews with domain experts and scientists** taking into account the past and future needs: **bandwidth, latency, connectivity, available storage volume and compute infrastructure**.

Downstream processing and analysis needs will be identified to serve as **upper boundary limits for information loss during compression**.

Detector manufacturers will be involved in the discussion to ensure the compatibility, and future integration in production, of the resulting components.”

“Research will also be conducted to **identify the techniques which generate the most data** and to develop metrics which assess the effect of data reduction on the quality of the final result. The most likely candidates are **serial crystallography and single-particle imaging**, multiple types of **tomography**, and most **high-resolution imaging** techniques.

Metrics will be developed to evaluate the data explosion problem and to assess the **impact of the data reduction and compression solutions** developed in task 7.3. The performance of the various reduction strategies will be investigated, with the goal of **achieving close to real-time performance** in data reduction.

This task will be carried out in consultation with **scientists and external industrial experts** in data reduction. Contacts have been already established with industrial experts in data compression and specifically the Blosc community, IBM and StreamHPC”

LEAPS-Innov WP7.2: main tasks

1) **Site-specific needs** (volume, bandwidth): already asked in email, we'll see the partners answers. A more formal/detailed poll later on ?

2) "**Upper boundary limits for information loss** during compression": **lossy compression/data reduction**. Need a specific survey among facilities for that. Or better, a poll per community (e.g. serial Xtallography, powder diffraction, tomography, etc..)

3) **Metrics**:

    a) global, technique-specific (volume reduction and compression + decompression speed)

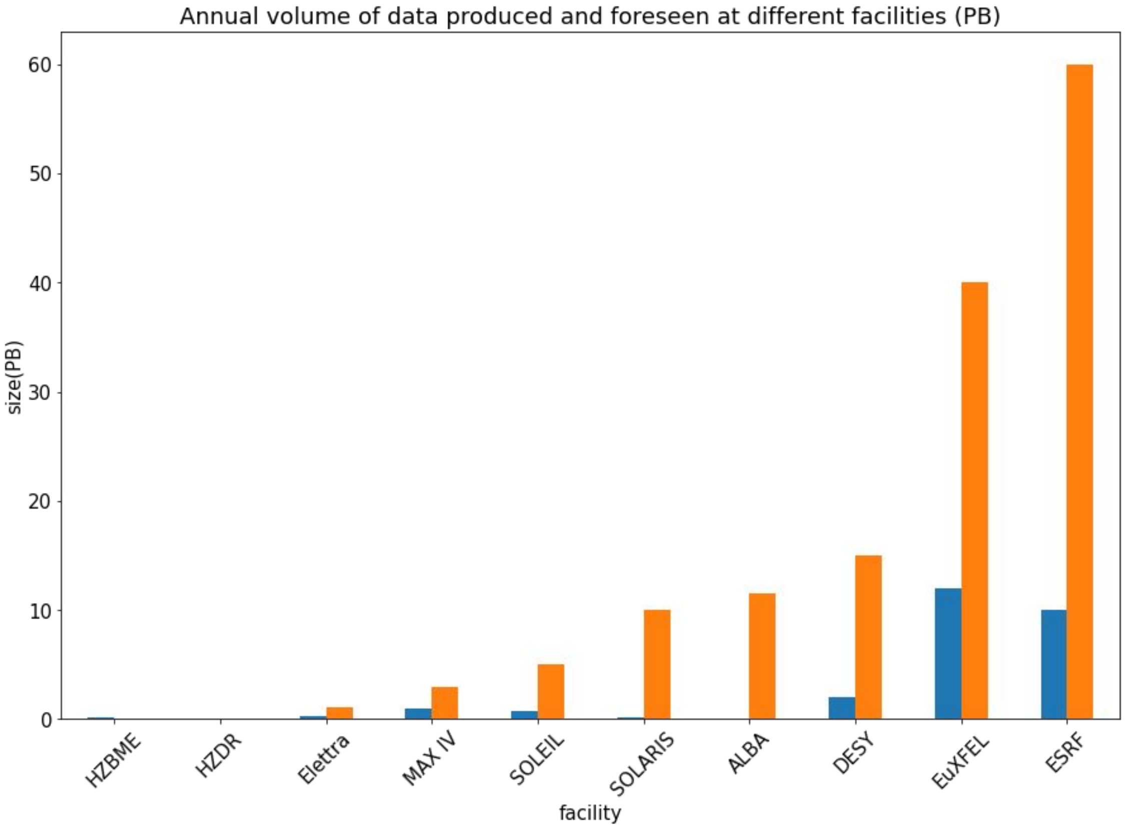
    b) loss of information (acceptability per technique/community)

Need from each partner

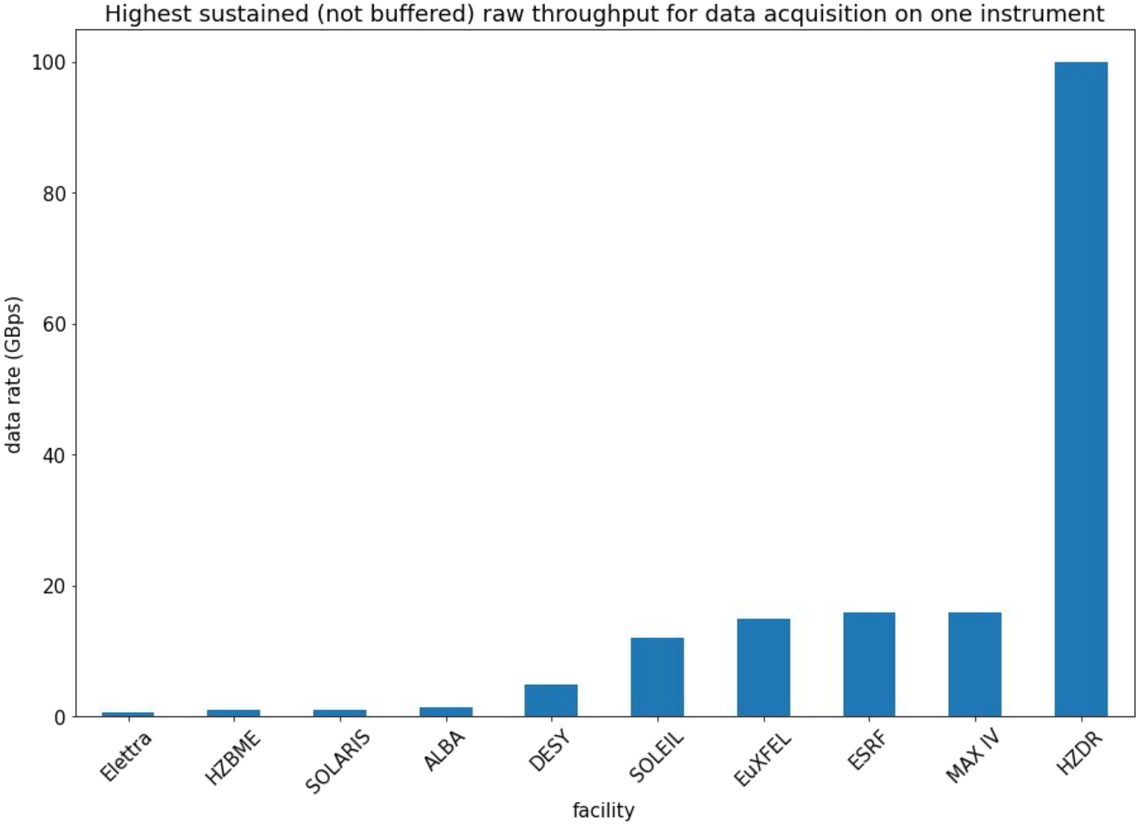
- Poll on **global needs**
- **What techniques** each partner wants to spend time on during WP7. Both development/datasets but also if they can get the scientific community involved for tests. Might be easier if partners would specialise on **1 or 2 techniques**.
- What **compression scheme** (lossy/not lossy) are already used/developed, or planned
- **Gather names/contacts** per technique for **a)** development and **b)** scientists. Once the ‘target’ techniques have actually been chosen ??

Poll results:
Techniques and compression needs

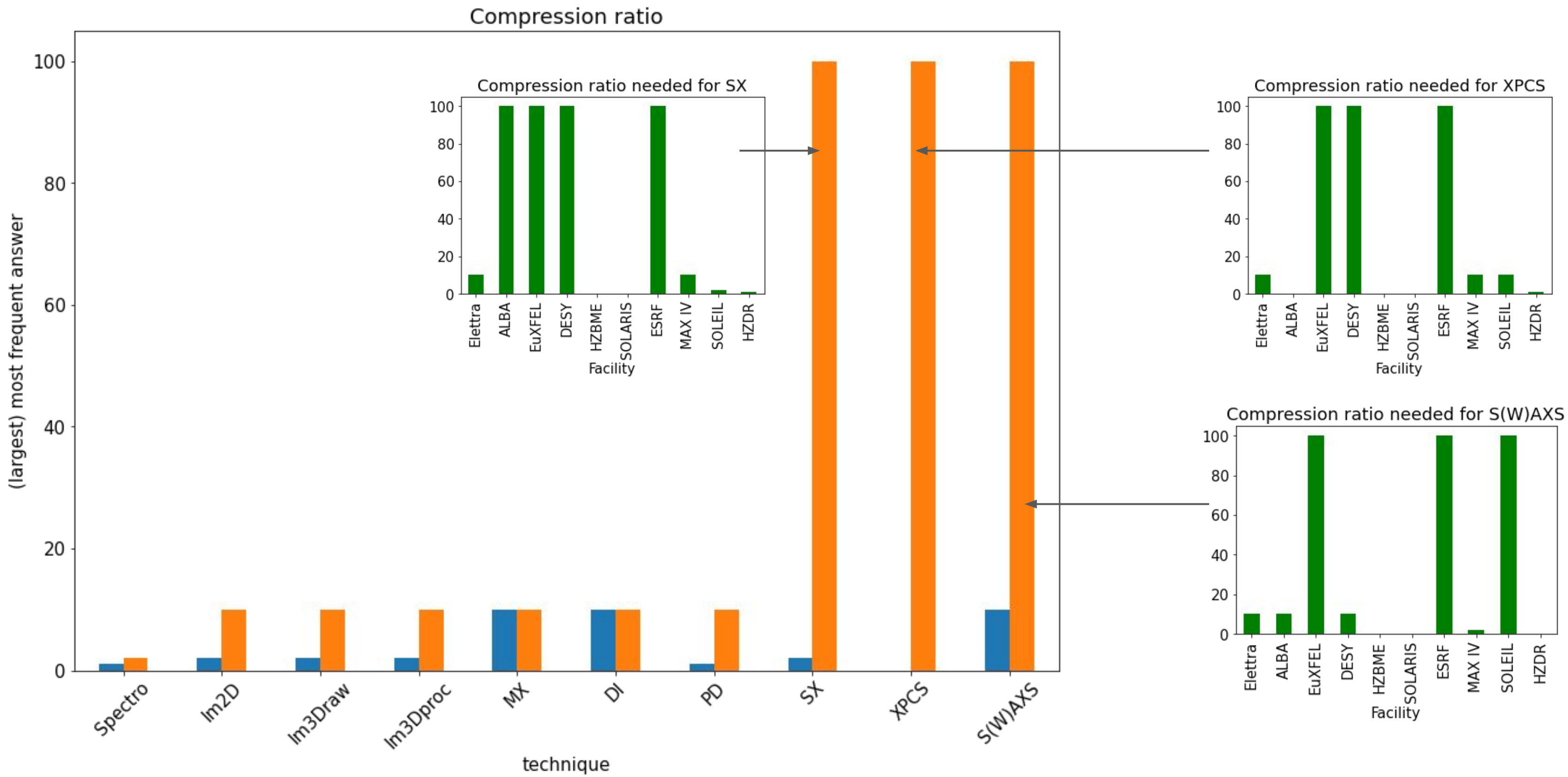
Prevision for data volume increase: 2020 - 2025



Highest sustained data rate (not buffered)



Compression ratio by technique, *achieved* vs *needed*



Compression ratio

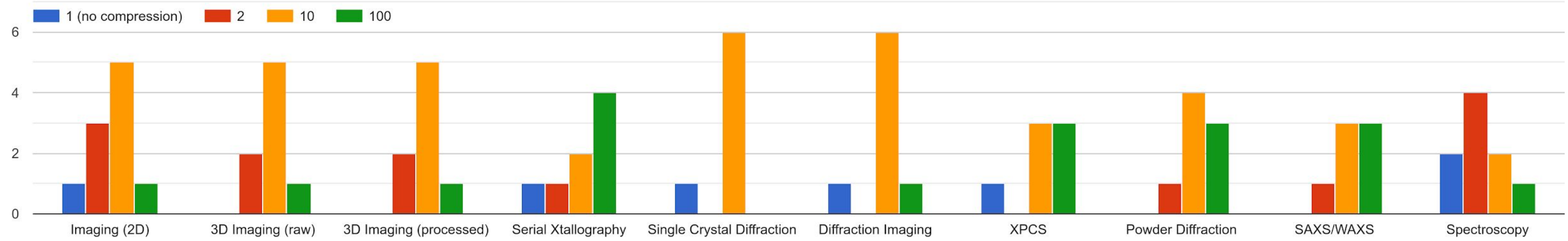
Compression ratio needed for SX

Compression ratio needed for XPCS

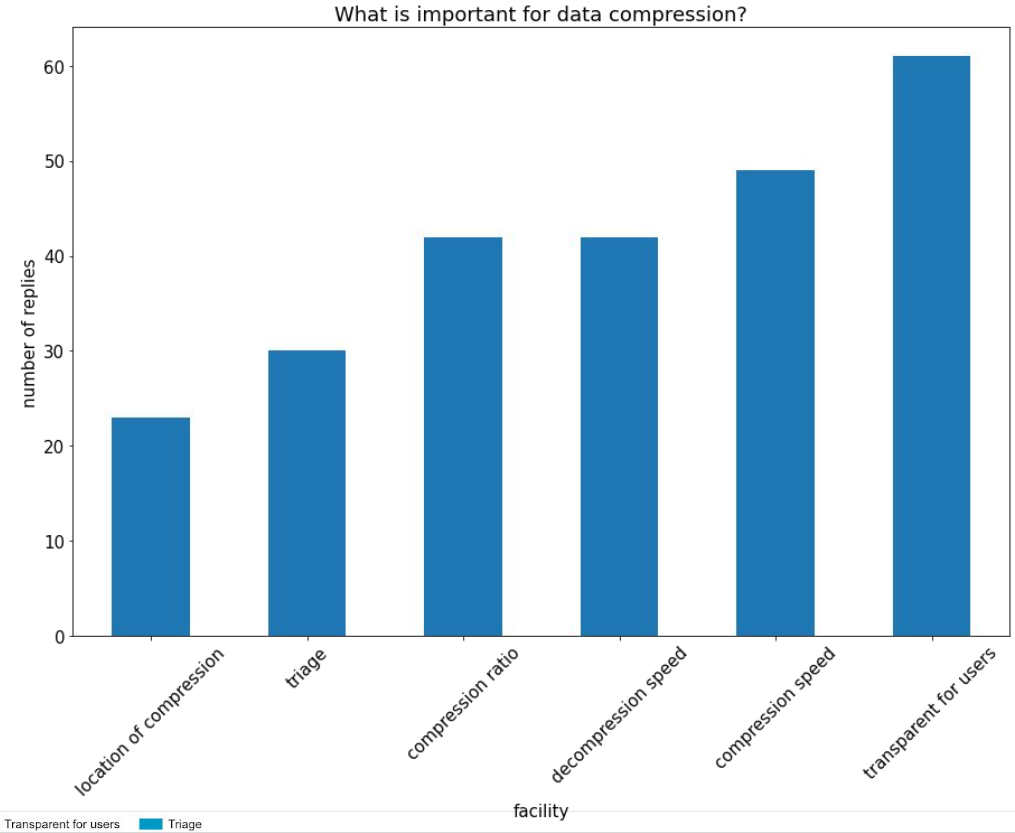
Compression ratio needed for S(W)AXS

Compression ratio by technique

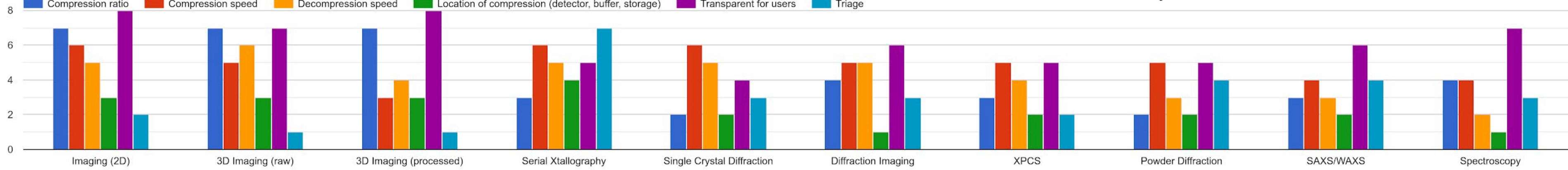
For the following techniques, what is the overall compression ratio (raw volume/archived volume) needed ? (overall=including compression, reduction and triage of data)



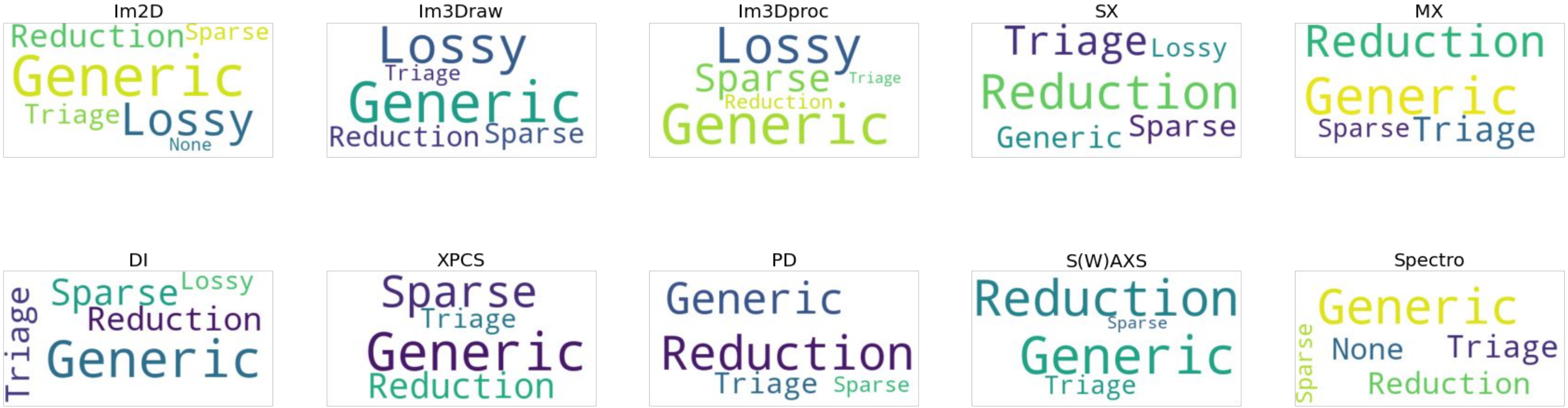
Concerning data reduction / compression, what is really important?



Regarding compression, what is really important (multiple answers possible)



Methods needed, per technique



Other techniques mentioned

Facility	Technique
Elettra	XRF, Ptychography, CDI, and FEL based techniques for our FEL FERMI
ALBA	cryo/material EM
EuXFEL	Single particle imaging (SPI), X-ray cross-correlation analysis (XCCA), Scanning transmission X-ray microscopy (STXM)
HZBME	SEM / FIB tomography
ESRF	Note for powder/SAXS/WAXS: the switch between 10x and 100x compression corresponds to the choice of keeping the original images, or not. The reduction software is already in production.
MAX IV	Event based data
SOLEIL	The spectroscopy item gathers ARPES, absorption, and elastic and inelastic/resonant spectroscopy, which may have different requirements

NB : For the last 4 techniques listed in the first table above (ie. XPCS, powder diffraction, SAXS/WAXS, spectroscopy), the answer for the overall compression ratio needed is between two proposed values : the number indicated is the high value, only one choice being possible

Existing development efforts in your facility or collaborators

Which compression or data reduction software or library are you developing?

EuXFEL Tools for marking/filtering dark frames in EuXFEL files, Python, Linux detector corrections, reduction & compression, Pyhon/C, Linux

ESRF Blosc (C) https://github.com/Blosc/c-blosc
hdf5plugin (python): https://github.com/silx-kit/hdf5plugin
Exploiting hardware compression (power9)
pyFAI (2D->1D data reduction for SAXS, powder diffraction..) https://github.com/silx-kit/pyFAI
dynamix (XPCS compression): https://github.com/silx-kit/dynamix (in development)
powergzip (hardware gzip compression) https://github.com/libnxz/power-gzip

MAX IV bitshuffle conda repackaging (python, C++): https://github.com/weninc/bitshuffle
azint - python version of MATFRAIA (C++, python): https://weninc.github.io/azint/
fpga azint (OpenCL, HLS, C++, python) (Intel, Xilinx):
https://www.esrf.fr/files/live/sites/www/files/events/conferences/2021/IFDEPS%20Virtual%20Thursdays%202021/2nd%20Thursday/5.1_Zdenek%20Matej_FlashTalk_pres_maxiv_zdenek.pdf

SOLEIL No development. Multiple parallel LZ4, ZSTD calls are used from /dev/shm for high throughput.
We rely on existing algorithms and prefer to use them as system calls/commands from within /dev/shm. This allows to use existing reduction/analysis software without modification of their source code, with maximum disk IO performance. In short, we recommend to copy data on the shared memory RAMdisk, compress/decompress there, and possibly treat the data in between. Compressed data can then be archived. Besides, SOLEIL data storage system natively includes LZ4 compression for all recorded files: compression/decompression is completely transparent to users.

HZDR Using several existing libraries to engineer high-bandwidth low-latency pipelines

Analysis pipelines including compression, data reduction or triage steps already in use or under development

EuXFEL Tools to identify frames exposed with X-Rays, Python, Linux
Batch Azimuthal Integration – framework for offline data transformation/reduction, Python, Linux
EXtra-Xwiz – framework for SFX based on CrystFEL, Python, Linux
EXtra-metro - pipeline for online analysis, Python, Linux
OnDA - pipeline for online analysis (collaborators)
Chitach/CrystFEL - pipeline for SFX offline analysis (collaborators)

DESY Crystfel (analysis including indexing) - https://www.desy.de/~twhite/crystfel/
Cheetah (includes distinguishing good and bad images) https://www.desy.de/~barty/cheetah/Cheetah/Welcome.html

HZBME Python, Fiji, IDL, Avizo

ESRF serial Xtallography (pyFAI sparsification) azimuthal integration (pyFAI / dahu)

MAX IV A collaboration on FPGA based DAQ schema for synchrotron serial crystallography with a JungFRAU detector with Filip Leonarski (PSI): Xilinx FPGA with HLS, CUDA, Power9, FPGA data processing, bs-lz4 compression, HDF5

SOLEIL MX beamlines use HDF5, HDF5plugins (Bitshuffle/LZ4), DIALS, XDS/Neggia, ADXV, Dectris ALBULA

HZDR Object detection (instance segmentation) and tracking; Foreground background segmentation

Validation software or scripts (providing metrics)

ESRF	Scripts for serial Xtallography (using XDS)
MAX IV	jnbv: https://gitlab.com/MAXIV-SCISW/JUPYTERHUB/jnbv hdf_pwrite3dc: https://github.com/zdemat/hdf_pwrite3dc mstruct: https://github.com/xray-group/mstruct rectv (lprec): https://github.com/nikitinvv/rectv_gpu jupyter-notebooks testing and validation set: https://gitlab.com/MAXIV-SCISW/JUPYTERHUB/jupyter-notebook-validation/-/tree/master/notebooks/maxiv

Possible contributions for WP7

Provide open source tools for data compression / reduction

EuXFEL Integration tool (azimuthal, along line), SAXS/WAXS, Spectroscopy, XPCS, Powder Diffraction, Python
Dark/lit frames identification or signal characterisation - any, statistics, Python/C
Peak/hit finding - SFX, Python/C
Hit finding - SPI, Python/C
Averaging/Moving averaging/Conditional averaging - SPI, Imaging, SAXS/WAXS, Spectroscopy, XPCS, Powder Diffraction, Python ML/AI for image clustering and classification (e.g. SPI)

DESY Existing tools for serial crystallography (continually being developed)
Crystfel (analysis including indexing) - https://www.desy.de/~twhite/crystfel/
Cheetah (includes distinguishing good and bad images) - https://www.desy.de/~barty/cheetah/Cheetah/Welcome.html

ESRF hdf5plugin: extension to new compression codecs. Serial Xtallography on-the-fly reduction & compression (pyFAI, LIMA2)
dynamix: sparse compression for XPCS, ...

MAX IV Thanks to adoption of Matrix based Azimuthal Integration (MATFRAIA) we are recently exploring various aspects of it, including in-cache coupled data decompression and reduction, FPGA pipelining with HLS and in particular we could be interested in compression/decompression on hw compute accelerators. This is also related somehow to general binning and histogramming algorithms.

Further simplification of distribution of HDF5 compression filters is important for us, including non-python and Windows and MacOS users.

Provide open-source tools/scripts for validation (metrics) on compressed data

EuXFEL We are planning to

ESRF 2D/3D imaging with lossy compression schemes (JPEG2000/XC)

MAX IV - There is a set of jupyter-notebooks used for validation of analysis environments with CI on top of it. Compression & reduction & triage notebooks could be tested in a similar way.

- We have performance benchmarks mainly for parallel-HDF5 and direct chunk write. Maybe a reference system could be established as the performance depends on many external factors and it is difficult to have an absolute measure.

- We have expertise in powder diffraction with a possibility to modify statistical factors for complex situation when reduced data are correlated, i.e. do not follow assumed Poisson distribution.

Provide use-case datasets

ALBA yes (serial crystallography, tomography)

EuXFEL We are in process of identifying a few data sets

DESY I expect we could provide use-case datasets for most synchrotron techniques by request to the appropriate beamline

HZBME Tomography datasets, synchrotron, neutron and SEM/FIB

ESRF
2D imaging
3D imaging (raw & and phased projections)
3D imaging
XPCS etc (open data available from http://data.esrf.fr)

MAX IV ExPaNDS reference datasets: https://doi.org/10.5281/zenodo.4558708
Tomobank: https://tomobank.readthedocs.io
Open Datasets indexed by OpenAIRE:
https://explore.openaire.eu/search/find/research-outcomes?type=%22datasets%22
or B2FIND: http://b2find.eudat.eu

SOLEIL
- Serial Xtallography (PX Beamlines)
- SAXS (SWING Beamline)
- 3D Imaging (processed) - Tomography (PSICHE and ANATOMIX Beamlines)
- Powder diffraction (CRISTAL Beamline)
- 3D Imaging and spectroscopy (NANOSCOPIUM Beamline)

HZDR ROFEX data, PiconGPU data

Input from Workshop 11Oct.

https://cxidb.org

Provide feedback from scientist

ALBA yes

EuXFEL We plan to provide

DESY Serial crystallography community

We have the usual variety of experiments at PETRA, and could naturally approach any experiment for feedback

We have other institutes onsite specialising in particular techniques: Protein crystallography at EMBL, 3D imaging at HZG

HZBME Yes, radiography and tomography, energy research, paleontology research, plant research

ESRF 2D/3D Imaging: study of JPEG{2000,XL} compression schemes
Serial Xtallography (mostly handled internally)

MAX IV Full field tomography & Imaging Synchrotron serial crystallography

SOLEIL Users' feedback: needs for large volume data transfer

HZDR Transparent integration in downstream processing in analysis pipelines (jupyter, matlab, fiji, etc.)

Summary / next steps

Highlight technique #1: tomography

- Largest data volume, high need for compression, high throughput
- Will (likely) be the focus of most of the work
- Work on compression of:
  - raw images (projections)
  - phased projections
  - 3D volume
- Various compression schemes: **lossy** or not
- The resulting images / volumes are given to the community, so it is essential that the methods used are accepted and as transparent as possible
- The acceptability of lossy compression schemes is the most difficult to evaluate, especially if we aim to archive only lossy-compressed images (irreversible). How can this be discussed within the community ?
- Upcoming seminar on the topic

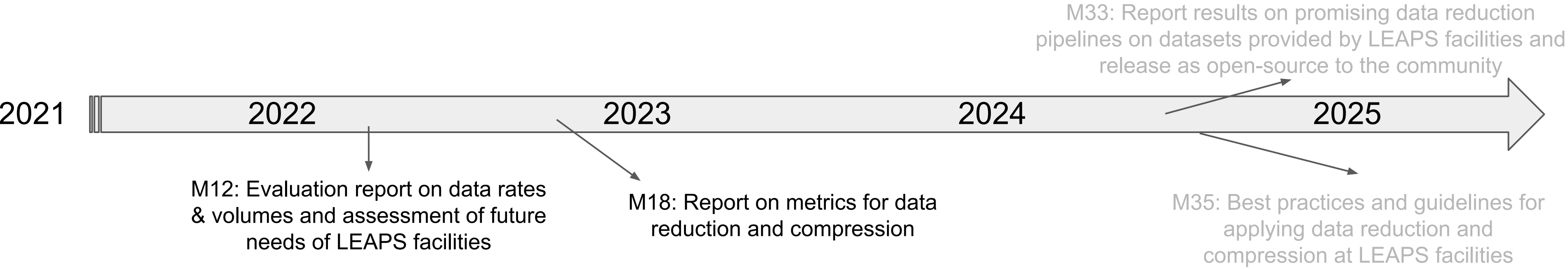
Highlight technique #2: serial Xtallography

- Very large raw data and throughput (16 Gbyte/s)
- Need on-the fly compression / reduction schemes with HPC solutions
- Not all facilities
- Solutions need to be tailored to facilities (high-throughput pipeline)
- End users do not see the compressed images. Only reduced data & the resulting model.
- Another seminar to be given to present solutions currently developed

Other techniques

- Powder diffraction / SAXS
  - Large raw data, but almost always reduced
  - Reduction schemes (2D->1D) have been established for a long time. 2D data is (mostly) irrelevant to end users
  - We can compare existing solutions/pipelines
  - Open question: do we need to keep raw data ?
- XPCS:
  - Need sparsification for the most efficient compression
- Spectroscopy:
  - Little demand for compression. Community is annoyed enough by hdf5 so generic, transparent compression should be enough
- MX, single crystal diffraction…

Timeline & deliverables



2021 2022 2023 2024 2025

M12: Evaluation report on data rates & volumes and assessment of future needs of LEAPS facilities

M18: Report on metrics for data reduction and compression

M33: Report results on promising data reduction pipelines on datasets provided by LEAPS facilities and release as open-source to the community

M35: Best practices and guidelines for applying data reduction and compression at LEAPS facilities

- Write-up of report from poll (ALBA / ESRF)
    - we may need to ask for a few additional questions
    - we’ll ask for some help writing up on specific techniques
- **Work on metrics**
    - Different types:
        - Compression ratio
        - Compression speed
        - Quality of data when using lossy compression
    - **Need to work together on various techniques** !
        - Small groups of volunteers
        - Some standard datasets notably for imaging
        - Contributions to specific techniques highly welcome