LEAPS-INNOV Workflow Co-Working Sprint Kickoff

Report of Contributions

https://events.hifis.net/e/2023-leaps-workflow-sprint

Contribution ID: **1** Type: **not specified**

Automated data processing pipelines for beamline P11

P11 at PETRA III (DESY, Hamburg) is a high-throughput instrument for macromolecular crystallography (1). P11 has tuneable photon energy between 5.5 - 28 keV having the Eiger2 X 16M as the stationary detector. The automatic sample changer at P11 has a total capacity of 23 sample pucks (total of 368 samples) having a mount-unmount cycle of approximately 36 s, which brings the beamtime spent per sample down to ca. 2min. All this enables enormous amounts of data being collected during every beamtime, which needs to be automatically processed preferably in real-time.

During the past two years mode of operation at P11 has changed from fully on-site to almost exclusively remote. Remote connection using FastX-access via a dedicated remote machine was established in 2020. Users have scientific accounts that can be used during and after the beamtime for manual data processing on Maxwell, where the autoprocessing is also migrated to dedicated P11 nodes. Currently there is only one, XDSAPP-based (2), autoprocessing pipeline running for each dataset through a simple script starting every time a data collection stops and a full dataset is detected. This workshop could help us establish parallel autoprocessing pipelines, for which there are numerous options available, depending on the type of the experiment (standard, characterisation, large-scale screening, etc.). Furthermore, finding an alternative way of detecting the full dataset, which is currently done by a workaround having a batch script inside the processing batch script, and transferring all the data for users to have access to, would be ideal.

[1] Burkhardt A., et al., Status of the crystallography beamlines at PETRA III. Eur. Phys. J. Plus 131, 56 (2016).
[2] Sparta KM, et al., XDSAPP2.0. J. Appl. Cryst. 49, 1085-1092 (2016).

Accelerator or Beamline

PETRA III

Team Contacts

Team Name

Workflow Goals

Programming Languages

Publications

Data Volume

Team Speaker

Primary author: TABERMAN, Helena (Deutsches Elektronen-Synchrotron DESY)

Co-authors: Dr GRUZINOV, Andrey (DESY); Dr POMPIDOR, Guillaume (DESY); Dr HAKANPÄÄ, Johanna (DESY); Dr CHATZIEFTHYMIOU, Spyros (DESY)

Presenter: TABERMAN, Helena (Deutsches Elektronen-Synchrotron DESY)

Contribution ID: **2** Type: **not specified**

Welcome

Friday 10 February 2023 10:00 (10 minutes)

A short introduction to the sprint

Presenter: STEINBACH, Peter (HZDR)

Contribution ID: **3** Type: **not specified**

Transparent, reproducible, and adaptable data analysis with Snakemake

Friday 10 February 2023 10:10 (30 minutes)

The Snakemake workflow management system is a tool to create transparent, reproducible and adaptable data analyses. Workflows are described via a human readable, Python based language. They can be seamlessly scaled to server, cluster, grid and cloud environments, without the need to modify the workflow definition. Finally, Snakemake workflows can entail a description of required software, which will be automatically deployed to any execution environment. With over 600,000 downloads and over 1600 citations (on average >7 per week), Snakemake is one of the most widely used systems for reproducible data analysis.

Primary author: KÖSTER, Johannes (Universitätsklinikum Essen)

Presenter: KÖSTER, Johannes (Universitätsklinikum Essen)

Contribution ID: **4** Type: **not specified**

An Introduction to Efficient and Scalable Pipeline Management with Nextflow

Friday 10 February 2023 10:40 (30 minutes)

Nextflow is an open-source workflow orchestration tool for data-intensive pipelines. It has rapidly become an industry standard, enabling scalable and reproducible scientific workflows. Nextflow has a vibrant community with thousands of bioinformaticians as part of the nf-core project, which provides ready-to-use pipelines, ready-to-plug-in modules, and sub-workflows. Nextflow simplifies the implementation and deployment of complex workflows across almost all compute infrastructures - from HPC job schedulers to all of the main cloud providers. It has built-in support for software packaging tools such as Docker, Podman, Singularity, conda, automatically managing workflow toolchains and facilitating scalable and reproducible scientific workflows.

In this talk I will introduce Nextflow and nf-core, explaining how the workflow manager works, the community tools available to streamline development, and how to get started building your own pipelines.

Write your pipeline once, and run it anywhere

Team Name

Team Speaker

Team Contacts

Accelerator or Beamline

Workflow Goals

Programming Languages

Publications

Data Volume

Presenter: RIBEIRO-DANTAS, Marcel (Seqera Labs)

Contribution ID: **5** Type: **not specified**

Introduction to CWL and Workflowhub

Friday 10 February 2023 11:15 (30 minutes)

Presenter: CRUSOE, Michael R.

Contribution ID: **6** Type: **Team proposal**

FAXTOR: the 14th beamline at ALBA for BIG DATA

The FAXTOR beamline at the third generation ALBA synchrotron will be dedicated to fast micro-tomography in the hard X-ray regime range. It will start its operation at the beginning of 2024, serving different user communities, including material science, biomedical, palaeontology, earth science, cultural heritage etc. At the start of operations, the spatial resolution (in terms of image pixel size) will range between 1 μm and 10 μm, thanks to the presence of different detection systems. The design of the endstation allows the possibility of acquiring imaging simultaneously at two different spatial resolutions if required. The possibility of performing dynamical studies with a temporal resolution < 1 s (per tomography) is foreseen. Both phase-contrast imaging (propagation-based) and absorption-based imaging techniques in radiography and tomography modes will be accessible during the experiments. The beamline is expected to present a high data throughput and making use of state-of-the art CMOS fast detectors, therefore a particular care is required in order to cope with the computing requirements.

FAXTOR's data workflow will be presented and discussed, with particular focus on data compression. Different known datasets will be made available for testing the compression algorithm performance as this procedure will become a routine at FAXTOR. Data compression will be essential for the processing and storage purpose of TB-sized datasets, typical dimension of data acquired during experiments aimed at capturing the sample dynamics or for rendering large samples at very high spatial resolution. The quality of the datasets, in terms of contrast to noise ratio, will play a major role on the compression ratio. Such a ratio will mostly depend on the experimental setting configuration used to acquire the data and on the reconstruction algorithm. A generalized algorithm capable to tune such a ratio depending on the data will be the desired solution.

Accelerator or Beamline

FAXTOR beamline at ALBA synchrotron (in design)

Team Contacts

apatera@cells.es, fcova@cells.es, gjover@cell.es

Team Name

FAXTOR team

Workflow Goals

Tomography raw data to be lossy compressed (parameters adjusted to a variety of different use cases)

Programming Languages

Python, Nextflow DSL

Publications

Data Volume

1 dataset reaches ranges from 1 to 100TB (extreme cases)

Team Speaker

Alessandra Patera, Frederico Cova

Primary authors: Dr PATERA, Alessandra (ALBA synchrotron); Dr COVA, Frederico (ALBA synchrotron)

Co-authors: Mr CENTENO, Emilio (ALBA synchrotron); Dr JOVER MAÑAS, Gabriel (ALBA synchrotron); Dr SOLER, Nicolas (ALBA synchrotron)

Presenters: Dr PATERA, Alessandra (ALBA synchrotron); Dr COVA, Frederico (ALBA synchrotron)

Contribution ID: **7** Type: **not specified**

Data post-processing and analysis pipeline

We have several python scripts for consequential data processing, DL-supported analysis, and, finally, database storage. It would be cool to see what people in community do, and how they tackle it with pipelines.

Programming Languages

Python

Publications

Data Volume

ca 10-100 TBs

Team Speaker

Accelerator or Beamline

KARA

Team Contacts

yaroslav.zharov@kit.edu

Team Name

Dunno

Workflow Goals

Data processing, data analysis, various

Primary author: ZHAROV, Yaroslav

Presenter: ZHAROV, Yaroslav

Contribution ID: **8** Type: **Team proposal**

X-ray diagnostics at European XFEL

Experiments at HED, XFEL can run up to 10Hz, providing image data from various diagnostics. In the recent years, scientists are typically using Jupyter labbooks to extract and reduce the data online and then work with the proceesed data more carefully. This reduction process however can be pretty complicated, dependent on many parameters and evolve through years, as an example the unwarping and flatfielding of SAXS scattering datata obtained by the SAXS mirror, or careful anayslis of x-ray spectra from the HAPG spectrometers. I think this process would be great candidate to use it to learn how to handle processes or pipelines in a good and fair way.

Accelerator or Beamline

European XFEL, HED instrument

Team Contacts

m.smid@hzdr.de

Team Name

Michal & Mikhail

Workflow Goals

Programming Languages

Python

Publications

https://aip.scitation.org/doi/abs/10.1063/5.0021691

Data Volume

some GBs

Team Speaker

Michal & Mikhail

Primary authors: Dr SMID, Michal (HZDR); MISCHENKO, Mikahil (European XFEL)

Presenters: Dr SMID, Michal (HZDR); MISCHENKO, Mikahil (European XFEL)

Contribution ID: **9** Type: **not specified**

Team Spotlight: P11 / DESY HH

Friday 10 February 2023 11:45 (7 minutes)

Primary author: TABERMANN, Helena

Presenter: TABERMANN, Helena

Contribution ID: **10** Type: **not specified**

Team Spotlight: FAXTOR / ALBA

Friday 10 February 2023 11:52 (7 minutes)

Presenter: PATERA, Alessandra

Contribution ID: **11** Type: **not specified**

Team Spotlight: KARA / KIT

Friday 10 February 2023 11:59 (7 minutes)

Primary author: ZHAROV, Yaroslav

Presenter: ZHAROV, Yaroslav

Contribution ID: **12** Type: **not specified**

Team Spotlight: Michal & Mikhail

Friday 10 February 2023 12:06 (7 minutes)

Presenter: SMID, Michal

Contribution ID: **13** Type: **not specified**

Team Pairing

Friday 10 February 2023 12:13 (5 minutes)

Presenter: STEINBACH, Peter (HZDR)