PAUL SCHERRER INSTITUT

PSI

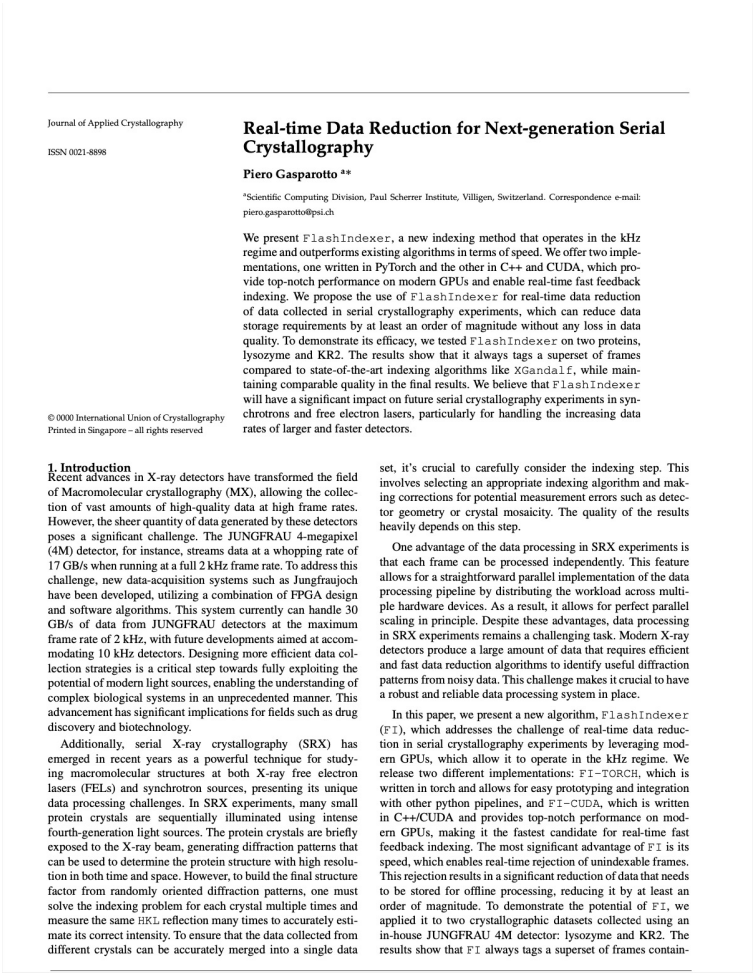
Markus Janousch ::

Activities Data Processing Development and Consulting (7902)

AWI Department Update, August 21, 2023

RED-ML

- Consolidation of previous work
- Continued with publications
- Further testing with existing data, check with established methods (XGandalf etc)
- Extension of Piero's stay until end of 2023



Journal of Applied Crystallography

ISSN 0021-8898

Real-time Data Reduction for Next-generation Serial Crystallography

Piero Gasparotto [a]*

[a]Scientific Computing Division, Paul Scherrer Institute, Villigen, Switzerland. Correspondence e-mail: piero.gasparotto@psi.ch

We present `FlashIndexer`, a new indexing method that operates in the kHz regime and outperforms existing algorithms in terms of speed. We offer two implementations, one written in PyTorch and the other in C++ and CUDA, which provide top-notch performance on modern GPUs and enable real-time fast feedback indexing. We propose the use of `FlashIndexer` for real-time data reduction of data collected in serial crystallography experiments, which can reduce data storage requirements by at least an order of magnitude without any loss in data quality. To demonstrate its efficacy, we tested `FlashIndexer` on two proteins, lysozyme and KR2. The results show that it always tags a superset of frames compared to state-of-the-art indexing algorithms like `XGandalf`, while maintaining comparable quality in the final results. We believe that `FlashIndexer` will have a significant impact on future serial crystallography experiments in synchrotrons and free electron lasers, particularly for handling the increasing data rates of larger and faster detectors.

1. Introduction

Recent advances in X-ray detectors have transformed the field of Macromolecular crystallography (MX), allowing the collection of vast amounts of high-quality data at high frame rates. However, the sheer quantity of data generated by these detectors poses a significant challenge. The JUNGFRAU 4-megapixel (4M) detector, for instance, streams data at a whopping rate of 17 GB/s when running at a full 2 kHz frame rate. To address this challenge, new data-acquisition systems such as Jungfraujoch have been developed, utilizing a combination of FPGA design and software algorithms. This system currently can handle 30 GB/s of data from JUNGFRAU detectors at the maximum frame rate of 2 kHz, with future developments aimed at accommodating 10 kHz detectors. Designing more efficient data collection strategies is a critical step towards fully exploiting the potential of modern light sources, enabling the understanding of complex biological systems in an unprecedented manner. This advancement has significant implications for fields such as drug discovery and biotechnology.

Additionally, serial X-ray crystallography (SRX) has emerged in recent years as a powerful technique for studying macromolecular structures at both X-ray free electron lasers (FELs) and synchrotron sources, presenting its unique data processing challenges. In SRX experiments, many small protein crystals are sequentially illuminated using intense fourth-generation light sources. The protein crystals are briefly exposed to the X-ray beam, generating diffraction patterns that can be used to determine the protein structure with high resolution in both time and space. However, to build the final structure factor from randomly oriented diffraction patterns, one must solve the indexing problem for each crystal multiple times and measure the same `HKL` reflection many times to accurately estimate its correct intensity. To ensure that the data collected from different crystals can be accurately merged into a single data set, it's crucial to carefully consider the indexing step. This involves selecting an appropriate indexing algorithm and making corrections for potential measurement errors such as detector geometry or crystal mosaicity. The quality of the results heavily depends on this step.

One advantage of the data processing in SRX experiments is that each frame can be processed independently. This feature allows for a straightforward parallel implementation of the data processing pipeline by distributing the workload across multiple hardware devices. As a result, it allows for perfect parallel scaling in principle. Despite these advantages, data processing in SRX experiments remains a challenging task. Modern X-ray detectors produce a large amount of data that requires efficient and fast data reduction algorithms to identify useful diffraction patterns from noisy data. This challenge makes it crucial to have a robust and reliable data processing system in place.

In this paper, we present a new algorithm, `FlashIndexer` (`FI`), which addresses the challenge of real-time data reduction in serial crystallography experiments by leveraging modern GPUs, which allow it to operate in the kHz regime. We release two different implementations: `FI-TORCH`, which is written in torch and allows for easy prototyping and integration with other python pipelines, and `FI-CUDA`, which is written in C++/CUDA and provides top-notch performance on modern GPUs, making it the fastest candidate for real-time fast feedback indexing. The most significant advantage of `FI` is its speed, which enables real-time rejection of unindexable frames. This rejection results in a significant reduction of data that needs to be stored for offline processing, reducing it by at least an order of magnitude. To demonstrate the potential of `FI`, we applied it to two crystallographic datasets collected using an in-house JUNGFRAU 4M detector: lysozyme and KR2. The results show that `FI` always tags a superset of frames contain-

SLS / SLS 2.0

Debye / SuperXAS

- Decision to use BEC framework for data processing
- Implementation of a new time-resolved detector (timePix3) for superXAS.

BEC

Involvement of Ivan with the GUI-framework (daiquiri, flint, pydm) ongoing

D3

Involvement on the data processing side

Training

Development of a course for BEC training together with the “Bildungszentrum” of PSI.