Megabytes in Murmuration: From 8-MB X-Ray Projections to Interpretable TB-Scale Structure through Automated HPC

1st Marina Nikolova
European Molecular Biology Laboratory
Hamburg, Germany
marina.nikolova@embl-hamburg.de

2nd Jonas Albers
European Molecular Biology Laboratory
Hamburg, Germany
jonas.albers@embl-hamburg.de

3rd Angelika Svetlove
European Molecular Biology Laboratory
Hamburg, Germany
angelika.svetlove@embl-hamburg.de

4th Fabio De Marco
European Molecular Biology Laboratory
Hamburg, Germany
fabio.demarco@embl-hamburg.de

5th Ksenia Denisova
European Molecular Biology Laboratory
Hamburg, Germany
ksenia.denisova@embl-hamburg.de

6th Stefan Fiedler
European Molecular Biology Laboratory
Hamburg, Germany
stefan.fiedler@embl-hamburg.de

7th Elizabeth Duke
European Molecular Biology Laboratory
Hamburg, Germany
elizabeth.duke@embl-hamburg.de

8th Gleb Bourenkov
European Molecular Biology Laboratory
Hamburg, Germany
gleb.bourenkov@embl-hamburg.de

Abstract—Synchrotron tomography beamlines routinely produce large sequences of two-dimensional X-ray projection images as their primary experimental output. Individually, these images often provide limited scientific value: the features of interest in biological samples typically emerge only after transformation into three-dimensional reconstructed volumes. This transformation is computationally demanding, requiring sustained high-throughput data transfers, and multiple processing steps – even more so for samples larger than the field of view. We describe the automated data acquisition and processing system implemented at the EMBL P14 beamline at PETRA III, Hamburg, Germany, and the underlying computing architecture that enables the beamline users to inspect interpretable three-dimensional representations of their samples shortly after X-ray exposure.

Keywords—synchrotron tomography, high-performance computing, beamline automation, X-ray imaging, tomographic reconstruction, tiled acquisition, 3D stitching

As reported in [1], a standard X-ray imaging run at P14 consists of the acquisition of four datasets at different sample-to-detector distances in rapid succession. Each dataset contains 100 flat-field frames and 1810 projections with the sample in the beam. The frames are 8 megapixels each and are acquired at the default camera rate of 100 Hz. The resulting data stream is transferred over a 40 Gb/s InfiniBand connection using ZeroMQ to an 88-core server equipped with 692 GB RAM. This server performs the first level of on-the-fly processing, e.g. flat-field correction, and raw data disk storage while keeping pace with the camera acquisition speed. Scaled-down corrected images are transferred via the beamline control system over TCP/IP and displayed live at 23 Hz. A second live display reads raw images at 25 Hz from a 1.6 PB BeeGFS storage system. Together, these displays provide an immediate view of the acquired data and allow users and beamline staff to identify problems during the experiment.

For many biological specimens, however, a three-dimensional reconstruction is the first truly meaningful representation of the data. At P14, the full reconstructed stack for a single 180-degree acquisition is produced approximately 30s after acquisition end. This is achieved by distributing the reconstruction over four cluster nodes, each with 88 CPU cores and 512 GB RAM, communicating via MPI. The reconstruction workflow is implemented in TOMO-CTF [2]. It performs flat-field correction, phase retrieval using the contrast transfer function [3], and tomographic reconstruction using the TOMOPY gridrec library [4] modified for improved throughput. The raw data are read from BeeGFS while the resulting reconstructed data are written as a TIFF stack to a dedicated 44 TB NVMe BeeGFS pool.

Extended-field-of-view acquisitions come with additional computational and algorithmic challenges. For these 360-degree scans, TOMO-CTF reconstructs a single dataset in approximately 1min 30s using two cluster nodes. The workflow includes determining the relative shift between projection pairs separated by 180 degrees, flipping one member of each pair, and combining the two frames into an extended projection before reconstruction. This procedure allows samples wider than the detector field of view to be reconstructed while preserving the automated execution model used for standard acquisitions.

Automation extends to tiled acquisitions of larger specimens. Multiple standard or extended-field-of-view acquisitions can be queued from the graphical beamline

interface, MXCuBE [5], and executed automatically in sequence. During tiled acquisitions, different regions of the sample are brought into the beam using motorized axes leaving sufficient overlap between neighboring tiles for later alignment. As soon as the final distance for a tile has been acquired, reconstruction of that tile is triggered automatically. Consequently, acquisition of the next tile proceeds in parallel with the reconstruction of the previous one, reducing the effective latency between experiment completion and availability of interpretable data.

After all reconstructions have completed, the stitching software SIBERIA [6] is launched on an 88-core cluster node to assemble the individual tiles into one aligned three-dimensional volume. SIBERIA reads data from the fast NVMe pool and writes the final stitched volume to BeeGFS. It has been used to stitch datasets of widely varying size, ranging from two standard 180-degree volumes to 105 extended-field-of-view volumes. The ready-to-be-viewed stitch is available within 20s after the reconstruction end of two standard volumes. For larger datasets, execution time increases with the number of tiles and overlaps between them, but remains within experimentally practical limits, e.g. 45min for 2×2×10 volumes, 4096×4096×2000 pixels each.

The current tomography acquisition and processing system at P14 provides beamline users with rapid 3D visualisation of complex biological samples, including punch biopsies, plants, insects, organoids, plankton, etc. By overlapping acquisition, preprocessing, reconstruction, and stitching, the system converts raw projections into interpretable volumes within the time frame of an active experiment. This automation improves experimental decision-making, supports efficient use of beamtime, and enables users to adjust acquisition strategies based on reconstructed data rather than on raw projections alone. In this way, each 8-MB projection begins as a single bird in flight; through correction, reconstruction, and stitching, thousands of such images gather into gigabyte- and terabyte-scale digital murmurations representing the biological structure.

ACKNOWLEDGMENT

ChatGPT is acknowledged for reviewing the abstract and improving sub-optimal non-native English wording. We are grateful to the EMBL Hamburg IT for support with the BeeGFS and cluster systems.

REFERENCES

[1] Albers J, Nikolova M, Svetlove A, Darif N, Lawson MJ, Schneider TR, Schwab Y, Bourenkov G and Duke E. High Throughput Tomography (HiTT) on EMBL beamline P14 on PETRA III. Journal of Synchrotron Radiation (2024). DOI: 10.1107/S160057752300944X

[2] Nikolova, M., Albers, J. and Bourenkov, G. (2026). TOMO-CTF Version 9.0. Tomographic Reconstruction Software for HiTT at P14. https://git.embl.org/grp-ximg/tomo-ctf/.

[3] [3] P. Cloetens, W. Ludwig, J. Baruchel, J.-P. Guigay, P. Pernot-Rejmánková, M. Salomé-Pateyron, M. Schlenker, J.-Y. Buffière, E. Maire, and G. Peix, “Hard x-ray phase imaging using simple propagation of a coherent synchrotron radiation beam,” J. Phys. D: Appl. Phys., vol. 32, pp. A145–A151, 1999.

[4] Gürsoy, D., De Carlo, F., Xiao, X. & Jacobsen, C. (2014). J. Synchrotron Rad. 21, 1188–1193.

[5] Oscarsson, M., Beteva, A., Flot, D., Gordon, E., Guijarro, M., Leonard, G., McSweeney, S., Monaco, S., Mueller-Dieckmann, C., Nanao, M., Nurizzo, D., Popov, A., von Stetten, D., Svensson, O., Rey-Bakaikoa, V., Chado, I., Chavas, L., Gadea, L., Gourhant, P., Isabet, T., Legrand, P., Savko, M., Sirigu, S., Shepard, W., Thompson, A., Mueller, U., Nan, J., Eguiraun, M., Bolmsten, F., Nardella, A., Milàn-Otero, A., Thunnissen, M., Hellmig, M., Kastner, A., Schmuckermaier, L., Gerlach, M., Feiler, C., Weiss, M. S., Bowler, M. W., Gobbo, A., Papp, G., Sinoir, J., McCarthy, A., Karpics, I., Nikolova, M., Bourenkov, G., Schneider, T., Andreu, J., Cuní, G., Juanhuix, J., Boer, R., Fogh, R., Keller, P., Flensburg, C., Paciorek, W., Vonrhein, C., Bricogne, G. & de Sanctis, D. (2019). J. Synchrotron Rad. 26, 393–405.

[6] Nikolova, M. & Bourenkov, G. (2026). SIBERIA. Stitching Software for HiTT at P14. Version 1.0. https://git.embl.org/grp-ximg/siberia