2023 European HDF User Group (HUG) plugins and data compression summit

Report of Contributions

https://indico.desy.de/e/ehug2023

Contribution ID: **1** Type: **Submitted talk**

GPU processing of HDF5 data

Tuesday 19 September 2023 11:30 (30 minutes)

HDF5 is the standard data format at most X-ray sources. The ESRF uses this format for both acquisition and processing of data. This contribution highlights the usage of direct-chunk read/write features of the HDF5 library and shows how it can be coupled with GPU processing.

For numerical analysis, GPUs are proven to be ~5 times faster than equally optimized CPU code on equivalent hardware. Compared to CPUs, GPUs benefit from a faster memory and from many more compute units. We present some performance comparisons for azimuthal integration, background extraction and peak localization when data is read from file using the standard HDF5-plugin pipeline or when the data is read via the direct-chunk read and decompressed on the device performing the subsequent analysis.

On the acquisition side, GPUs are now used as part of the LIMA2-framework to perform real-time reduction of data. The compression of data on the GPU can be coupled with direct-chunk write and alleviates the bottleneck of the memory bandwidth observed on computer driving the fastest detectors.

This contribution focuses on the Bitshuffle-LZ4 compression plugin, used by Eiger detectors from Dectris.

Website

https://github.com/silx-kit/pyFAI/blob/main/doc/source/usage/tutorial/Parallelization/GPU-decompression.ipynb

Primary author: Dr KIEFFER, Jerome (ESRF)

Co-authors: Dr WRIGHT, Jonathan (ESRF); Mr HOMS PURON, Alejandro (ESRF); Mr DEBIONNE, Samuel (ESRF); Dr VINCENT, Thomas (ESRF)

Presenter: Dr KIEFFER, Jerome (ESRF)

Session Classification: Day 1

Contribution ID: **2** Type: **Submitted talk**

Expanding HDF5 capabilities to support multi-threading access and new types of storage

Tuesday 19 September 2023 10:00 (30 minutes)

Enabling multi-threaded access to data stored in HDF5 and efficient storage of sparse and variable-length data are long-standing requests from the HDF5 user community. Lifeboat, LLC has been working closely with The HDF Group on design and implementation of the new capabilities.

In our talk we will report on the progress we made toward multi-threaded concurrency in HDF5 since the last European HUG at ITER in May 2022. We will also present proposed extensions to the HDF5 File format and public APIs to support sparse and variable-length data storage in HDF5. The proposed sparse storage is agnostic to memory structures used to represent sparse data in RAM (e.g., sparse matrix), and provides storage savings and portability between different memory formats. New implementation of variable-length data in HDF5 will allow significant improvements in I/O performance and will finally enable compression of the variable-length data in HDF5.

Website

Primary authors: Ms POURMAL, Elena (Lifeboat, LLC); Mr MAINZER, John (Lifeboat, LLC)

Presenter: Ms POURMAL, Elena (Lifeboat, LLC)

Session Classification: Day 1

Contribution ID: **3** Type: **Submitted talk**

NetCDF Compression Improvements

Tuesday 19 September 2023 13:30 (30 minutes)

New compression features have been added to the netCDF C and Fortran libraries, including lossy compression, zstandard, and parallel I/O support.

These features will help science data producers such as NOAA, NCAR, NASA, and ESA process, store and distribute the large scientific datasets produced by higher-resolution models and instruments.

The Community Codec Repository (CCR) will be used to bring additional compression filters to the netCDF community. Continuing to improve compression technology for the Earth science community requires collaboration and consultation to select which filters to support.

Website

Primary author: Mr HARTNETT, Edward (CIRES/NOAA)

Presenter: Mr HARTNETT, Edward (CIRES/NOAA)

Session Classification: Day 1

Contribution ID: **4** Type: **Submitted talk**

hdf5plugin: Use HDF5 compression filters from Python

Tuesday 19 September 2023 14:00 (30 minutes)

hdf5plugin is a Python package (1) providing a set of HDF5 compression filters (namely: Blosc, Blosc2, BitShuffle, BZip2, FciDecomp, LZ4, SZ, SZ3, Zfp, ZStd) and (2) enabling their use from the Python programming language with h5py a thin, pythonic wrapper around libHDF5.

This presentation illustrates how to use hdf5plugin for reading and writing compressed datasets from Python and gives an overview of the different HDF5 compression filters it provides. Finally it concludes on improvement suggestions for the HDF5 compression plugin mechanism.

Website

http://www.silx.org/doc/hdf5plugin/latest/

Primary authors: VINCENT, Thomas (ESRF); Mr SOLE JOVER, Vicente Armando (ESRF); KIEFFER, Jerome (ESRF)

Presenter: VINCENT, Thomas (ESRF)

Session Classification: Day 1

Contribution ID: **5** Type: **Submitted talk**

Current and upcoming challenges for data packaging of DECTRIS X-ray detectors

Tuesday 19 September 2023 14:30 (30 minutes)

DECTRIS X-Ray detectors are utilized at synchrotrons and laboratories around the world, where they strongly contribute to a growing accumulation of data. As we move toward the introduction of next-generation detectors, we expect a rise in both framerates and datarates. Our current pipelines that heavily rely the HDF5 data format and its corresponding software framework, wheres these pipelines have already shown limitations. These challenges are set to grow with future developments. In this presentation, we will discuss both the present difficulties and future concerns associated with high-reliability data pipelines in X-ray detectors across various applications. We will pinpoint the immediate bottlenecks we face and outline where potential future issues might arise. Our aim is simple: to initiate a meaningful discussion about key architectural decisions and to motivate the community to collaboratively address these upcoming data challenges.

Website

www.dectris.com

Primary authors: BURIAN, Max (DECTRIS Ltd.); Mr GAEMPERLE, Diego (DECTRIS Ltd.)

Co-authors: Mr GRIMM, Sascha (DECTRIS Ltd.); Ms HOTZ, Sophie (DECTRIS Ltd.)

Presenters: BURIAN, Max (DECTRIS Ltd.); Mr GAEMPERLE, Diego (DECTRIS Ltd.)

Session Classification: Day 1

Contribution ID: **6** Type: **Submitted talk**

Recent improvements in the HDF5/Blosc2 plugin systems

Tuesday 19 September 2023 11:00 (30 minutes)

Recently, the hdf5plugin (https://www.silx.org/doc/hdf5plugin) has gained support of the Blosc2 library. This allows for HDF5/h5py to use many of the technologies that Blosc2 already supports.

In our talk, we will be describing recent work that we have conducted in enhancing Blosc2, namely:

1) A new dynamic plugin system, that can be easily installed via Python wheels.

2) A new dynamic plugin for the HTJ2K (https://github.com/osamu620/OpenHTJ2K) codec. This codec has better performance and image quality scores than e.g. JPEG (https://htj2k.com/htj2k-versus-ye-olde-jpeg/).

3) Support for Blosc2 NDim inside HDF5/PyTables. Blosc2 NDim leverages a double partition (chunks and blocks) for storing data, allowing for a better utilization of L1/L2/L3 cache hierarchy in CPU caches. This makes for increased performance when reading general slices in multi-dimensional datasets. This implementation can be leveraged to do a port for h5py, and will provide hints on doing this.

4) We will briefly introduce Btune (btune.blosc.org), a tool for automatically selecting the best combination of codecs and filters based on a user-specified tradeoff between compression ratio and speed.

Most of these enhancements should be available for the HDF5/h5py via the hdf5plugin, with minimal modifications.

Website

www.blosc.org

Primary author: Mr ALTED, Francesc (Blosc project)

Presenter: Mr ALTED, Francesc (Blosc project)

Session Classification: Day 1

Contribution ID: **7** Type: **Submitted talk**

NexusCreator & ICAT - Helmholtz-Zentrum Berlin applying FAIR data management.

Wednesday 20 September 2023 10:00 (30 minutes)

https://gitlab.com/helmholtz-berlin/nexuscreator
https://gitlab.com/helmholtz-berlin/nexuscreatorpy

The research data management group at Helmholtz-Zentrum Berlin is applying FAIR data management. Data starts to be moved from specific file formats into NeXus/HDF5 files. The standardization program involves the conversion of already generated data, and the automation for the creation of NeXus files from new experiments. Our tool, NexusCreator, allows to separate the standarization process in two parts: 1) defining instruments via NeXus standard, and 2) creation of file converters or automated generation processes. NexusCreator comes in two flavours, python and javascript.

Website

Primary authors: Dr PEREZ PONCE, Hector (Helmholtz-Zentrum Berlin für Materialien und Energie GmbH); Mrs GÖRZIG, Heike (Helmholtz-Zentrum Berlin für Materialien und Energie GmbH); Mr KRAHL, Rolf (Helmholtz-Zentrum Berlin für Materialien und Energie GmbH)

Presenter: Dr PEREZ PONCE, Hector (Helmholtz-Zentrum Berlin für Materialien und Energie GmbH)

Session Classification: Day 2

Contribution ID: **8** Type: **Submitted talk**

openPMD - the Open Standard for Particle-Mesh Data

Wednesday 20 September 2023 09:30 (30 minutes)

The Open Standard for Particle-Mesh Data (openPMD) is a F.A.I.R. metadata standard for tabular (particle/dataframe) and structured mesh data in science and engineering.
We show the basic components of openPMD, its extensions to specific domains, applications from laser-plasma physics, particle accelerators, material physics to imaging and the ability to bridge multiple heterogeneous scientific models with a commonly-understood markup.

The openPMD-api builds upon established portable I/O formats such as HDF5 and ADIOS2, enabling workflows that scale from single-user computers up to Exascale simulations, in-transit data processing, 3D visualization, GPU-accelerated data analytics and AI/ML. openPMD links into the existing ecosystems of its scalable I/O backends and extends them with tooling that understands the openPMD data markup.
An overview over the openPMD ecosystem and community is shown.

Attention is given to recent developments in openPMD that interplay with HDF5, including mesh refinement and the Helmholtz Metadata Collaboration's HELPMI project which aims for an easier integration of openPMD with other HDF5-based standards, this way bringing openPMD closer to experiment workflows.

References:
1 Axel Huebl, Remi Lehe, Jean-Luc Vay, David P. Grote, Ivo F. Sbalzarini, Stephan Kuschel, David Sagan, Christopher Mayes, Frederic Perez, Fabian Koller, and Michael Bussmann. "openPMD: A meta data standard for particle and mesh based data,"DOI:10.5281/zenodo.591699 (2015)
2 Homepage: https://www.openPMD.org
3 GitHub Organization: https://github.com/openPMD
[4] Projects using openPMD: https://github.com/openPMD/openPMD-projects
[4] Reference API implementation: Axel Huebl, Franz Poeschel, Fabian Koller, and Junmin Gu. "openPMD-api 0.14.3: C++ & Python API for Scientific I/O with openPMD,"DOI:10.14278/rodare.1234 (2021)
https://openpmd-api.readthedocs.io
[5] Selected earlier presentations on openPMD:
https://zenodo.org/search?page=1&size=20&q=openPMD&type=presentation
[6] Axel Huebl, Rene Widera, Felix Schmitt, Alexander Matthes, Norbert Podhorszki, Jong Youl Choi, Scott Klasky, and Michael Bussmann. "On the Scalability of Data Reduction Techniques in Current and Upcoming HPC Systems from an Application Perspective,"ISC High Performance 2017: High Performance Computing, pp. 15-29, 2017. arXiv:1706.00522, DOI:10.1007/978-3-319-67630-2_2
[7] Franz Poeschel, Juncheng E, William F. Godoy, Norbert Podhorszki, Scott Klasky, Greg Eisenhauer, Philip E. Davis, Lipeng Wan, Ana Gainaru, Junmin Gu, Fabian Koller, Rene Widera, Michael Bussmann, and Axel Huebl. Transitioning from file-based HPC workflows to streaming data pipelines with openPMD and ADIOS2, Part of Driving Scientific and Engineering Discoveries Through the Integration of Experiment, Big Data, and Modeling and Simulation, SMC 2021, Communications in Computer and Information Science (CCIS), vol 1512, 2022. arXiv:2107.06108, DOI:10.1007/978-3-030-96498-6_6
[8] The Helmholtz Metadata Collaboration's ongoing HELPMI project: https://helmholtz-metadaten.de/de/inf-projects/helpmi-helmholtz-laser-plasma-metadata-initiative

Website

https://github.com/openPMD

Primary author: POESCHEL, Franz (CASUS/HZDR)

Co-authors: HUEBL, Axel (LBNL); BUSSMANN, Michael (CASUS / Helmholtz-Zentrum Dresden - Rossendorf)

Presenter: POESCHEL, Franz (CASUS/HZDR)

Session Classification: Day 2

Contribution ID: **9** Type: **Submitted talk**

[Tutorial] Use of Btune for finding best codecs/filters for Blosc2

Thursday 21 September 2023 09:00 (3h 30m)

Btune (https://www.blosc.org/pages/btune/) is a dynamic plugin for Blosc2 that can help you find the optimal combination of compression parameters for datasets compressed with Blosc2. Blosc2 can easily be used from HDF5/h5py via the hdf5plugin (https://www.silx.org/doc/hdf5plugin).

Depending on your needs, Btune has three different tiers of support for tuning datasets:

- **Genetic (Btune Free)**: A genetic algorithm tests different combinations of compression parameters to meet the user's requirements for both compression ratio and speed for each chunk in the dataset.
- **Trained (Btune Models)**: The Blosc development team train neural network models that enable Btune to predict the best compression parameters for user's datasets.
- **Fully managed (Btune Studio)**: Enables users for doing on-site training of an unlimited number of datasets. Requires a license.

In this tutorial, we will use Btune in these three different modes. Users wanting to explore the best compression codecs/filters for their cases are advised to bring their own datasets and use the techniques learnt for finding them out.

Time for the tutorial: 3.5 hours (including a 30 min break).

Website

https://www.blosc.org/pages/btune/

Primary author: Mr ALTED, Francesc (Blosc project)

Presenter: Mr ALTED, Francesc (Blosc project)

Session Classification: Day 3

Contribution ID: **10** Type: **Submitted talk**

h5cpp and pninexus c++ libraries

Wednesday 20 September 2023 09:00 (30 minutes)

The h5cpp library developed by DESY and ESS is a c++ wrapper for the HDF5 library. Using modern c++ features it simplifies creation of HDF5 files. The pninexus library adds a set of advance tools, e.g. a file structure builder from XML configuration. The libraries with their python binding are heavily used by the PETRA III experiment @ DESY in our detector Tango servers and our NeXus metadata framework.

Website

Primary author: KOTANSKI, Jan (DESY. FS-EC)

Presenter: KOTANSKI, Jan (DESY. FS-EC)

Session Classification: Day 2

Contribution ID: **12** Type: **Submitted talk**

Lossless and Lossy Compression for Photon Science

Wednesday 20 September 2023 12:00 (30 minutes)

High bandwidth instruments (data production rates of GB/s) have proliferated in photon science experimental facilities in the last years across the globe. Some of them are planned to be operated 24/7. Data volumes thus produced exceed both the budget of storage facilities and sometimes even the ingest capacities of hardware.

In this talk, I'd like to highlight key challenges when considering both lossless and lossy compression in photon science. I will highlight data science approaches to characterize or preprocess data. The talk will also showcase advances in finding optimal encoding parameters to achieve high data ingest bandwidths at high compression ratios. In addition, I'd like to introduce challenges for lossy compression with respect to good scientific practice and our advances to mitigate them without regressing to data quality metrics.

Website

Primary author: STEINBACH, Peter (HZDR)

Presenter: STEINBACH, Peter (HZDR)

Session Classification: Day 2

Contribution ID: **13** Type: **Keynote talk**

Welcome and introduction

Tuesday 19 September 2023 09:00 (20 minutes)

Welcome to participants
Goals of workshop
Organizational matters

Presenters: PENNICARD, David (FS-DS (Detektorsysteme)); HEBER, Gerd (The HDF Group)

Session Classification: Day 1

Contribution ID: **14** Type: **not specified**

HDF5 and plugins - overview and roadmap

Tuesday 19 September 2023 09:20 (40 minutes)

Presenter: Dr ROBINSON, Dana (HDF Group)

Session Classification: Day 1

Contribution ID: **15** Type: **not specified**

Discussion

Tuesday 19 September 2023 16:00 (45 minutes)

Session Classification: Day 1

Contribution ID: **16** Type: **not specified**

Discussion

Wednesday 20 September 2023 15:30 (1h 30m)

Session Classification: Day 2

Contribution ID: **17** Type: **Submitted talk**

Processing HDF5 data with FPGAs

Tuesday 19 September 2023 12:00 (30 minutes)

HDF5 format is used to store experimental data from photon and neutron sources worldwide. Field programmable gate arrays (FPGAs) are recently finding applications in data acquisition on accelerator-based photon sources. FPGAs can be used also as regular compute accelerators similarly to general purpose graphical processing units. Options for feeding FPGA data reduction algorithms with compressed data and HDF5 format are discussed. X-ray scattering data compressed with popular bslz4 filter are of a particular interest.

Website

https://gitlab.com/MAXIV-SCISW/compute-fpgas

Primary author: MATEJ, Zdenek (MAX IV Laboratory, Lund University)

Co-author: Dr SALNIKOV, Andrii (MAX IV Laboratory, Lund University)

Presenter: MATEJ, Zdenek (MAX IV Laboratory, Lund University)

Session Classification: Day 1

Contribution ID: **18** Type: **not specified**

DESY tour (PETRA / FLASH)

Thursday 21 September 2023 15:00 (1 hour)

Contribution ID: **19** Type: **not specified**

Discussion

Wednesday 20 September 2023 14:00 (1 hour)

Submitted topics for discussion:

–

Thomas Vincent - Managing compression filters in the mid- to long-term

–

Erik Maansson - Run-length encoding for mostly black images

About ongoing work where a much simpler compression algorithm (customized run-length encoding at the application-level) may be faster than the built-in ones.

For multi-detector covariance analysis, e.g. between photoelectron energy and mass of ionic fragments of a molecule, it is necessary to save raw data (e.g. images and spectra) separately for each laser shot, rather than only saving the average over many laser shots. This can lead to higher rates of raw data than a typical computer or storage medium can handle at the required repetition rate of the experiment, unless suitable (lossy) compression is applied.

In my application, we use a CMOS camera to acquire images at 1 kHz from which angularly-resolved electron velocities can be determined. The largest square image size of 1024x1024 pixels gives 2.1 TB/s of raw data, which in my experience is too much to save locally, with or without available HDF5 compression libraries, by a single computer (via pytables in Python & numba, on an Intel Xeon E5-1620 v4 3.5 GHz from 2018). For mass spectra, sampled waveform data (ADC) is also acquired, but this is two orders of magnitude less data and therefore not setting the speed limit.

However, by knowing that our kind of image normally contains less than a few hundred bright spots (detected electrons), each covering a few pixels, it becomes worthwhile to find a more efficient (lossy) encoding that still maintains the scientifically meaningful information. After subtracting a dark image, pixels darker than a threshold value are therefore set to zero. Currently the resulting image is then passed to HDF5 for compression with LZO to about 1/10th to 1/50th of its raw its size.

The required 1 kHz continuous processing and saving rate is achieved by letting the camera bin groups of 2x2 pixels, so that the software only sees 512x512 pixels. It would be desirable to be able to use the full 1024x1024 pixels, and perhaps the standard compression algorithms are wasting CPU-time by trying to be "smart"when the main way that my mostly-black images can be compressed is to get rid of all the zeroes. I have begun implementing a run-length encoding scheme where a run of successive values below a user-chosen threshold are encoded by a negative value (the length of the run) in the array of signed 16-bit integers. Bright pixels remain as positive values. This yields several times higher compression ratios than LZO, and compiled to machine-code with numba (LLVM) it runs at a speed where it seems interesting to implement for full-scale testing in the acquisition program. HDF5's variable-length array does not seem performant enough to store the compressed result from each individual image, so solutions concatenating the compressed form of many (or all) images will be explored.

Website

Session Classification: Day 2

Contribution ID: **21** Type: **Submitted talk**

Compression Plugins in h5wasm (javascript/webassembly)

Tuesday 19 September 2023 15:30 (30 minutes)

H5wasm is a webassembly-based library for reading and writing HDF5 files, which can be used natively in a web browser or in a local nodejs environment. The library has no external runtime dependencies, and is used in some online HDF5 viewers that don't require server-side processing: https://h5web.panosc.eu/h5wasm and https://myhdf5.hdfgroup.org/

The community has requested more compression plugins (e.g. ZSTANDARD) for h5wasm beyond the (included) DEFLATE, SHUFFLE, FLETCHER32 and SCALEOFFSET filters. In my talk I will discuss issues associated with adding plugins to h5wasm
For collaborative work on h5wasm, a change from single-maintainer in a private organization (github/usnistgov)
Incomplete support for dynamic linking in emscripten (MAIN_MODULE/SIDE_MODULE)
Complex dependency chains for some plugins (all libraries have to be compiled to WASM)
Browser limitations (e.g. max 4KB dynamic WASM loading in Chrome)
I will demonstrate a proof-of-concept build of h5wasm including a ZSTANDARD plugin, and discuss why I was not able to easily build an LZ4 plugin.
We can discuss a shared effort on building a repository for h5wasm like the the h5py plugins at https://github.com/silx-kit/hdf5plugin, also based on https://github.com/HDFGroup/hdf5_plugins.
We could use people with skills in CMake, Emscripten, TypeScript and of course the HDF5 C API.

Website

Primary author: MARANVILLE, Brian (NIST)

Presenter: MARANVILLE, Brian (NIST)

Session Classification: Day 1

Contribution ID: **22** Type: **not specified**

Travel to restaurant (for those joining)

Tuesday 19 September 2023 16:45 (1h 15m)

Portuguese restaurant, Ola Lisboa. https://ola-lisboa.de/
https://goo.gl/maps/V2i15XbgpuhAYPbE6

We will be going by public transport (self-paid) - if you don't yet have a ticket you can buy one on the bus.
People from DESY will help with directions. The route is below:
- Go to bus stop outside the DESY main entrance, on the opposite side of the road
- Take Bus 1 (Bf. Altona) until S Othmarschen –approximately 10 minutes
- At the S-Bahn station, take S1 in the direction Airport/Poppenbuttel
- Get off at Landungsbruken –approximately 13 minutes
- Short walk to the restaurant

Contribution ID: **23** Type: **not specified**

Restaurant Ola Lisboa (for those joining)

Tuesday 19 September 2023 18:00 (2 hours)

*** Please let me know by email if you wish to join –david.pennicard@desy.de. If more people than expected want to join I will try to book more places, but this isn't guaranteed, in which case it's first come, first served! ***

We will be going to a Portuguese restaurant, Ola Lisboa. It offers a range of dishes, including vegetarian, with seafood being a speciality.
https://ola-lisboa.de/

This is in the "Portuguese quarter" of the city, at Hamburg harbour. So, there will also be the opportunity to see the riverside, which is one of the main attractions of the city.
https://goo.gl/maps/V2i15XbgpuhAYPbE6

Since the workshop does not have a registration fee, you will have to pay for yourselves, and we will be travelling by public transport; there is a subway station next to the restaurant, so it is reasonably convenient for travel back to your hotels.

Ditmar-Koel-Straße 18, 20459 Hamburg

Website

Contribution ID: **24** Type: **Submitted talk**

Using Sparse Arrays for Synchrotron 3D-XRD-CT Data Reduction.*

Wednesday 20 September 2023 11:00 (30 minutes)

The materials science beamline, ID11, at the ESRF, was upgraded in 2020 to get a Dectris Eiger 4M pixel detector. This can record diffraction frames at 500 Hz while samples are rotated and scanned in a tiny (~150 nm) X-ray beam. Reconstruction of the diffraction data can eventually give detailed images of all the crystals inside the materials. The large quantities of data can be problematic to process, a single scan may contain millions of frames. This contribution will review our experience in the last few years working with these data. Raw frames are recorded into hdf5 files using bitshuffle and lz4 compression. For many of the larger datasets, the diffraction data are very sparse, so converting to a sparse format helps a lot. While this first step is bounded by IO and decompression, but the format conversion can be done in parallel over frames. The rest of our processing is based on these sparse data rather than full images.

Website

https://github.com/jonwright/bslz4_to_sparse

Primary author: WRIGHT, Jonathan (ESRF)

Presenter: WRIGHT, Jonathan (ESRF)

Session Classification: Day 2

Contribution ID: **25** Type: **Submitted talk**

Data reduction in serial crystallography

Wednesday 20 September 2023 11:30 (30 minutes)

Serial crystallography (SX) has become an established technique for protein structure determination, especially when dealing with small or radiation-sensitive crystals and investigating fast or irreversible protein dynamics. The advent of newly developed multi-megapixel X-ray area detectors, capable of capturing over 1000 images per second, has brought about substantial benefits. However, this advancement also entails a notable increase in the volume of collected data. Today, up to 2 PB of raw data per experiment could be easily obtained under efficient operating conditions. The combined costs associated with storing data from multiple experiments provide a compelling incentive to develop strategies that effectively reduce the amount of data stored on disk while maintaining the quality of scientific outcomes. Lossless data compression methods are designed to preserve the information content of the data but often struggle to achieve a high compression ratio when applied to experimental data that contains noise. Conversely, lossy compression methods offer the potential to greatly reduce the data volume. Nonetheless, it is vital to thoroughly assess the impact of data quality and scientific outcomes when employing lossy compression, as it inherently involves discarding information. The evaluation of lossy compression effects on data requires proper data quality metrics.

Our focus here is to evaluate different lossless and lossy data compression methods and determine the appropriate metrics for evaluating the impact of lossy compression on the final SX data quality. Our research found that effective strategies for lossy data reduction in SX are: non-hits rejection (in the case of strongly diffracting crystals), binning (in the case of crystals with small unit cell) and reduction in the precision of the measured diffraction pattern intensities, especially in a non-uniform way (saving only several most significant bits). At the same time, we demonstrate the potential risks associated with particular lossy data reduction schemes, such as: reduction in the number of stored diffraction patterns, saving only the intermediate results (.mtz files), or saving only the regions around the detectable Bragg peaks.

Some of the lossy compression schemes, that we have developed and/or tested, can be implemented either in hardware or as HDF5 plugins for application in crystallography as well as for data generated using other techniques.

Website

Primary author: GALCHENKOVA, Marina (FS-CFEL-1 (Forschung mit Photonen Experimente 1))

Co-authors: CHAPMAN, Henry (FS-CFEL-1 (Forschung mit Photonen Experimente 1)); YEFANOV, Oleksandr (FS-CFEL-1 (Forschung mit Photonen Experimente 1))

Presenter: GALCHENKOVA, Marina (FS-CFEL-1 (Forschung mit Photonen Experimente 1))

Session Classification: Day 2

Contribution ID: **26** Type: **Topic for discussion**

Run-length encoding for mostly black images

About ongoing work where a much simpler compression algorithm (customized run-length encoding at the application-level) may be faster than the built-in ones.

For multi-detector covariance analysis, e.g. between photoelectron energy and mass of ionic fragments of a molecule, it is necessary to save raw data (e.g. images and spectra) separately for each laser shot, rather than only saving the average over many laser shots. This can lead to higher rates of raw data than a typical computer or storage medium can handle at the required repetition rate of the experiment, unless suitable (lossy) compression is applied.

In my application, we use a CMOS camera to acquire images at 1 kHz from which angularly-resolved electron velocities can be determined. The largest square image size of 1024x1024 pixels gives 2.1 TB/s of raw data, which in my experience is too much to save locally, with or without available HDF5 compression libraries, by a single computer (via pytables in Python & numba, on an Intel Xeon E5-1620 v4 3.5 GHz from 2018). For mass spectra, sampled waveform data (ADC) is also acquired, but this is two orders of magnitude less data and therefore not setting the speed limit.

However, by knowing that our kind of image normally contains less than a few hundred bright spots (detected electrons), each covering a few pixels, it becomes worthwhile to find a more efficient (lossy) encoding that still maintains the scientifically meaningful information. After subtracting a dark image, pixels darker than a threshold value are therefore set to zero. Currently the resulting image is then passed to HDF5 for compression with LZO to about 1/10th to 1/50th of its raw its size.

The required 1 kHz continuous processing and saving rate is achieved by letting the camera bin groups of 2x2 pixels, so that the software only sees 512x512 pixels. It would be desirable to be able to use the full 1024x1024 pixels, and perhaps the standard compression algorithms are wasting CPU-time by trying to be "smart"when the main way that my mostly-black images can be compressed is to get rid of all the zeroes. I have begun implementing a run-length encoding scheme where a run of successive values below a user-chosen threshold are encoded by a negative value (the length of the run) in the array of signed 16-bit integers. Bright pixels remain as positive values. This yields several times higher compression ratios than LZO, and compiled to machine-code with numba (LLVM) it runs at a speed where it seems interesting to implement for full-scale testing in the acquisition program. HDF5's variable-length array does not seem performant enough to store the compressed result from each individual image, so solutions concatenating the compressed form of many (or all) images will be explored.

Website

https://atto.cfel.de

Primary author: MAANSSON, Erik (FS-ATTO (Attosecond Science and Technology))

Contribution ID: **27** Type: **Submitted talk**

Highly Scalable Data Service (HSDS) Tutorial

Thursday 21 September 2023 13:30 (1 hour)

Hands on tutorial for running HSDS. HSDS is a RESTful service for HDF data that can be used in cloud, desktop, or HPC environments. Tutorial will cover:
HSDS architecture
Installing HSDS
Configuration Options
HSDS command line tools
HSDS compression
Accessing HSDS with REST, Python, and C (rest-vol)

TO JOIN THE TUTORIAL, YOU NEED:
Install Anaconda Python from: https://www.anaconda.com/download
Create an anaconda environment: \ $condcreate-nhsdsworkshoppython = 3.9Activatetheenvironment :$
$\$condaactivatehsdsworkshop$
$Installhsds :$ \ pip install hsds Install h5py: \$ pip install h5py
Install h5pyd: \\$ pip install h5pyd

Website

Primary author: READEY, John (HDF Group)

Presenter: READEY, John (HDF Group)

Session Classification: Day 3

Contribution ID: **28** Type: **Submitted talk**

HDF Compression for data service architectures

Wednesday 20 September 2023 13:30 (30 minutes)

HSDS (Highly Scalable Data Service) is a REST-based web service that supports most of the features of the HDF library, but running as a service. HSDS supports the standard HDF compressors as well as BLOSC-based compressors "out of the box". In addition, HSDS supports parallel compression/decompression and supports using compression with variable length datatypes. This talk will cover how the HSDS architecture supports these features and some ideas for future development.

Website

Primary author: READEY, John (HDF Group)

Presenter: READEY, John (HDF Group)

Session Classification: Day 2

Contribution ID: **29** Type: **not specified**

Videos are also available - link available in this contribution's ”materials” section