

Introducing HDF5 - A new storage format for your data

Chris Hall
IMBL

Our current storage demands

- IMBL CT requires > 1000 images to be collected and stored.
- IMBL allows moderate resolution for each image over a wide area \Rightarrow many pixels per image
- The SR beam and detector limits means individual projections can be made of several images.
- \Rightarrow Many arrays need to be stored.
- E.g. Elvis the rhino: raw data 1.31 TB in 30 CT sets: 336,390 files.

Future storage issue

- IMBL is designed for moderate spatial resolution imaging of large objects. We have not yet fully exploited this capability
- As well as X, Y, theta, & Z we might also add time, and even a spectral dimension to the data stream
- => The quantity and speed of data that needs to be stored is going to increase

Data storage efficiency

- Storing each image in a single file is very inefficient.
 - Data storage is slow
 - Directories take long times to list
 - Management of the data is slow and awkward
 - It is separated from the 'meta-data' (everything you would normally record in your experiment log book.)

Towards a data store standard...

- There have been several attempts to agree on a tomography data storage format e.g.
 - APS : Scientific Data Exchange
 - International: NeXus and PANData Formats
- Many rely on the concept of a data ‘container’ which keeps relevant data together in one place.

Hierarchical Data Format 5 (HDF5)

- An HDF5 file is a container for storing various data
- An HDF5 file is self describing... You can figure out where and what the data is by looking at it
- HDF5 is mature, and used in many other areas e.g. financial services
- It is composed of two primary types of objects: groups and datasets.
 - **HDF5 group:** a grouping structure containing zero or more HDF5 objects, together with supporting metadata
 - **HDF5 dataset:** a multidimensional array of data elements, together with supporting metadata

Introduction to HDF5

- Any HDF5 group or dataset may have an associated attribute list.
 - An **HDF5 attribute** is a user-defined HDF5 structure that provides extra information about an HDF5 object.
- Working with groups and datasets is similar in many ways to working with directories and files in Linux. As with Linux directories and files, an HDF5 object in an HDF5 file is often referred to by its **full path name** (also called an **absolute path name**).
 - / signifies the root group.
 - /blah signifies a member of the root group called blah.
 - /blah/blah signifies a member of the group blah (which in turn is a member of the root group blah)

The goal for HDF5 on IMBL:

- Each sample will have all its raw data saved in a single HDF5 file. (This will eventually include the calibration images (F&D), but probably not in the first instance)
- Serial scans and other protocols will be kept in an N-dimensional data array within the HDF5 file
- Stitched and corrected projection images will be stored in a separate HDF5 file
- Reconstructed data will be stored in a third HDF5 file.

Why this protocol?

- Raw data can be collected at the highest speeds into a single file
- Kept in the /input tree of the filestore this will be archived automatically.
- Processed data requires programs to read the data. Currently these work on TIFF file stacks. It is easy to unload data from the HDF5 file to TIFF stacks.

The role of AreaDetector

- All our imagers use an EPICS AreaDetector system to control, read, and store image data
- AreaDetector has a plugin which will take the data and save it to the HDF5 file along with the instrument attributes
- Storing to HDF5 works in either Stream or Capture mode

An example HDF5 set-up

NDFileHDF5.adl

13SIM1:HDF1:

asyn port FileHDF1 Plugin type NDFileHDF5 ver1.10.0 ADCore version 2.6.0 Plugin version 2.6.0 Array port SIM1 SIM1 Array address 0 0 Enable Enable Enable Min. time 0.000 0.000 Callbacks block No No Queue size/free 20 20 Array counter 0 720 Array rate 197.00 Execution time 2.199 msec Dropped arrays 0 0 # dimensions 2 Array Size 1024 1024 0 Data type UInt8 Color mode Mono Bayer pattern RGGB Unique ID 61668 Time stamp 849993152.877 Attributes file Array callbacks Disable Disable asyn record 	File path /home/epics/scratch/ Exists: Yes File name test Create dir. depth 0 0 Help Next file # 21 21 Temp. suffix Auto increment Yes Yes Lazy open No No Filename format %s_%3.3d.h5 Example: %s_%3.3d.h5 Last filename /home/epics/scratch/test_020.h5 Save file Save Read Auto save No No Write mode Stream Stream # Capture 1000 1000 720 Capture Start Stop Delete driver file No No Write status Write OK Write message
Rows per chunk 0 1024 Columns per chunk 0 1024 Frames cached per chunk 0 1 Boundary alignment 0 0 Boundary threshold 65536 65536 Flush on N'th frame 0 1 Fill value 0.0 0.0	Compression zlib None # data bits 8 8 Data bits offset 0 0 SZip # pixels 16 16 Zlib level 6 6 Store performance Yes Yes Store attributes Yes Yes Run time 0.012 I/O speed 664.3 Default layout selected Exists: Yes XML File name

SWMR Support
 SWMR supported **Supported**
 SWMR mode **Off** **Off**
 SWMR active **Off**
 SWMR callbacks **0**
 More

Example HDF5 on IMBL

The screenshot displays the HDFView application window, which is used for inspecting and visualizing HDF5 files. The interface includes a menu bar (File, Window, Tools, Help), a toolbar, and a 'Recent Files' list showing '/scratch/tmp/test_1.h5'.

The main panel is divided into several sections:

- Left Panel (Tree View):** Shows the hierarchical structure of the HDF5 file. The root is 'test_1.h5', which contains a group 'entry'. Under 'entry', there are groups 'data', 'instrument', and 'performance'. The 'data' group contains a dataset 'data' with various attributes like 'AcquireTime', 'CameraManufacturer', 'CameraModel', 'E', 'GetHistory', 'ID_Energy', 'ID_Energy_EQU', 'ImageCounter', 'MaxSizeX', 'MaxSizeY', 'NDArrayEpicsTSsec', 'NDArrayEpicsTSsec', 'NDArrayTimeStamp', 'NDArrayUniquid', 'PI', 'RingCurrent', 'RingCurrent_EQU', 'Ten', 'detector', 'NDAttributes', 'ColorMode', and 'data'. The 'performance' group contains a dataset 'timestamp'.
- Central Panel (ImageView):** Displays a grayscale image of a detector. The title bar indicates 'ImageView - data - /entry/data/ - /scratch/tmp/test_1.h5 - 800.0%'. The image is a 10x10 grid of pixels, with a color bar on the right showing values from 0.00E0 to 1.00E2.
- Right Panel (TableViews):** Contains three table views:
 - TableView - ImageCounter:** Displays a 10x10 grid of values, likely representing the image data.
 - TableView - NDArrayTimeStamp:** Displays a 10x10 grid of values, likely representing the timestamp data.
 - TableView - data:** Displays a 10x10 grid of values, likely representing the data array.
- Bottom Panel (Log Info / Metadata):** Shows metadata for the 'data' dataset (5644). The metadata includes:
 - 8-bit unsigned character, 10 x 40 x 60
 - Number of attributes = 5
 - NDArrayDimBinning = 1, 1
 - NDArrayDimOffset = 0, 0
 - NDArrayDimReverse = 0, 0
 - NDArrayNumDims = 2
 - NK_class = SDS
 - signal = 1

Take home message:

- If you want to collect a CT set with a short exposure times
- If you are interested in switching to a more manageable data format
- **Ask us about using HDF5**