Eiger, HDF5, XDS and autoPROC

Clemens Vonrhein[*], Kay Diederichs[¶],
Claus Flensburg[*], Andreas Förster[§],
Gérard Bricogne[*]

[*]Global Phasing Ltd., UK
[¶]University Konstanz, DE
[§]Dectris Ltd., CH

High Data-Rate Macromolecular Crystallography
NSLS-II Brookhaven National Laboratory
26th - 28th May 2016

XDS and autoPROC - Introduction

- **XDS** is a data-processing package developed by Wolfgang Kabsch (MPI for Medical Research, Heidelberg, DE) and Kay Diederichs (University Konstanz, DE):
  - http://xds.mpimf-heidelberg.mpg.de/
  - http://strucbio.biologie.uni-konstanz.de/xdswiki/

- **autoPROC** adds automation and expertise to data processing with XDS as well as additional capabilities (e.g. visualisation) and is developed by Global Phasing Ltd. (Cambridge, UK):
  - uses XDS, POINTLESS, AIMLESS (P. Evans) and CCP4
  - http://www.globalphasing.com/autoproc/

XDS – current handling of HDF5 (1)

- **Meta data** need to be extracted beforehand:
  - different tools/systems and scripts available to read *_master.h5 file
  - autoPROC: *imginfo* reads all supported image formats and extracts required information for processing
- Alternatively, write XDS.INP file directly from beamline control software:
  - Advantages: all data available and any local configuration change can be accommodated
  - Disadvantage: difficult to transfer processability to other programs or program versions and different installations (especially if local configuration and calibration data is not archived and/or associated with the data itself)
- Supplied to XDS via **XDS.INP** file to describe **single-axis** MX experiment

- **Nothing changed here: standard XDS design for any data format.**

XDS and HDF5 history

From 2007, the Pilatus detector enabled improved ways to collect data in macromolecular crystallography. These allow the faster solution of more accurate crystal structures from biological macromolecules - the driving force for the improvement of techniques in X-ray crystallography.

Finely-sliced (0.1° rotation) datasets typically consist of thousands of 6MB frames (Pilatus 6M), covering 180°-360°. Together with short exposure times (0.1s), this leads to “high” data rates. The Eiger detector can deliver even higher data rates.

The result of an earlier meeting (Dectris Eiger Workshop, 2013) was the decision to store the data of the Eiger detector in the HDF5 archiving format.

Problems and workarounds

The official HDF5 library does not efficiently support parallel (threaded) reading of files. This deficiency prevents fast processing with **XDS**, which employs **two levels of parallelization**: **OpenMP** and **multiple clients**. An attempt to use the official HDF5 library (in 2013/2014) revealed HDF5 I/O as the bottleneck.

At DEW 2013, the XDS team (Wolfgang Kabsch and KD) agreed to support an **external conversion program**, later provided by Dectris. This "**H5ToXDS**" reads the name of the HDF5 master file, a frame number, and the name of the output file from its command arguments. It produces a **CBF file without header** which is then read by XDS.

XDS with support for H5ToXds (i.e. running H5ToXds for each frame) became available in **November 2014**. Parallel processing scales well with the number of threads, but the **overhead of writing and reading an intermediate data file** is substantial (~100%). The "Eiger" article in XDSwiki describes how to reduce this overhead to ~24%, compared to processing of CBF files.

XDS – current handling of HDF5 (2)

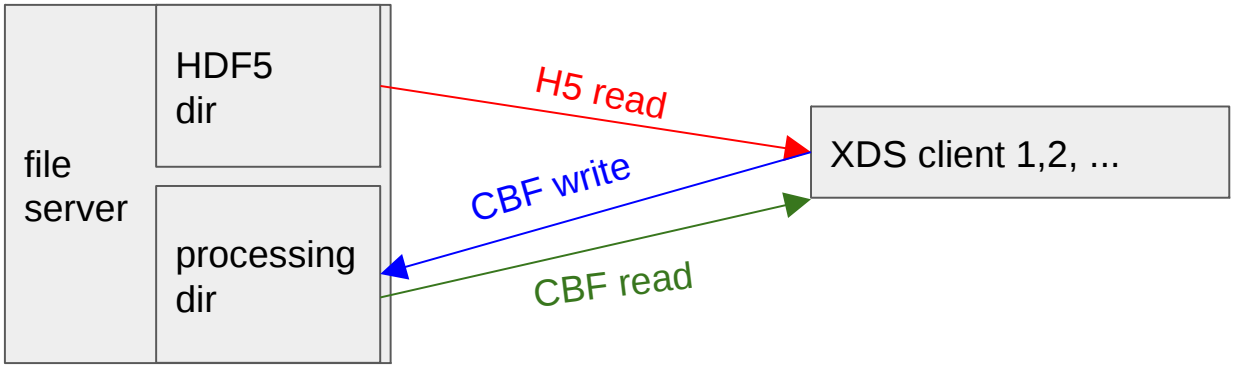
On-the-fly conversion of HDF5 to mini-cbf file:

- mini-cbf header content irrelevant (since not read by XDS)
- implemented via system-call from Fortran (XDS developers)
- expects external program ***H5ToXds*** (Dectris) in $PATH
- converted (temporary) mini-cbf file removed after reading
- tricks for speeding this up:
  - writing files into shared memory or fast local disk/SSD (instead of networked FS)
  - sensible alignment of nodes/threads with *DELPHI=* parameter and number of images to be processed

- see also e.g.:
  - **http://strucbio.biologie.uni-konstanz.de/xdswiki/index.php/Eiger**
  - **http://www.globalphasing.com/autoproc/wiki/index.cgi?DataProcessingHdf5#hdf5converter**

Working around the overhead of H5ToXds

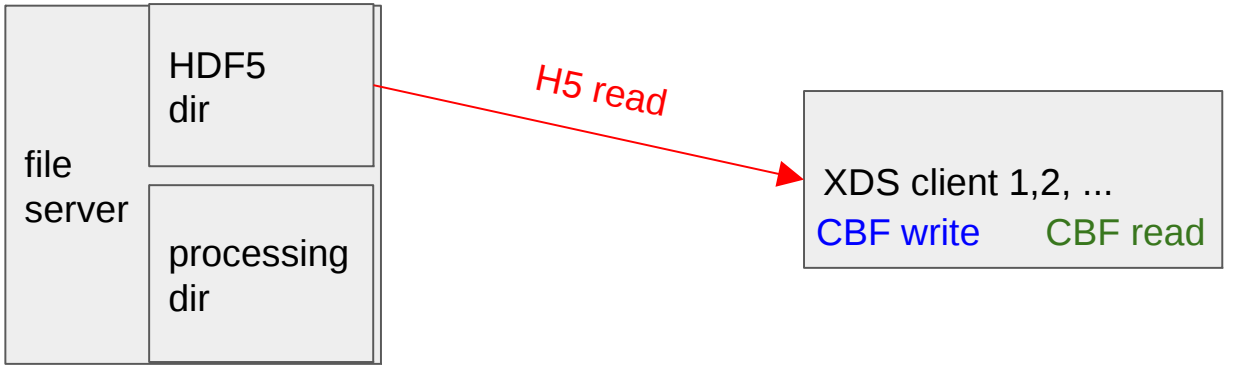
By default, each frame of a dataset traverses the network three times:




```
read  HDF5 from fileserver
write CBF  to   fileserver
read  CBF  from fileserver
```

(for compatibility with multi-client processing, the processing dir cannot be local!)

A 3-line wrapper script (see “Eiger” article) reduces this to one-time traversal:



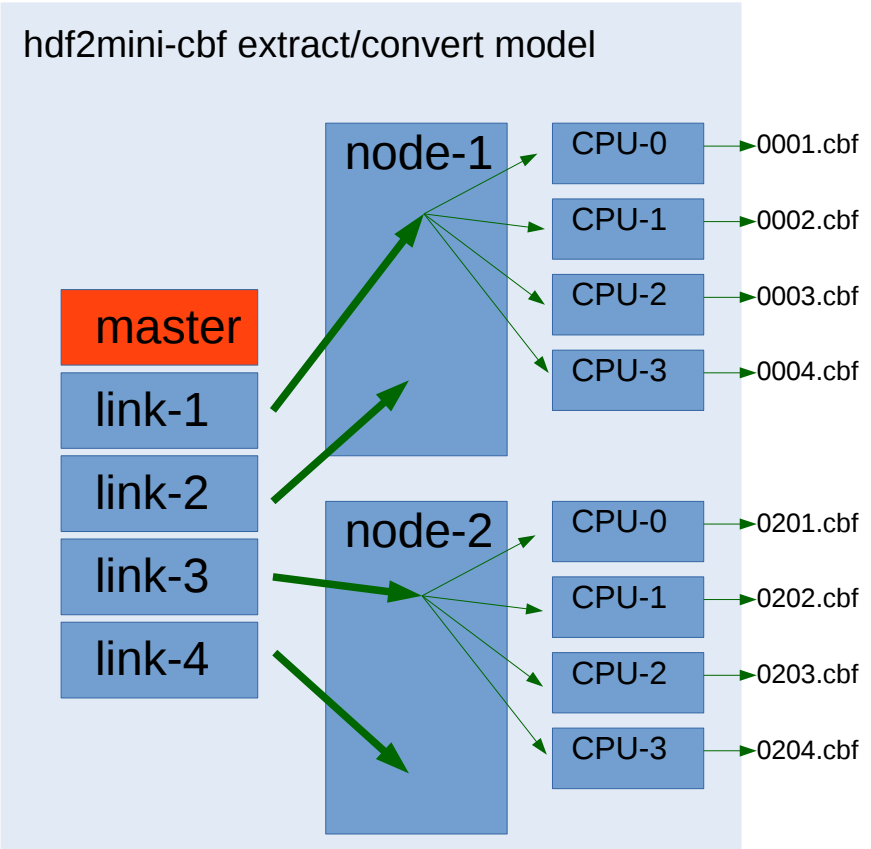

```
read  HDF5 from fileserver
write CBF  to   RAM (/dev/shm)
read  CBF  from RAM
```

this is compatible with multi-client processing, and **reduces overhead to ~24%**

autoPROC converter – *hdf2mini-cbf (1)*

- requested in June 2015 by commercial users of autoPROC:
  - fast conversion of complete datasets to fit into
    - existing data transfer procedures
    - archiving protocols
    - internal processing pipelines (autoPROC)
  - support for OsX
- existing tools at the time didn't provide that

- parallel-1: divide external links across compute nodes
  - sequential: read complete external link into memory
  - parallel-2:
    - apply pixel_mask
    - detect defective pixels not yet masked
    - byte-offset compression
    - writing mini-cbf file



- alternative: *H5ToXds* emulation mode (single-image extraction/conversion)
- public release: autoPROC 20160225

autoPROC converter – *hdf2mini-cbf (2)*

```
USAGE: aP_convert_hdf5 [-h] [-v] [-t] [-compressed] -m <master> -o <prefix>

       -h                    : show help

       -v                    : increase verbosity (default = 0)

       -t                    : don't actually run the command(s), but show
                               what would be done

       -m <master>           : HDF5 *_master.h5 file

       -o <prefix>           : prefix (default=__146215_) for output mini-cbf
                               files (default=__146215__######.cbf)

       -compressed           : (optional) write *.cbf.gz files directly
```

```
1800 16M images (18 external links): 52 sec on 8 nodes at SLS
```

parameters:

```
autoPROC_Hdf2Cbf_NodesFile

autoPROC_Hdf2Cbf_Nodes

autoPROC_Hdf2Cbf_SshCommand
```

wrapper to distribute efficiently among cluster nodes

actual C tool (HDF5 1.8.16):

Bitshuffle - Filter for improving compression of typed binary data.
Copyright (c) 2014 Kiyoshi Masui (kiyo@physics.ubc.ca)
LZ4 - Fast LZ compression algorithm.
Copyright (C) 2011-2012, Yann Collet.
LZ4/HDF5 FILTER IMPLEMENTATION.
Copyright (C) 2011-2013, Dectris Ltd.

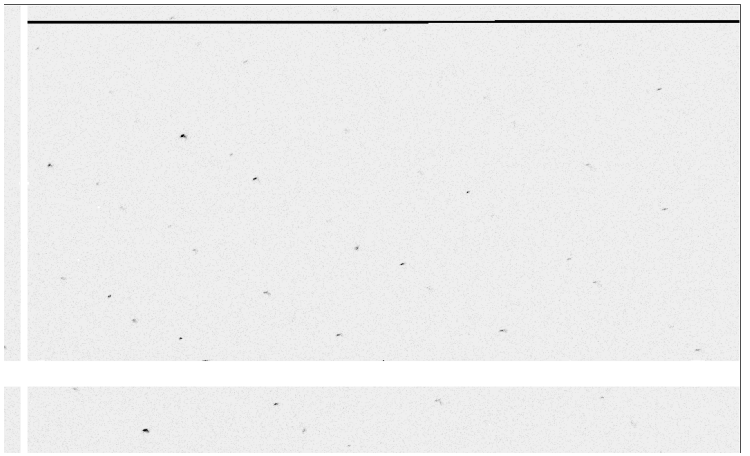
```
USAGE: hdf2mini-cbf [-h] -m <master.h5> [-linkrange <i1> <i2>] [-image <imgnum>] [-o <prefix>] \
       [-uncompressed|-compressed|-gzip-level <level>] [-minpixval <val>] \
       [-noapply_pixel_mask] [-noapply_intmax]
```

Eiger/HDF5 data – content issues

- Varied experiences with HDF5 datasets from Eiger detectors at PX-I/SLS, MASSIF-3/ESRF, LS-CAT/APS, Proxima-2/Soleil, BL1A/PF, BL32XU/Spring-8
- Missing meta-data:
  - Crystal-detector distance
  - Beam centre
  - Omega and oscillation angle
- Incorrect meta-data:
  - Sensor thickness (0.45 m)
- Inconsistent (?) meta data:
  - /entry/instrument/detector/detectorSpecific/nimages = 2400, but 72001 Omega values
  - /entry/data/data_112000 External Link {x7_1_data_112000.h5//entry/data/data} (560000 degree of data)

```
DATASET "/entry/instrument/detector/sensor_thickness"
{
   DATATYPE  H5T_IEEE_F32LE
   DATASPACE  SCALAR
   DATA {
   (0): 0.45
   }
   ATTRIBUTE "units" {
      DATATYPE  H5T_STRING {
         STRSIZE 1;
         STRPAD H5T_STR_NULLTERM;
         CSET H5T_CSET_ASCII;
         CTYPE H5T_C_S1;
      }
      DATASPACE  SCALAR
      DATA {
      (0): "m"
      }
   }
}
```

- Other issues that need to be handled correctly before data-processing:
  - Failing chips (before getting replaced)
  - Failing pixels (before pixel-mask update)
  - “Failed” (marked) lines of pixels (4K pixels, detector safety measure)



What we don't want ...

Synchrotron	Beamline	Date (YYYYMMDD)	Arguments to *process*	Remark
ALBA	XALOC - BL13	20121120		Pilatus 6M
ALS	4.2.2	20110929	`ReversePhi="yes"`	NOIR-1 detector
	5.0.1	20031212	`BeamCentreFrom=header:x,-y autoPROC_TwoThetaAxis="-1 0 0"`	ADSC Q4 (S/N 402)
	5.0.2	20110424	`BeamCentreFrom=header:x,-y`	
		20130111	`BeamCentreFrom=header:y,x`	
	8.2.1	20110908	`BeamCentreFrom=header:x,-y autoPROC_TwoThetaAxis="-1 0 0"`	For older ADSC Q210 detector (S/N 445)
		20140115	`BeamCentreFrom=header:y,x`	For newer ADSC detector (S/N 905)
	8.2.2	20121017	`BeamCentreFrom=header:y,x`	
		20071004	`BeamCentreFrom=header:x,-y`	detector (S/N 905)
	8.3.1	20121219	`BeamCentreFrom=header:x,-y autoPROC_TwoThetaAxis="-1 0 0"` (or: `-M Als831`)	see Note 3 and here
	12.3.1	20121205	`BeamCentreFrom=header:x,-y autoPROC_TwoThetaAxis="-1 0 0"` (or: `-M Als1231`)	see Note 3
APS	14-BM-C	20081014	`ReversePhi=yes`	ADSC QUANTUM 315, S/N 910
	17-ID / IMCA-CAT	20100121	`autoPROC_TwoThetaAxis="-1 0 0"`	
	19-ID	20101210	`ReversePhi=yes beam="1570 1491"`	ADSC QUANTUM 315R; see Note 1
		20110609, 20130727, 20140602	`ReversePhi=yes beam="1577 1496"`	ADSC QUANTUM 315R (S/N 458); see Note 1
		20150209	`ReversePhi=yes beam="1584 1497"`	ADSC QUANTUM 315R (S/N 458); see Note 1
	19-BM	20081210	`ReversePhi=yes beam="1579 1520"`	SBC-3 detector; see Note 1
		20090406	`beam="1052 1003"`	ADSC QUANTUM 210R, binned; see Note 1
	21-ID-D	20130423		MarCCD, S/N 023
		20160310	see here	Eiger 9M
	21-ID-F	20080725	`beam="1521 1526"`	see Note 1
		20130310		MarCCD, S/N 019
	21-ID-G	20130209		MarCCD, S/N 025
	23-ID-B	20081011		
	23-ID-D	20071025		
Australian Synchrotron	24-ID	20081010	`BeamCentreFrom=header:x,-y`	ADSC QUANTUM 315, S/N 911
	MX1	20130909	`BeamCentreFrom=header:y,x ReversePhi=yes`	ADSC QUANTUM 210, S/N 457
	MX2	20130909	`BeamCentreFrom=header:y,x ReversePhi=yes`	ADSC QUANTUM 315 928, S/N; very old images have no date in image header
Bessy	14.1	20121120		MarCCD detector (MX-225)
	14.2	20121120		MarCCD detector (MX-225)
CLSI	08ID-1	20110920		
ESRF	BM-14	20070909	`BeamCentreFrom=header:y,x`	MarCCD detector storing beam-centre in mm instead of standard pixels (automatically adjusted
				storing beam-centre in mm instead of standard pixels (automatically adjusted within autoPROC)
	ID14-1	20070821		ADSC, S/N 444
	ID14-4	20121206	`BeamCentreFrom=header:y,x`	
	ID29		`-M EsrfId29`	
	MASSIF-3 (ID30A-3)	20160307	see here	Eiger 4M
NSLS/BNL	X25	20110319		Some older datasets have distance stored as mm when it should be m
	X26C	20130422	`BeamCentreFrom=header:y,x`	
Photon Factory	BL-1A	20160309	see here	Eiger 4M
	BL-17A	20111209	`BeamCentreFrom=header:y,x`	
SLS	PX-I/X06SA			Pilatus 6M
		20160307	see here	Eiger 16M
	PX-II/X10SA			
	PX-III/X06DA		`-M SlsPXIII`	Pilatus 2M with PRIGo goniometer
Soleil	Proxima1	20110630	`BeamCentreFrom=header:y,x`	ADSC S/N 927
		20111027	`-M SoleilProxima1`	Pilatus 6M detector
	Proxima2	20131004	`BeamCentreFrom=header:y,x`	ADSC S/N 927
		20160314	see here	Eiger 9M
SPring-8	all MX beamlines	20160510	`ReversePhi=yes`	(Thanks to Keitaro Yamashita)
SSRF	BL17U-1	20151229		(Thanks to Qingjun Ma)
SSRL	BL19U-1	20151229	`ReversePhi=yes`	(Thanks to Qingjun Ma)
	all			see Note 2
	BL7-1			see Note 4

… but already (again) (partially) have:
http://www.globalphasing.com/autoproc/wiki/index.cgi?BeamlineSettings

Transferability of processing

- Archiving raw data (see IUCr DDDWG):
  - http://www.iucr.org/resources/data/dddwg

- No use in archiving if re-processing can only be done in the same way it was done originally with the same software

- There is a difference between the intention and the actual experiment.
- It is important to record both:
  - any deviation from the plan tells us something (beam fluctuation, goniostat jitter, synchronisation issues, poor crystal centering, radiation damage, …)
  - it will enable us to improve instruments, beamline control software, crystal handling, experimental design, data analysis ...

- Pre-generated meta-data (image headers) for the sake of speed are problematic 1 year (1 month, 1 week, 1 day) later:
  - no record what actually happened (interleaving, new position on crystal etc) … and no time to write it down into a notebook.

Experiment

- (Initial) plan of experiment
- Execution of experiment – results in data:
  - Detector data: pixel counts, detector events ("masked" pixels), readout timings, energy-dependent countrate cutoff, ...
  - Goniostat data: goniostat settings, rotation axis, speed, crystal-detector distance, translational motor settings, ...
  - Source data: flux, wavelengths, refill/top-up, shutter,...
  - Crystal data: fluorescence scan, pictures, position, overlap with previously exposed positions, interleaving, helical scan, multi-wavelength, ...
  - Project data: compound, sequence, soak conditions, (likely) SG/cell, ...
- Analysis of experiment:
  - **Online** (while crystal is still mounted or available)
  - **Remote** (to plan next construct, cloning, expression, purification, crystallisation, crystal handling, soaking)

X-ray data collection is the last experimental step of the analysis, but it is not a mere technicality and should be treated as an important scientific process. *Dauter, Z. (1999). Acta Cryst. D55, 1703-1717.*

Who is responsible for writing data items?
Who is responsible for checking data?

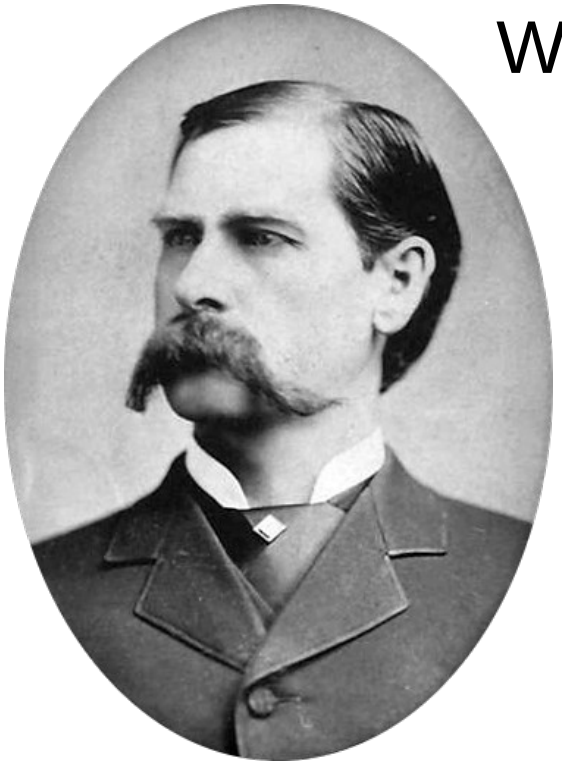
Goal: transferability of processing

Special thanks to:

- Kay Diederichs (Uni Konstanz)
- Wolfgang Kabsch (MPI Heidelberg, DE)
- Andreas Foerster, Stefan Brandstetter (Dectris)
- Meitian Wang, Ezequiel Panepucci, Justyna Wojdyla, Vincent Olieric (SLS)
- Dirk Reinert (Boehringer-Ingelheim)
- Joachim Diez (Expose)
- David von Stetten (ESRF)
- Joe Brunzelle (LS-CAT)
- Martin Savko, Bill Shepard (Soleil)
- Nicholas Keep (Birkbeck, London)
- Claus Flensburg, Gerard Bricogne (Global Phasing Ltd., UK)
- David von Stetten (ESRF)
- Joe Brunzelle (LS-CAT)
- Martin Savko, Bill Shepard (Soleil)
- Nicholas Keep (Birkbeck, London)
- Takaaki Fukumi (Chugai Pharmaceuticals)
- Keitaro Yamashita (Spring-8)
- Kevin Bataille (IMCA-CAT)
- Phil Evans (LMB/MRC, Cambridge, UK)
- CCP4

http://strucbio.biologie.uni-konstanz.de/xdswiki/index.php/Eiger
http://www.globalphasing.com/autoproc/wiki/index.cgi?DataProcessingHdf5
http://www.globalphasing.com/autoproc/wiki/index.cgi?BeamlineSettings

“Fast is fine, but accuracy is final.”



Wyatt Earp (1848-1929)

“If everything seems under control, you're not going fast enough.”

Mario Andretti (1940-)