Lossless compression

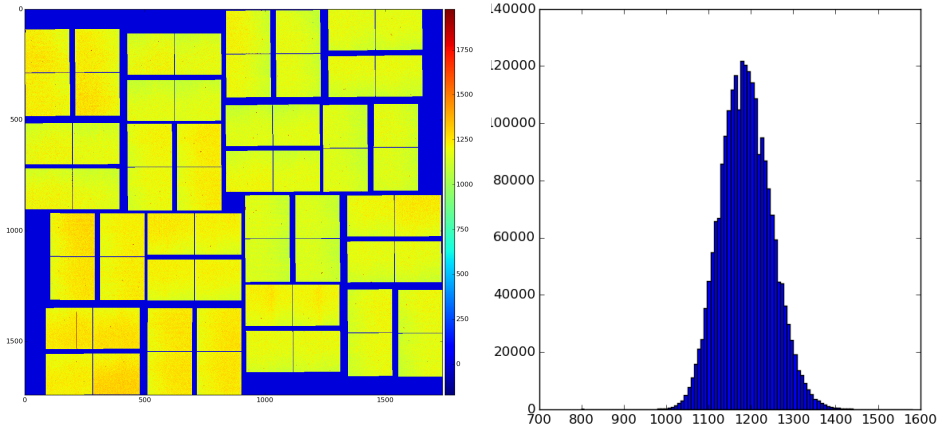
Mikhail Dubrovin

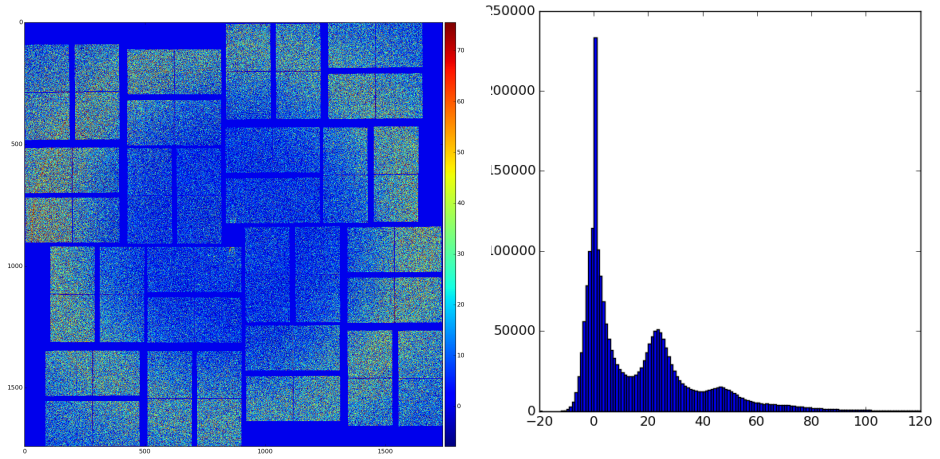
CSPAD data

- shape=(32,185,388)
- size=2296960 pix
- dtype=(u)int16

Raw data entropy H=7.95

Calibrated data H=5.84





Lossless compression - GZIP

- gzip-CLI – most popular and productive file-to-file compressor
  - gzip -c test.xtc > test.xtc.gz
  - compression factor = 1.89, time 3sec/event…
    - not optimal – xtc contains data & metadata
    - slow
- zlib-(gzip)API compression for single CSPAD image:

```
zlib level=0:  data size (bytes) in/out = 4593957/4594663 = 1.000  time(sec)=0.025749  t(decomp)=0.005665
zlib level=1:  data size (bytes) in/out = 4593957/2922633 = 1.572  time(sec)=0.108629  t(decomp)=0.026618
zlib level=2:  data size (bytes) in/out = 4593957/2908156 = 1.580  time(sec)=0.125363  t(decomp)=0.029112
zlib level=3:  data size (bytes) in/out = 4593957/2884917 = 1.592  time(sec)=0.170814  t(decomp)=0.027699
zlib level=4:  data size (bytes) in/out = 4593957/2886850 = 1.591  time(sec)=0.158719  t(decomp)=0.029466
zlib level=5:  data size (bytes) in/out = 4593957/2885665 = 1.592  time(sec)=0.261296  t(decomp)=0.030550
zlib level=6:  data size (bytes) in/out = 4593957/2834066 = 1.621  time(sec)=0.597133  t(decomp)=0.027355
zlib level=7:  data size (bytes) in/out = 4593957/2828951 = 1.624  time(sec)=0.609569  t(decomp)=0.026842
zlib level=8:  data size (bytes) in/out = 4593957/2828951 = 1.624  time(sec)=0.636173  t(decomp)=0.027226
zlib level=9:  data size (bytes) in/out = 4593957/2828951 = 1.624  time(sec)=0.611562  t(decomp)=0.027042
```

  - compression time rises with level, but factor almost flat

Lossless compressors in HDF5

GZIP and LZF compression in HDF5

- gzip default compression_opts level=4
- input size=4594000(byte)
- time includes saving in file

raw: gzip t=0.216324(sec) ratio=1.583 shuffle=False
raw: gzip t=0.146706(sec) ratio=1.958 shuffle=True

calib: gzip t=0.168040(sec) ratio=2.072 shuffle=False
calib: gzip t=0.182965(sec) ratio=2.187 shuffle=True

raw: lzf t=0.108339(sec) ratio=1.045 shuffle=False
raw: lzf t=0.075530(sec) ratio=1.698 shuffle= True

calib: lzf t=0.100822(sec) ratio=1.351 shuffle=False
calib: lzf t=0.086916(sec) ratio=1.473 shuffle= True

Compression filters szip, lzo, blosc, bzip2 are unavailable in our inst of HDF5

Lossless compressors for LCLS data

Igor Gaponenko - compressor for LCLS detector int16 data:

- estimates dataset spread,
- use 16- and 8-bit words to save data with positions coded in metadata
- Features
  - Optimized to work with 16-bit detector data only (not with xtc or hdf5 files containing metadata).
  - By design compression factor ≤2.
  - Single array of data is split and processed in multi-threads (inside compression algorithm).
  - Igor's statement: up to ~two order of magnitude faster than gzip.
  - Igor thinks that further specialization of data (separation of signal and background regions between threads) may improve compression factor.

Matt Weaver – compressors Hist16 & HistN

- Available in package pdsdata/compress
- Hist16 - the same as Igor's compressor, but does not use multi-threading - slower than Igor's
- HistN – uses 16-bit and 8,7,6...-bit words, compression factor HistN upto ~2.