Using the HDF5 Translator

LCLS Users Workshop

David Schneider
davidsch@slac.stanford.edu
Oct 8, 2014

U.S. DEPARTMENT OF ENERGY
Office of Science

SLAC NATIONAL ACCELERATOR LABORATORY

Outline

HDF5 Files

- Datasets vs. Events
- Dataset Alignment
- Hierarchy, where the data goes
- Translation speed
- Filtering translation
  - Filtering types or sources
  - Dynamically filtering events with a Psana Module
- Adding user data to the translation
- Split scan translator

Reference:
https://confluence.slac.stanford.edu/display/PSDM/The+XTC-to-HDF5+Translator

Data File Fromat: XTC vs. HDF5

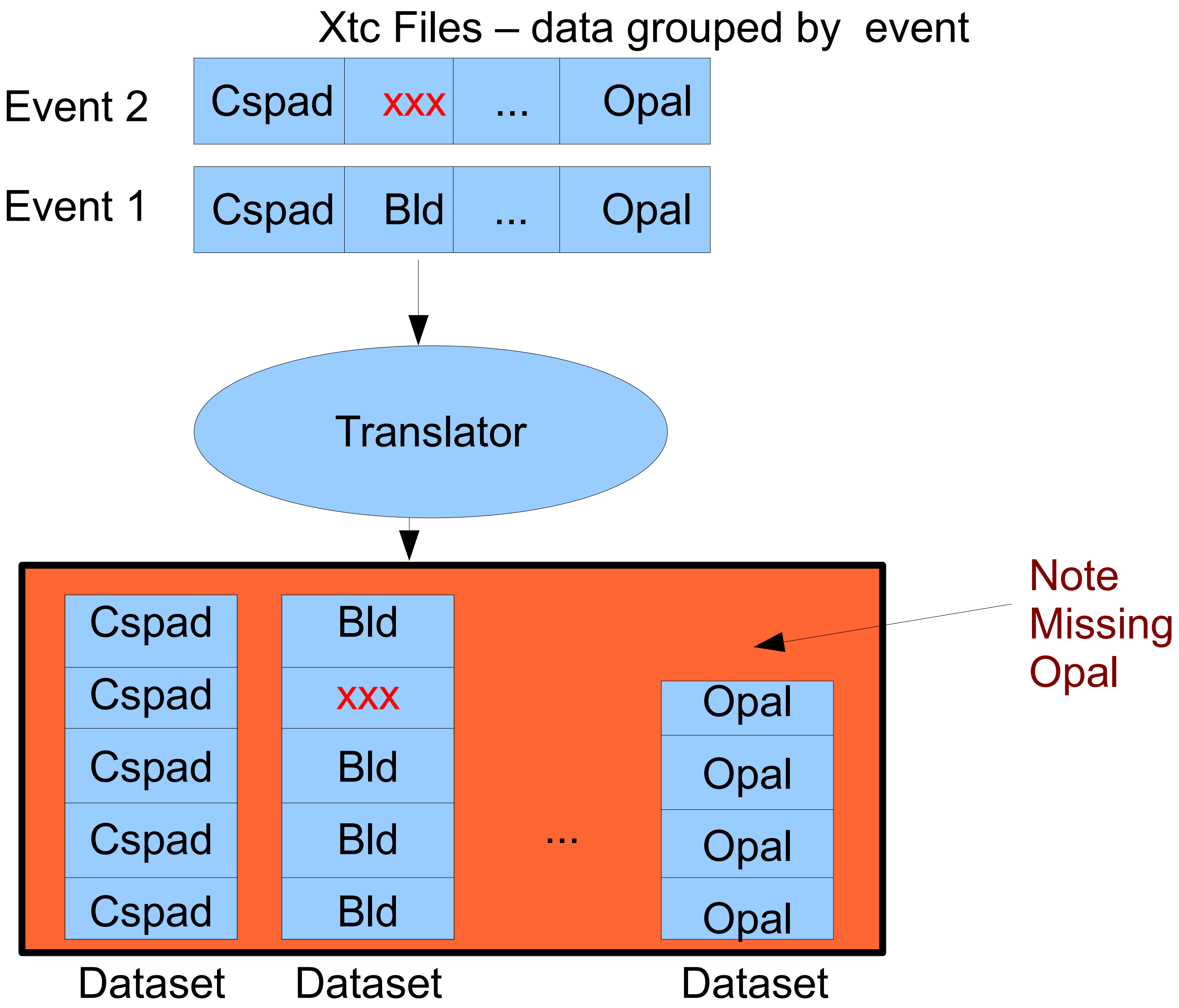
XTC Files:

- Format developed at SLAC
- Native format for experiment files
- Requires use of LCLS software to analyze (Psana/AMI)
    - Or write your own xtc parser using the C++ API

HDF5 Files:

- General scientific format
- Many software choices for analysis
    - Matlab, IDL, Python libraries (h5py, tables)

Translation Process



Xtc Files – data grouped by event

Event 2 Cspad xxx ... Opal

Event 1 Cspad Bld ... Opal

Translator

Cspad Cspad Cspad Cspad Cspad

Bld xxx Bld Bld Bld

...

Opal Opal Opal Opal

Dataset Dataset Dataset

Note Missing Opal

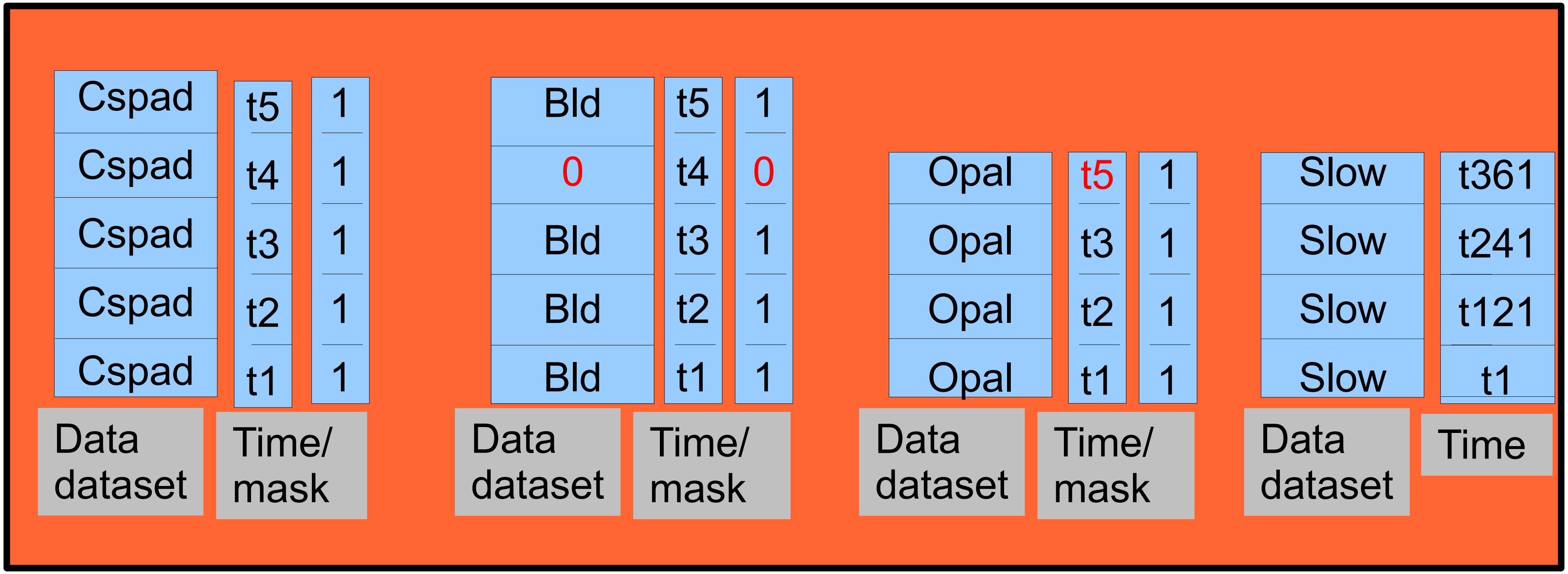
Hdf5 File

data grouped
By type into
datasets

HDF5 Dataset Alignment

- Use **time** datasets to “Event build” from the HDF5 file
  time contains seconds, nanoseconds

- Check **mask** datasets to identify undamaged data



Cspad	t5	1	Bld	t5	1					
Cspad	t4	1	0	t4	0	Opal	t5	1	Slow	t361
Cspad	t3	1	Bld	t3	1	Opal	t3	1	Slow	t241
Cspad	t2	1	Bld	t2	1	Opal	t2	1	Slow	t121
Cspad	t1	1	Bld	t1	1	Opal	t1	1	Slow	t1
Data dataset	Time/ mask		Data dataset	Time/ mask		Data dataset	Time/ mask		Data dataset	Time

Navigating the HDF5 Group Hierarchy

- /Configure:0000/Run:0000/CalibCycle:0000/{type}/{source}
- Examples of {type}/{source}
  - Ipimb::DataV2/XppSb2_Ipm
  - CsPad::ElementV2/CxiDs1.0:Cspad.0
- Epics
  - uses three groups: {type}/{source}/{pvname}
  - Epics::EpicsPv/EpicArch.0:NoDevice.0/BEAM:LCLS:ELEC:Q
- Configuration Data
  - /Configure:0000/{configType}/{source}
- Calibration constants
  - /CalibStore/{type}/{source}
  - Example   CsPad::CalibV1/CxiDsd.0:Cspad.0

Exploring HDF5 Files

- HDF5 Group tools
  - hdfview
  - h5ls -r
  - h5dump
- Psana
  - Does event building
  - High level Python objects
- h5py (Python)
  - datasets are numpy arrays

Translation speeds

- High data rates: ~ 1GB/sec
  - Cxi with 2 cspad, opal camera, acqiris
- Translation averages 50MB/sec
- Profiling translation shows:
  - Compression: up to 70% of the time
  - Calibrating cspad: significant
- Faster?
  - Turn off compression (will write more to disk)
  - Filter whole events
  - Filter data in the event
  - Parallel translation - split scan – seeing 180MB/sec

Turning off Compression and Filtering

Translator is a psana module. Reads configuration file.

Write file translate.cfg:

```
[Translator.H5Output]
deflate = -1
Cspad =False
type_filter=exclude Cspad Frame
src_filter=exclude DetInfo(CxiDs1.0:Cspad.0)
```

How do I know what source to use?

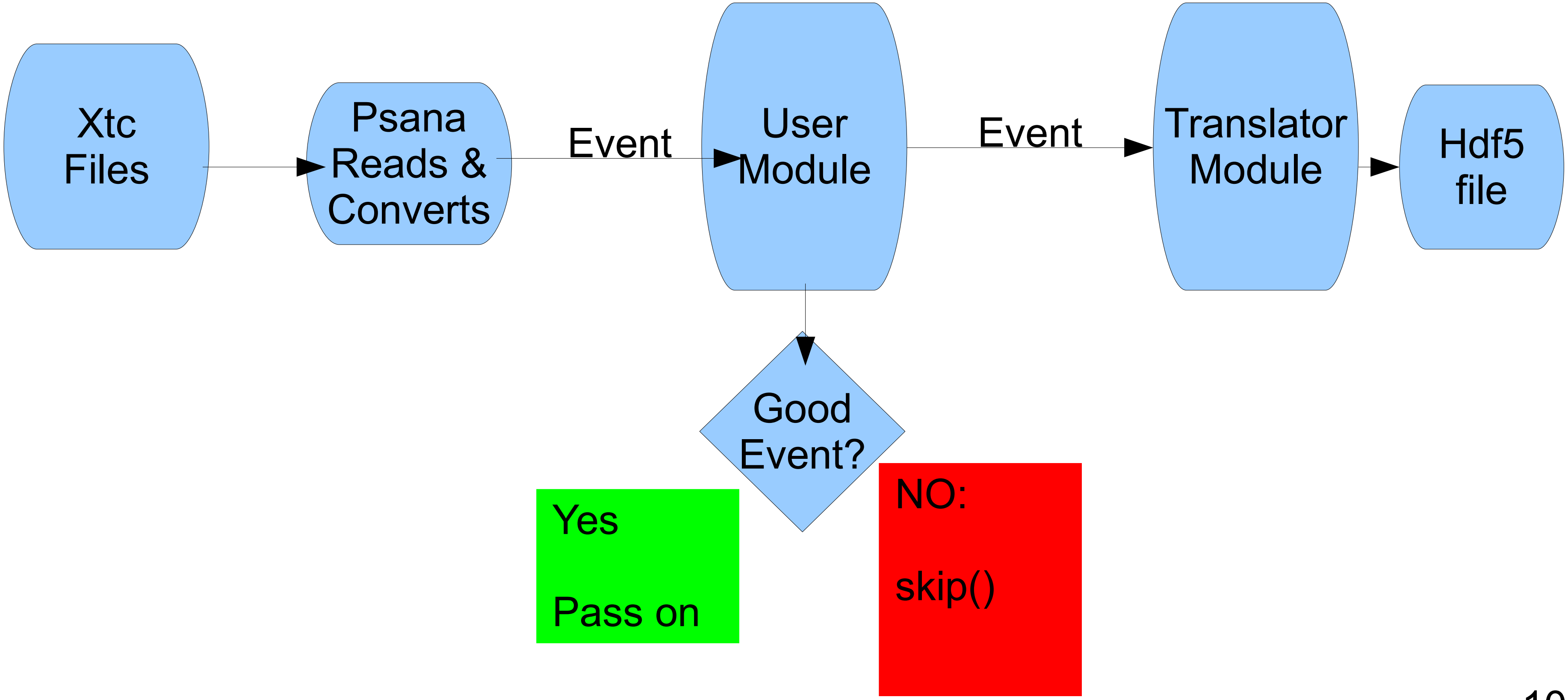
```
psana -m EventKeys -n 1 exp=xpptut13:run=179
```

shows all sources in configuration and first event

Now run translator

(see reference:
XTC-to-HDF5 Translator)

Dynamic Filtering with Psana Modules



Event Filtering Example – config file

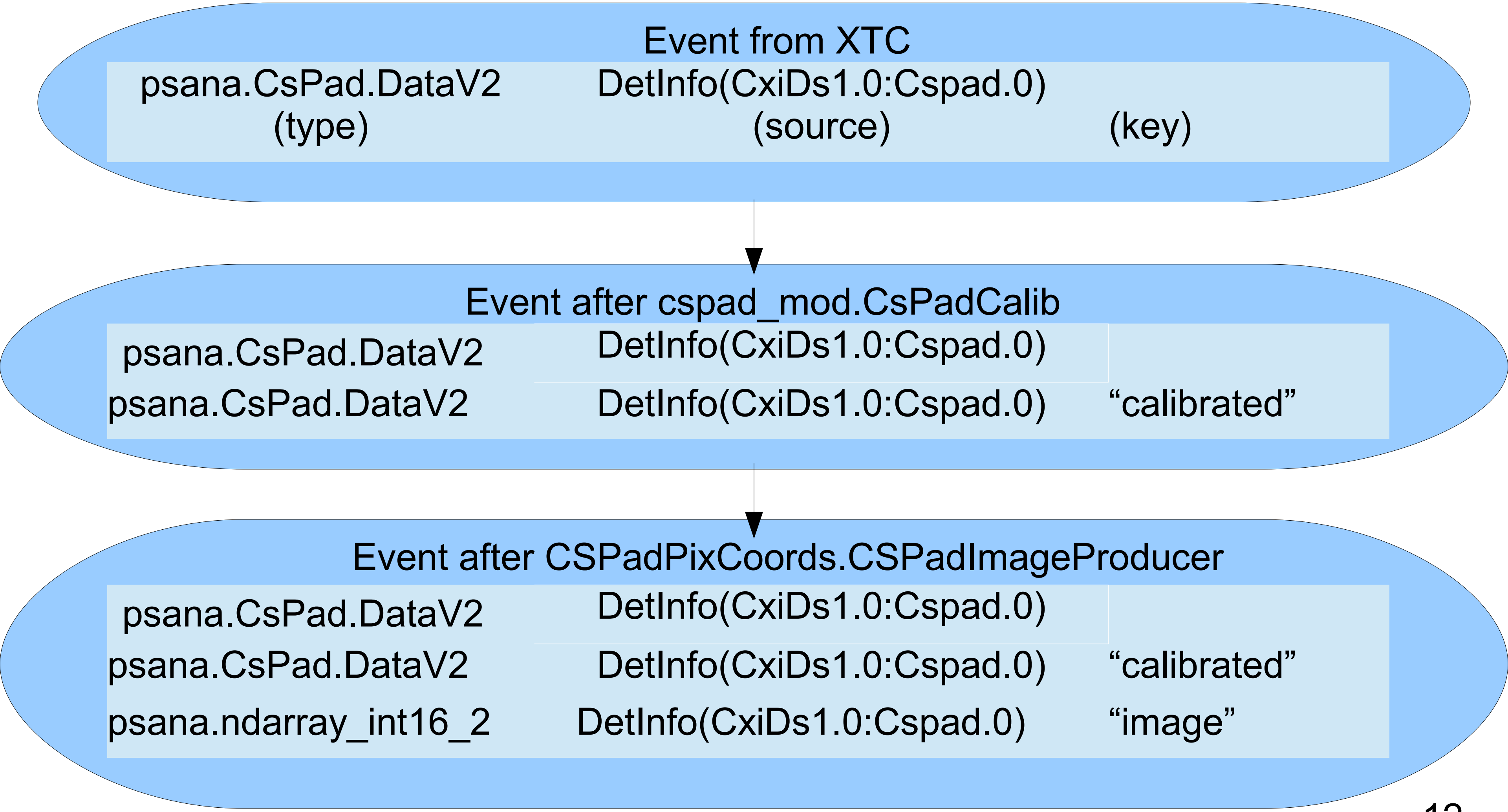
Go over example from XTC to HDF5 Translator page on confluence
psana configuration file: trans.cfg:

```
[psana]
modules = cspad_mod.CsPadCalib \
          CSPadPixCoords.CSPadImageProducer \
          mypkg.mymod \
          Translator.H5Output
files = exp=cxitut13:run=1150

[CSPadPixCoords.CSPadImageProducer]
source        = DetInfo(CxiDs1.0:Cspad.0)
key           = calibrated
imgkey        = image

[Translator.H5Output]
Cspad=exclude
deflate=-1
output_file=cxitut13-run1150-filt.h5
```

4 long
Psana module
chain

Next slide,
see how modules
change event

Psana Modules and the Event



Event from XTC

psana.CsPad.DataV2	DetInfo(CxiDs1.0:Cspad.0)	
(type)	(source)	(key)

Event after cspad_mod.CsPadCalib

psana.CsPad.DataV2	DetInfo(CxiDs1.0:Cspad.0)	
psana.CsPad.DataV2	DetInfo(CxiDs1.0:Cspad.0)	“calibrated”

Event after CSPadPixCoords.CSPadImageProducer

psana.CsPad.DataV2	DetInfo(CxiDs1.0:Cspad.0)	
psana.CsPad.DataV2	DetInfo(CxiDs1.0:Cspad.0)	“calibrated”
psana.ndarray_int16_2	DetInfo(CxiDs1.0:Cspad.0)	“image”

Event Filtering Example – user module

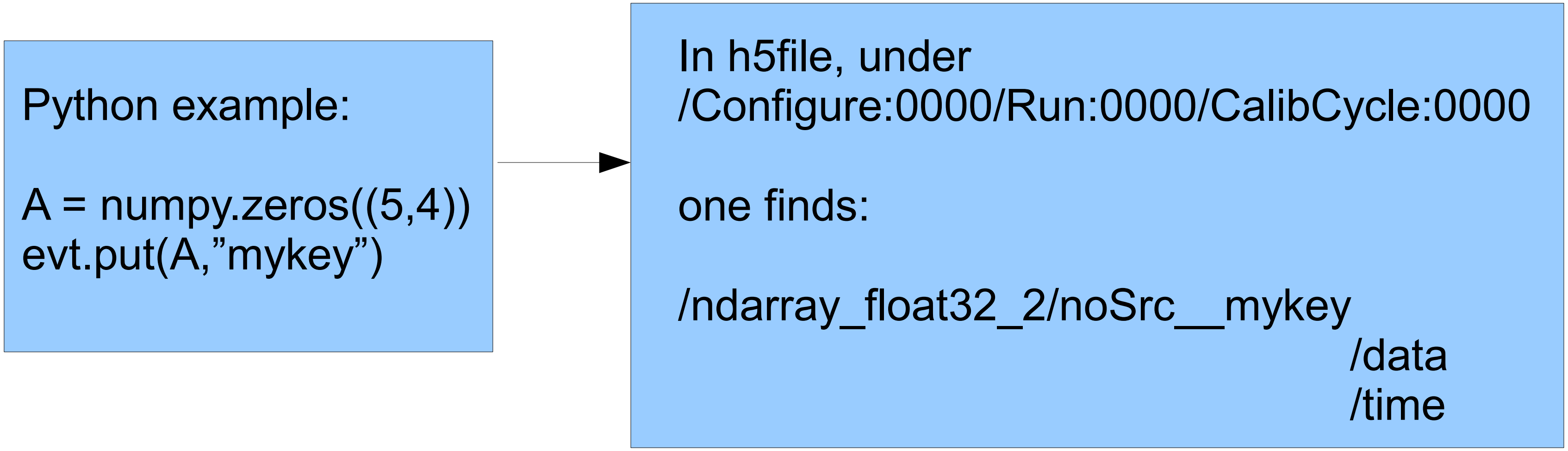
Psana Python module, mypkg/src/mymod.py:

note, filename and class name are same

```
import psana

class mymod(object):
    def __init__(self):
        self.threshold = self.configInt('threshold',int(1e9))
        self.source = self.configSrc('source','DetInfo(CxiDs1.0:Cspad.0)')

    def event(self, evt, env):
        image = evt.get(psana.ndarray_int16_2,
                        self.source,
                        'image')
        if image is None: return
        count = image[:].sum()
        if count < self.threshold:
            self.skip()
```

Adding Data to the Translation

- You can translate arrays.
- From Python, add numpy array
- From C++ add ndarray
- Dimensions 1 to 4
- 8,16,32,64 bit signed/unsigned ints, float and double



Python example:

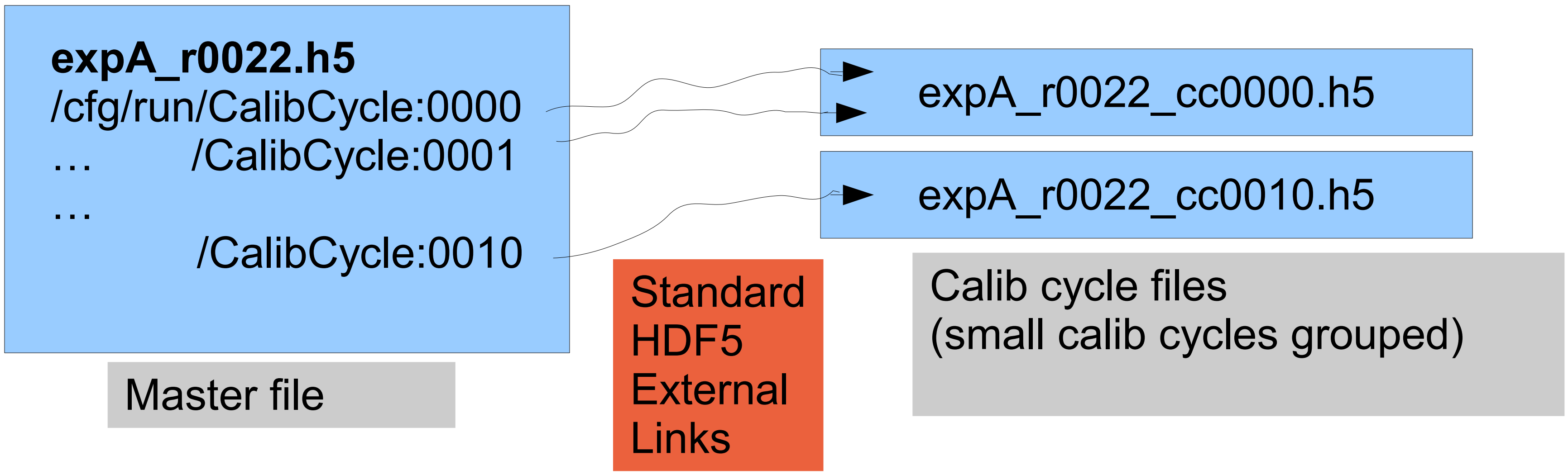
```
A = numpy.zeros((5,4))
evt.put(A,”mykey”)
```

In h5file, under
/Configure:0000/Run:0000/CalibCycle:0000

one finds:

```
/ndarray_float32_2/noSrc__mykey
                               /data
                               /time
```

Split Scan Translation



expA_r0022.h5
/cfg/run/CalibCycle:0000
… /CalibCycle:0001
…
/CalibCycle:0010

Master file

Standard HDF5 External Links

expA_r0022_cc0000.h5

expA_r0022_cc0010.h5

Calib cycle files
(small calib cycles grouped)

Reasons for Split Scan

- Faster due to parallel translation of calib cycles
- Manage several smaller files vs. one large HDF5 File
- Limited support to read hdf5 during Translation

Note: only applicable to experiments with multiple calib cycles