

Experiences with Virtual Datasets

Thomas Kluyver, European XFEL

HDF5 User Meeting 2021

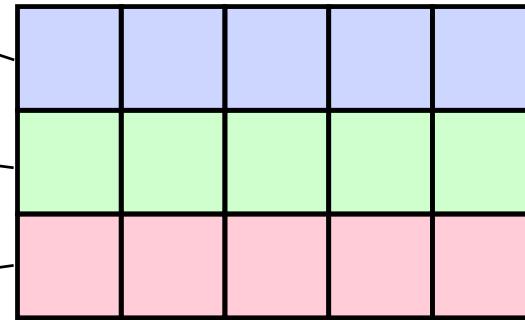


European
XFEL

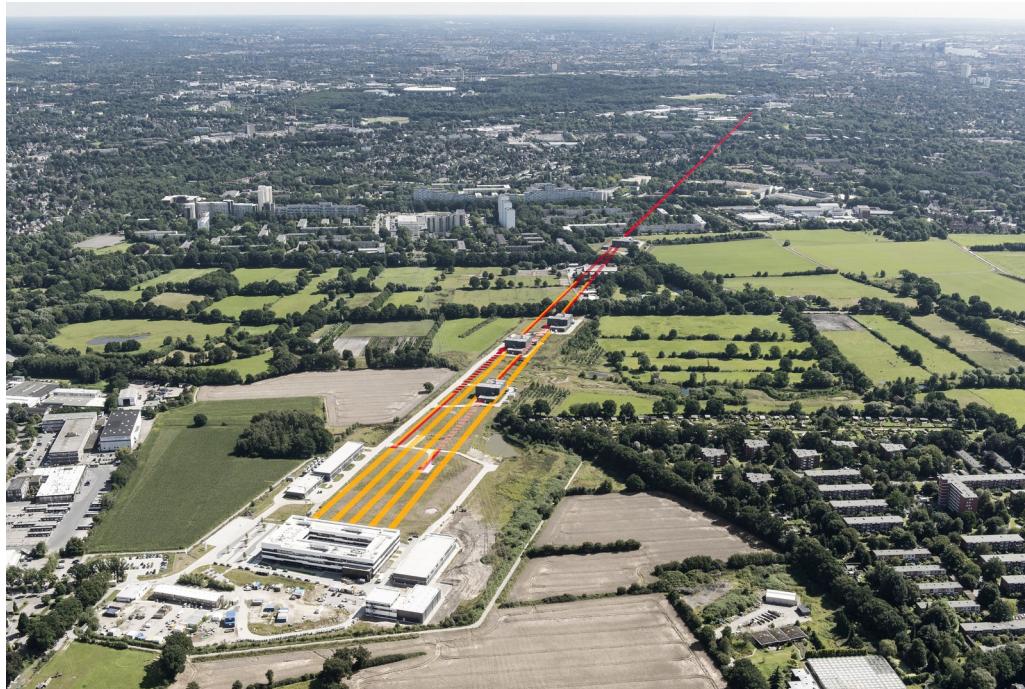
Source datasets



Virtual dataset

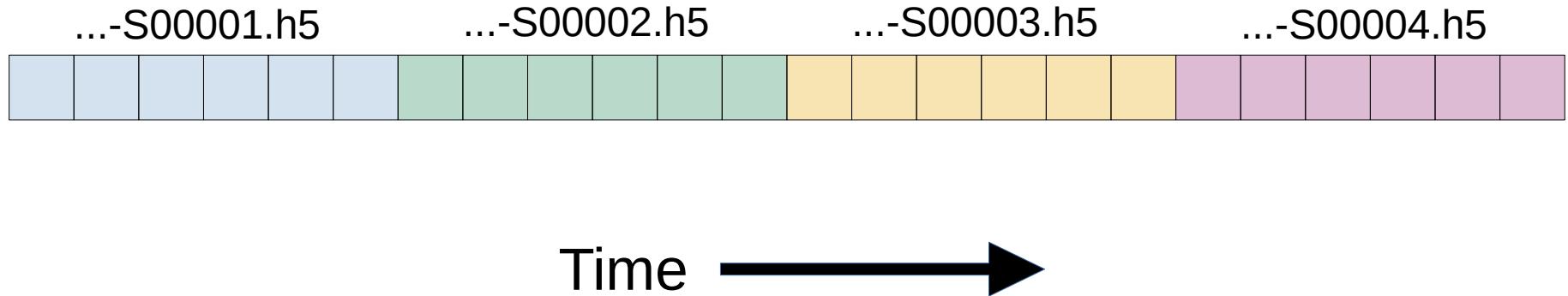


European XFEL

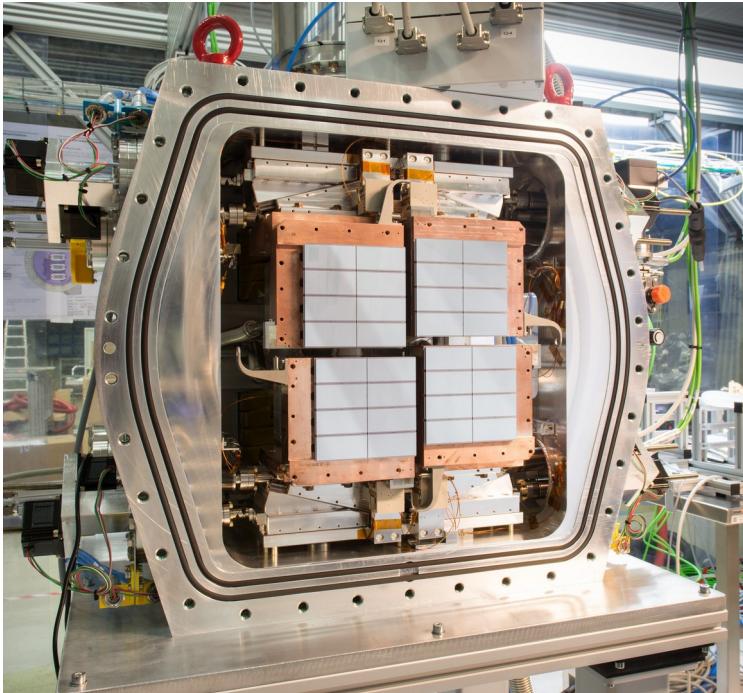


European
XFEL

Sequence files

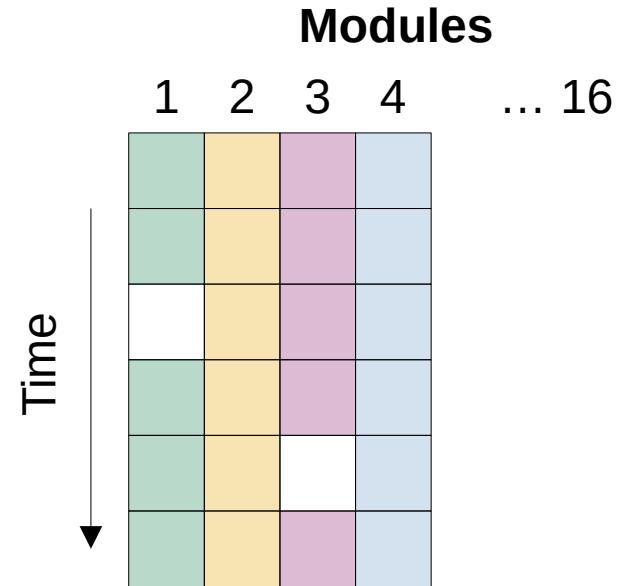


Multi-module detectors



DSSC detector

Image © DESY / Karsten Hansen



Making a VDS: Python code

```
import h5py

f = h5py.File('vds.h5', 'w')

with f.build_virtual_dataset(
        'VDS', shape=(4, 6), dtype='i4', fillvalue=-1
) as layout:

    for i, name in enumerate(['A', 'B', 'C']):
        layout[i] = h5py.VirtualSource(f'{name}.h5', name, (6,))
```

Making a VDS: C code

```
space = H5Screate_simple(RANK2, vdsdims, NULL);

/* Set VDS creation property. */
dcpl = H5Pcreate(H5P_DATASET_CREATE);
status = H5Pset_fill_value(dcpl, H5T_NATIVE_INT, &fill_value);

/* Initialize hyperslab values. */
start[0] = 0;
start[1] = 0;
count[0] = 1;
count[1] = 1;
block[0] = 1;
block[1] = VDSDIM1;

/*
 * Build the mappings.
 * Selections in the source datasets are H5S_ALL.
 * In the virtual dataset we select the first, the second and the third rows
 * and map each row to the data in the corresponding source dataset.
 */
src_space = H5Screate_simple(RANK1, dims, NULL);
for (i = 0; i < 3; i++) {
    start[0] = (hsize_t)i;
    /* Select i-th row in the virtual dataset; selection in the source datasets is the same.
*/
    status = H5Sselect_hyperslab(space, H5S_SELECT_SET, start, NULL, count, block);
    status = H5Pset_virtual(dcpl, space, SRC_FILE[i], SRC_DATASET[i], src_space);
}

/* Create a virtual dataset. */
dset = H5Dcreate2(file, DATASET, H5T_NATIVE_INT, space, H5P_DEFAULT, dcpl, H5P_DEFAULT);
```

From Dataset objects

```
import h5py

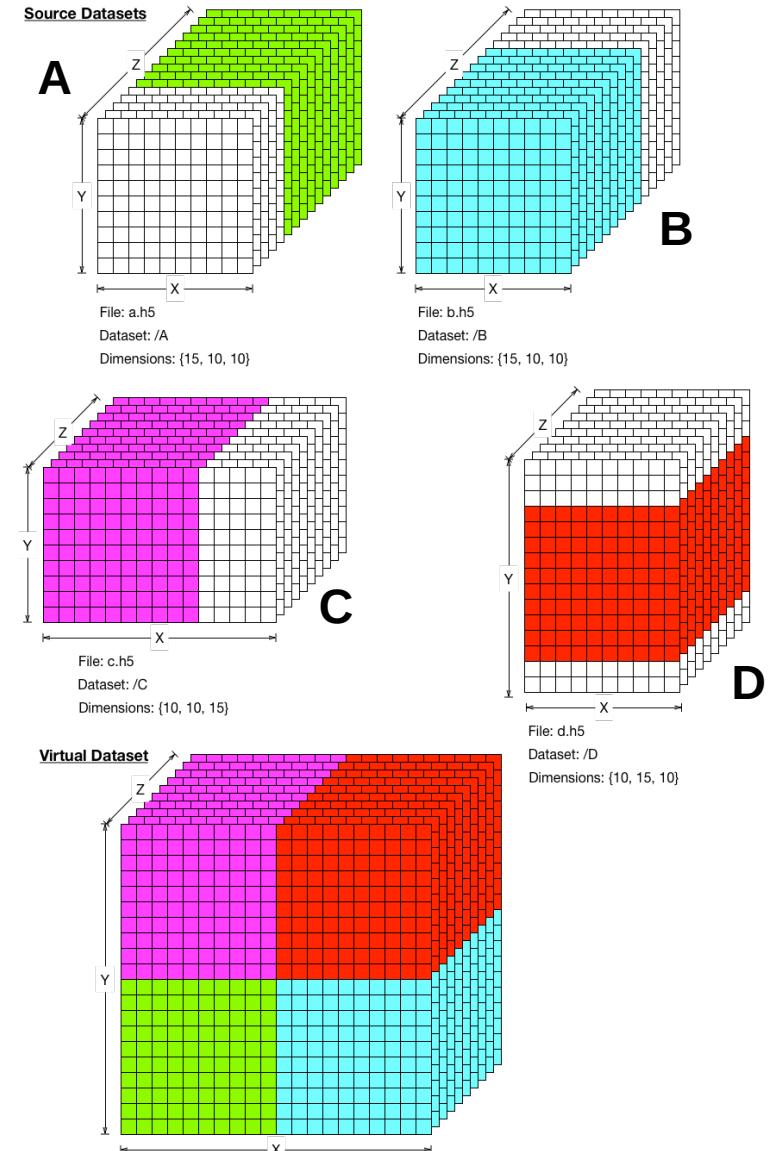
f = h5py.File('vds.h5', 'w')

with f.build_virtual_dataset(
        'VDS', shape=(4, 6), dtype='i4', fillvalue=-1
) as layout:

    for i, name in enumerate(['A', 'B', 'C']):
        with h5py.File(f'{name}.h5') as src_file:
            layout[i] = h5py.VirtualSource(src_file[name])
```

Slicing

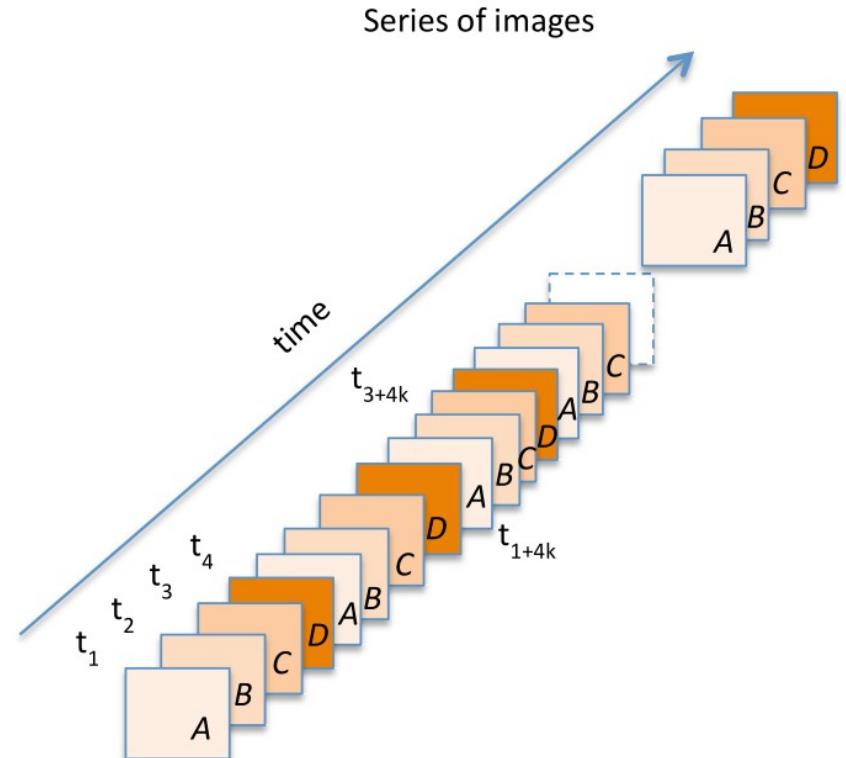
```
layout[:, :10, :10] = src_a[-10:]
layout[:, :10, 10:] = src_b[10]
layout[:, 10:, :10] = src_c[:, :, 10]
layout[:, 10:, 10:] = src_d[:, 2:12]
```



Illustrations from HDF Group's
RFC for virtual datasets

Steps & interleaving

```
layout[0::4] = src_a  
layout[1::4] = src_b  
layout[2::4] = src_c  
layout[3::4] = src_d
```



Illustrations from HDF Group's
[RFC for virtual datasets](#)

Issues

- Errors → empty data
 - Files not found
 - Permission problems
 - Read-write file pointing to read-only source files

hdf5-vds-check

Command-line tool

Check if sources are available

Clear information if not



pip install hdf5_vds_check



GitHub:
[European-XFEL/hdf5-vds-check](https://github.com/European-XFEL/hdf5-vds-check)

```
$ hdf5-vds-check vds.h5
Found 1 virtual datasets to check.
Checking virtual dataset: VDS
  3/3 sources accessible

All virtual data sources accessible
$ rm C.h5
$ hdf5-vds-check vds.h5
Found 1 virtual datasets to check.
Checking virtual dataset: VDS
C.h5:
  [Errno 2] No such file or directory: 'C.h5'
  2/3 sources accessible

ERROR: Access problems for virtual data sources
```

Other issues

- Documentation
- System open file limit
- Performance uncertainty