HDF5 Roadmap 2019-2020
Summer ESIP 2019

Elena Pourmal
The HDF Group EED2 Team Lead
epourmal@hdfgroup.org
Outline

• New features in Hierarchical Data Format 5 (HDF5) Release 1.12.0
• Support for HDF5 versions 1.8 and 1.10
• Summary of HDF5 Roadmap
HDF5 References

HDF5 dataset may store references to other objects or references to selected elements of a dataset. References to objects or dataset regions stored in other files are coming in HDF5 1.12.0.

Dataset in B.h5 stores references to datasets stored in A.h5
HDF5 Reference to Attribute

HDF5 1.12.0 will allow to reference an attribute of a dataset or a group. Dataset in A.h5 stores references to the attributes in A.h5 and B.h5.
HDF5 New Architecture

- HDF5 1.12.0 enables HDF5 to store data on new storage using Virtual Object Layer (VOL) connectors and Virtual File Drivers (VFDs).
- VOL connectors developed by The HDF Group and its collaborators are available from public repository https://bitbucket.hdfgroup.org/projects/HDF5VOL
HDF5 New Architecture (cont’d)
S3 VFD

• HDF5 1.12.0 will have a VFD to access HDF5 file via Amazon Simple Storage Service (Amazon S3)
• Requires minimum changes to the application code
• h5dump and h5ls tools have a flag to specify the driver to access HDF5 file on S3
  h5ls --vfd=ros3
  https://s3.us-east-2.amazonaws.com/file.h5
S3 VFD (cont’d)

• Uses “range get” commands to get “bytes” from HDF5 file stored on S3

• New API to set up S3 VFD

```c
herr_t H5Pset_fapl_ros3(hid_t fapl_id, H5FD_ros3_fapl_t *fa)
```

• Credentials are passed via parameter to the function

• Demo
HDFS VFD

- HDF5 1.12.0 will have a VFD to access HDF5 file on Hadoop Distributed File System (HDFS).
- New API to access HDF5 file on HDFS
  
  ```c
  herr_t H5Pset_fapl_hdfs(hid_t fapl_id);
  ```

- HDF5 command line tools with enabled HDFS VFD allows to extract metadata and raw data from HDF5 and netCDF4 files on HDFS, and use Hadoop streaming to collect data from multiple HDF5 files.

- Demo
New default for string encoding

• UTF-8 will be default string encoding instead of ASCII starting with HDF5 1.12.0
  – Names of groups, datasets, attributes
  – Names of compound datatypes fields, enums, names of the files that are stored in HDF5 according to the File Format spec (VDS, VFDs family and split) have to be ASCII.
Support for HDF5 1.8 and 1.10

• The HDF Group will continue support for HDF5 1.8.* and 1.10.* releases
  – Security patches
  – Critical bug fixes
  – Performance improvements
  – Thorough testing of HDF5 1.8.* for forward compatibility with HDF5 1.10.* and 1.12.0

• Files created by applications that are built with the HDF5 version later than 1.8.* and that do not use any new features of that version are readable by 1.8.*
### HDF5 Roadmap

#### HDF5 1.8
- **Q1 2019**
- **Fall 2019**
  - Security patches
  - Forward Compatibility with 1.12
- **Today**
- **Summer 2020**
  - Security patches
  - Forward Compatibility with 1.12

#### HDF5 1.10
- **Nov 1.10.6**
- **May 1.10.7**
- **Nov 1.10.8**
- **May 1.10.9**
- **Address degradation of performance between major HDF5 releases 1.6 – 1.10**
- **VDS HDF5 performance**
- **Public performance benchmark test suit**
- **Dynamically loaded VFDs**
- **S3, HDFS, Spark VFD connectors**

#### HDF5 1.12
- **July 2019**
  - VOL architecture
  - References
  - Parallel compression (enhanced)
- **Nov 2019**
  - VOL plugins (DAOS)
  - VFD plugins
  - Dynamically loaded VFDs
  - S3, HDFS, SPARK VFD connectors
- **2020-2021**
  - Query and Indexing
  - Full SWMR
  - Mirror VFD
  - Provenance (Onion) VFD

---

HDF5 2.0 instead of 1.12.0?
Storage beyond the current File Systems and HDF5 file format
This work was supported by NASA/GSFC under Raytheon Co. contract number NNG15HZ39C.