Introduction to HDF5

https://github.com/ResearchComputing/Final_Tutorials

February 18, 2016

Timothy Brown

Research Computing
UNIVERSITY OF COLORADO **BOULDER**

Overview

What is HDF5

Data Model

Datasets

Tools

IO Sequence

Example

Overview

What is HDF5

Data Model

Datasets

Tools

IO Sequence

Example

What is HDF5

Hierarchical **D**ata **F**ormat version 5 (HDF5).

- Designed for scientific, high volume data.
- Is a file format to manage data.
  - multidimensional arrays
  - tables
  - compounded structures
  - images
- Software library and tools that provide access to manage data in these files.
- Gives the developer access to manipulate groups and datasets rather than binary streams.

Why Use HDF5

Have you ever asked yourself

- How to handle petabytes of data?
- How to access your data?
  - across a cluster
  - remotely
  - on different platforms
- Majority of granting agencies require:
  - data management plan
  - quality assurance plan
  - open access to the data

Overview

What is HDF5

Data Model

Datasets

Tools

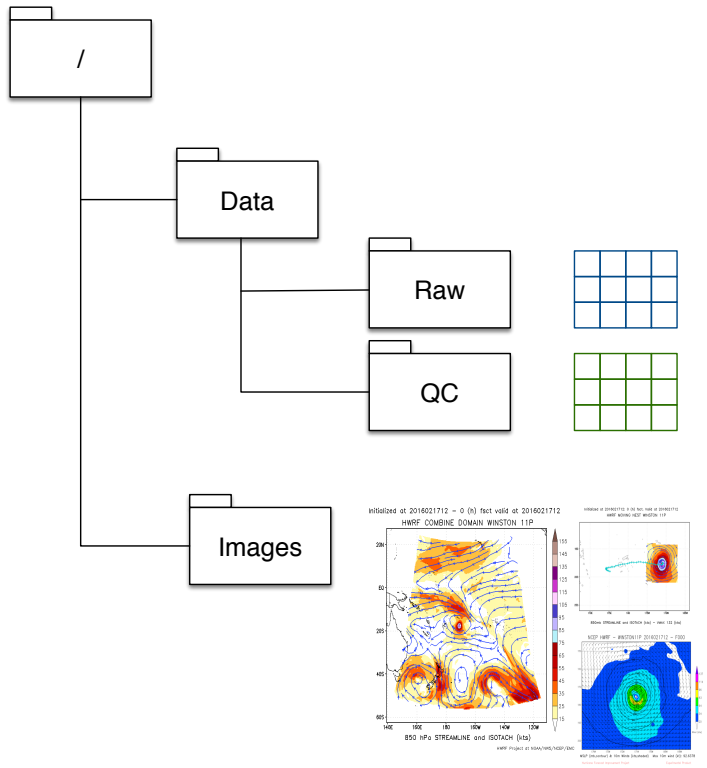
IO Sequence

Example

HDF5 Data Model

A HDF5 file is a container that can have groups, links and datasets.

- File - a contiguous string of bytes in a computer store (memory, disk, etc.), and the bytes represent zero or more objects of the model.
- Group - a collection of objects (including groups).
- Link - the way objects are connected.
- Dataset - a multi-dimensional array of data elements with attributes.
- Image - a dataset that are intended to be interpreted as an image.



- Table - a compounded data type to represent a table.
- Dataspace - a description of the dimensions of the dataset.
- Datatype - a description of a specific class of data element including its storage layout.
- Attribute - a named data value associated with a group, dataset, or named datatype.
- Property List - a collection of parameters (some permanent and some transient) controlling options in the library.

Overview

What is HDF5

Data Model

Datasets

Tools

IO Sequence

Example

HDF5 Datasets

HDF5 Datasets organize and contain your data. They consist of:

Metadata

- datatype (real, integer, …)
- layout (rank, rows, columns)
- properties (units)

Data

```
HDF5 "MIELLAJOKKA.h5" {
GROUP "/" {
   GROUP "010708-MIELLANJOKKA-1-3D" {
      DATASET "Emission" {
         DATATYPE  H5T_IEEE_F64LE
         DATASPACE  SIMPLE { ( 636 ) / ( 636 ) }
         DATA {
         (0): 240, 240.5, 241, 241.5, 242, 242.5, 243, 243.5,
         ...
         (630): 555, 555.5, 556, 556.5, 557, 557.5
         }
         ATTRIBUTE "Units" {
            DATATYPE  H5T_STRING {
               STRSIZE 2;
               STRPAD H5T_STR_NULLTERM;
               CSET H5T_CSET_ASCII;
               CTYPE H5T_C_S1;
            }
            DATASPACE  SCALAR
            DATA {
            (0): "nm"
            }
         }
      }
```

Attributes

Attributes are something you attach to a dataset that provides extra information.

- Describes the intended use of the dataset or group.
- User defined.
- Optional.

For example the location of a reading or the temperature when the reading occurred.

Virtual File Layers

HDF5 provides a virtual file layer which you can extend.

- POSIX
- STDIO
- MPI-IO

You do not need to be an MPI expert to use the parallel IO layer in HDF5.

Overview

What is HDF5

Data Model

Datasets

Tools

IO Sequence

Example

HDF5 on Janus

On Janus since the modules are hierarchical, we need to load the prerequisites.

- Serial

```
login04 ~$ ml intel
login04 ~$ ml hdf5
```

- Parallel (MPI-IO)

```
login04 ~$ ml intel
login04 ~$ ml impi
login04 ~$ ml hdf5
```

HDF5 Tools I

Which provides the following programs

Command	Description
`h5cc`	Simplifies compiling C programs (`h5pcc`)
`h5fc`	Simplifies compiling Fortran programs (`h5pfc`)
`h5debug`	Debugs an existing HDF5 file at a low level
`h52gif`	Converts a HDF5 image to a GIF
`gif2h`	Converts a GIF file to a HDF5 file
`h5diff`	Compares two HDF5 files
`h5dump`	Dumps a HDF5 file to ascii
`h5import`	Import ascii or binary data to a HDF5 file

HDF5 Tools II

Command	Description
`h5ls`	List information about a HDF5 file
`h5repack`	Repacks a file w/o compression/chunking
`h5repart`	Repartitions a file or family of files
`h5copy`	Copies objects to a new HDF5 file
`h5mkgrp`	Makes a group in a HDF5 file
`h5stat`	Display object and metadata information

Overview

What is HDF5

Data Model

Datasets

Tools

IO Sequence

Example

HDF5 IO Sequence

Very similar to normal IO sequence, only a few additional items need to be specified.

- open/create a file
- specify the dataspace
- create the dataset
- write the data
- close the file

HDF5 Fortran API

The fortran API is the same as the C API, however subroutines have a `_f` suffix and the last parameter is the return status.

C	Fortran
ierr = H5open(void)	H5open_f(ierr)

Overview

What is HDF5

Data Model

Datasets

Tools

IO Sequence

Example

Example

Write a 20×20 matrix to a file.

The code is provided on Janus in
`/projects/tibr1099/meetup/Intro_HDF5`

`Makefile`	Build definitions
`kinds.f90`	Precision definitions
`hdf_swrite.f90`	Serial example
`hdf_pwrite.f90`	Parallel example

Questions?

Online Survey

<Timothy.Brown-1@colorado.edu>

License