RFC: Virtual Object Layer

Mohamad Chaarawi

Quincey Koziol

This RFC describes a new abstraction layer within the HDF5 library that enables different methods for accessing data and objects that conforms to the HDF5 data model. This new layer, called Virtual Object Layer (VOL), is similar to the Virtual File Layer (VFL) abstraction where different Virtual File Drivers (VFD) can be plugged in to access data in the file in different ways; however the VOL layer would be at a higher level, to allow abstract operations on objects rather than blocks of bytes. A parallel I/O Metadata Server (MDS) plugin and a non-HDF5 storage backend plugin are used to demonstrate how the VOL abstraction would work with forms of data storage outside the typical HDF5 file paradigm.

# Introduction

The HDF5 data model is composed of two basic objects, groups and datasets. The data model in itself is very powerful and widely used by many applications, including High Performance Computing (HPC) applications. The main challenge currently in HDF5 is not in the data model itself, but in the native HDF5 single file format that has performance issues that vary widely over different platforms and has some limitations where all objects actually have to be accessible locally. Furthermore, in parallel applications, file access by several processes has to be coordinated in order to maintain strict consistency semantics to avoid file corruption.

In this RFC we propose to add a new abstraction layer internally to HDF5, implemented just below the public API. We call this new layer the Virtual Object Layer (VOL). The VOL intercepts all HDF5 API calls that could potentially access objects in the file and forwards those calls to a plugin “object driver”. The plugins could actually store the objects in variety of ways. A plugin could, for example, have objects be distributed remotely over different platforms, provide a raw mapping of the model to the file system, or even store the data in other file formats (like native netCDF or HDF4 format). The user still gets the same data model where access is done to a single HDF5 “container”; however the plugin object driver translates from what the user sees to how the data is actually stored.

# Design/Architecture

To implement a VOL interface capable of accepting any plugin a developer wishes to add, a modular design has to be put in place, similar to how the VFL interface in HDF5 is designed to accept different VFDs. The VOL is a higher level abstraction than the VFL abstraction. The following figure shows where the VOL is layered and how the data is accessed in the file:



## Calling in the VOL layer

The VOL intercepts all HDF5 API calls that potentially modify data on disk. Calls to one of these API functions only run sanity checks on the arguments passed in, and then immediately call the associated VOL callback for the API function. Operations, such as creating a group or retrieving a hyperslab from a dataset, will be captured and routed through the selected plugin that knows how the data is actually stored and capable of producing the results needed by the operations. For example, a call to H5Dcreate would be implemented within the HDF5 library as:

H5Dcreate(hid\_t loc\_id, const char \*name, hid\_t type\_id, hid\_t space\_id,

 hid\_t lcpl\_id, hid\_t dcpl\_id, hid\_t dapl\_id) {

 /\* Check arguments \*/

 check\_location (loc\_id);

 check\_type (type\_id);

 check\_dataspace (space\_id);

 check\_property\_lists (lcpl\_id, dcpl\_id, dapl\_id);

 /\* call the corresponding callback for H5Dcreate \*/

 result = H5\_VOL\_create (TYPE\_DATASET, …);

 /\* return result to user (yes the dataset is created, or no here is the error \*/

 return result;

}

The H5Dcreate call will inform the user that a dataset with the returned ID has been created in the HDF5 container being accessed with the chosen VOL plugin. Depending on the plugin selected, the H5Dcreate function might have created a dataset object in an HDF5 file, a netCDF file, a file on a remote machine, or any way a plugin is designed to create a dataset.

## Plugin Selection

The user should be able to select the VOL plugin and possibly provide some information to the corresponding plugin. The best way to do that is through a FAPL. For each plugin registered with the HDF5 library, a FAPL routine can be added to set the VOL plugin. If we consider the MDS plugin:

hid\_t fapl = H5Pcreate(H5P\_FILE\_ACCESS);

H5Pset\_fapl\_mds\_vol(fapl, …);

hid\_t file = H5Fcreate("foo.h5", H5F\_ACC\_TRUNC, H5P\_DEFAULT, fapl);

H5Pclose(fapl);

This would indicate that the MDS object layer plugin is to be used. All accesses to objects within the file opened using that FAPL would be routed through the MDS plugin.

In addition to plugins for the VOL class that will be added in the HDF5 library, any developer may write their own customizable plugin and tell the HDF5 library to use it at runtime. The plugin should be registered using:

H5VOLregister(H5VOL\_class\_t \*cls)

The *cls* parameter is a structure containing function pointers to the routines that implement all the callback routines the VOL class specifies. Those callbacks are discussed in more detail in the next section.

Other support routines to unregister a plugin and get information about a plugin should be provided:

H5VOLunregister(hid\_t driver\_id)

H5Pget\_vol\_plugin\_info(hid\_t plist\_id)

## Plugin Examples

### Different File Format Plugins

We mentioned earlier that a developer could design a plugin such that the file that stores data on disk could be of different format than the H5 format. This is very important to users that would like to use the HDF5 API and data model, but also desire the portability to use their files across different libraries or would like to access data from pre-existing files that are not in the H5 format.

The plugin, which understands the HDF5 data model and how the data is stored, will do the necessary “translation” needed for this kind of model. The following figure explains more on how this is achieved:



### Raw Format Plugin

The flexibility of the virtual object layer provides developers with the option to abandon the one file, binary format like the native HDF5 implementation. A “raw” file format could map HDF5 objects (groups, datasets, etc …) to file system objects (directories, files, etc …). The entire set of raw file system objects created would represent one HDF5 container as opposed to one HDF5 file with the native plugin. The actual mapping between all HDF5 objects to file system objects is subject to further research.

This plugin would allow the PLFS package (<http://institute.lanl.gov/plfs/>) to be applied to applications that use the HDF5 API, where access seems to be done to a single file from the high level, but is mapped by the raw plugin to multiple files/directories for performance benefits that were demonstrated by PLFS.

### Remote Access Plugin

The current HDF5 implementation requires applications to access a single HDF5 file that is also local on the file system where the application resides. A remote VOL plugin would allow access to files located remotely. The plugin could have an HDF5 server module located where the HDF5 file resides and listens to incoming requests (HDF5 operations that are routed through the VOL) from a remote process.

Remote visualization is an important use case for this plugin. Large, remote datasets are very expensive to migrate to the local virtualization system. It would be faster to just enable *in situ* virtualization to remotely access the virtualization data using the HDF5 API. This is however not possible at the moment with the native implementation for HDF5 as the file is required to be present locally.

### MDS Plugin

The MDS plugin uses the proposed strategy in the *Independent Metadata* RFC (http://www.hdfgroup.uiuc.edu/RFC/HDF5/VOL/2012-03-27-RFC\_independent\_metadata.docx), which allows processes in HPC applications to call HDF5 operations that modify metadata independently (the current HDF5 semantics require those calls to be collective). The MDS is a set-aside process that manages access to the file’s metadata to avoid corruption from simultaneous access by several processes. The design strategy is explained in details in the *Independent Metadata* RFC.

The MDS strategy puts all HDF5 metadata in a separate file, different than the raw data file. Only the designated MDS process will read/write data to that file. All HDF5 calls made by other processes that potentially need to access the metadata or raw data file must go through the MDS process to serialize access to the metadata and acquire the necessary locks for proper synchronization. The following figure shows the flow of the operation from the VOL layer to the MDS plugin:



All H5F, H5D, H5A, H5O, H5G, and H5L API calls are mapped to MDS specific implementations. An H5Dopen call, for example, would forward the parameters of the routine to the MDS that actually opens the dataset and returns the ID to the calling process along with some metadata for that dataset. The user gets back the ID of the opened dataset. An H5Dwrite call would ask the MDS process for a shared lock on the dataset being written to, to avoid having another process modify the metadata of the dataset, and then would actually write the data from the API call.

## Interchanging and Stacking VOL plugins

Accessing a file created with a different VOL plugin than the one it was created with would be a valid operation as long as the underlying file format is the same. This would be the user’s responsibility to ensure that the different plugins are interchangeable.

It would be also possible to stack VOL plugins on top of each other. This notion is similar to the idea of the split VFD, where underneath the split VFD itself, two file drivers would be used, one for the file storing the metadata and another for raw data. Some stackings make sense and others would be erroneous. For example, stacking the native HDF5 plugin on top of a non-HDF5 backend plugin does not make sense and is erroneous. The following figure showes a stacking of the remote plugin, where data is distributed remotely, on top of the native h5 plugin, where servers that store the data at remote locations use the h5 file format.



Another useful design option is to allow a mirroring plugin, where the HDF5 API calls are forwarded through a mirror plugin to two or more VOL plugins. This is an extention to the stacking feature. The following figure shows an example of a VOL mirror that maps HDF5 API calls to an h5 backend plugin and an XML backend plugin:



Another possible VOL plugin could be a statistics plugin that just gathers information on HDF5 API calls and records statistics associated with the number of calls to a specific API functions and corresponding parameters. This plugin would be very useful for profiling purposes. The statistics plugin would be stacked on top of another VOL plugin that actually performs the required access to the file.

# Implementation

The VOL class consists of data structure(s) that contain configuration variables that are set by each plugin and a collection of function pointers that maps all HDF5 API operations that access data on disk to the plugin’s corresponding implementations. Internally, each plugin is free to implement the underlying storage of objects in any way desired. The main goal is that the user still gets the same HDF5 API and data access model no matter which plugin is used. This section highlights some implementation details and the preliminary API for the VOL. We also give two examples for VOL plugins: a non H5 backend plugin and a Metadata Server plugin.

## VOL Callbacks

A VOL plugin is initialized based on the user selection with a file access property list API call or the default that the library chooses. The plugin is an instance of the VOL class and contains an implementation for the function callbacks specified by the VOL class.

The VOL class includes all the higher level logical callbacks that potentially touch data (metadata or raw data) on disk. All the API routines should map to one of those callbacks. The HDF5 API is too large however to have a 1 to 1 mapping from API routines to VOL callbacks, so we introduce the notion of a *generic* function call that developers could use to implement certain API routines that do not map to one of the VOL callbacks. This is similar to the UNIX ioctl() routine.

The HDF5 library currently does not support nonblocking I/O operations. However, since it is a very desirable feature, future versions of the library most probably will add nonblocking routines. In order to be able to support those API routines and not to double the number of VOL callback routines, it would be wiser to make all VOL callbacks nonblocking compatible, by adding an HDF5 request parameter (*hid\_t*) to all VOL callbacks. If the high level operation issued by the user is nonblocking, then the request would just be forwarded to the VOL plugin and the plugin itself would be responsible to support the nonblocking behavior. If the high level operation is a blocking operation, then a non-active request is forwarded to the plugin which detects that the request is inactive and implements the operation in a blocking behavior.

This is a preliminary set of VOL callbacks:

|  |  |
| --- | --- |
| Attribute Specific | API Calls |
| create | H5Acreate(\_by\_name) |
| open | H5Aopen(\_idx, \_by\_name/idx) |
| read: | H5Aread |
| write: | H5Awrite |
| get | H5Aget\_XXX |
| remove | H5Adelete(\_by\_idx/name) |
| close | H5Aclose |
|   |   |
| Dataset Specific | API Calls |
| create | H5Dcreate/H5Dcreate\_anon |
| open | H5Dopen |
| set\_extent | H5Dextent, H5Dset\_extent |
| read: | H5Dread |
| write: | H5Dwrite |
| get | H5Dget\_XXX |
| set\_extent | H5Dset\_extent, H5Dextend |
| close | H5Dclose |
|   |   |
| Datatype Specific | API Calls |
| commit | H5Tcommit |
| open | H5Topen |
| close | H5Tclose |
|   |   |
| File Specific | API Calls |
| create | H5Fcreate |
| open | H5Fopen |
| flush | H5Fflush |
| get | H5Fget\_XXX |
| misc |   |
| optional |   |
| close | H5Fclose |
|   |   |
| Group Specific | API Calls |
| create | H5Gcreate(\_anon) |
| open | H5Gopen |
| get | H5Gget\_XXX |
| close | H5Gclose |
|   |   |
| Link Specific | API Calls |
| move | H5Lcopy,H5Lmove,H5Gmove, H5Gmove2 |
| create | H5Lcreate\_(hard/soft/external/ud), H5Glink, H5Olink |
| remove | H5Ldelete(\_by\_idx), H5Gunlink |
| get |   |
|   |   |
| Object Specific | API Calls |
| open | H5Oopen(\_by\_idx/addr) |
| copy | H5Ocopy |
| lookup |   |
| free\_loc |   |
| get | H5Oget\_XXX |
| misc |   |
| optional |   |
| close | H5Oclose |

One could argue that the interface can be simplified by having, for example, a common open routine for all objects (like the H5Oopen) routine. We found however that having a common interface would be detrimental to performance in certain plugin implementations, since it would require querying the VOL plugin twice, once for looking up the object location and once for opening the object. Having an optional callback for every object and a generic one would solve this issue. The VOL layer would check if the specific object callback is implemented in the VOL plugin, and if it is not, it would fall back to the generic object callback routine. This way we don’t sacrifice performance for generality.

Another design issue comes to mind when looking at the large set of functions, and that is whether to lump all the functions together into one data structure as the VOL class, or to have a more object oriented approach, where we can either have a general class that contains all common functions, and then children of that class that contain functions specific to certain HDF5 objects, or for each object have a set of callbacks that are specific to that object. Switching from one design to another is not that difficult if we decide to change our minds.

## H5Ocopy & H5Ocompare

The HDF5 API has two routines that will require extra attention to be compatible with the VOL layer, and those are the object copy and compare routines:

* herr\_t H5Ocopy( hid\_t src\_loc\_id, const char \*src\_name, hid\_t dst\_loc\_id, const char \*dst\_name, hid\_t ocpypl\_id, hid\_t lcpl\_id );
* htri\_t *H5Ocompare(hid\_t loc1\_id, const char \*name1, hid\_t lapl1, hid\_t loc2\_id, const char \*name2, hid\_t lapl2, hid\_t cmppl\_id, H5O\_cmp\_cb\_t \*cb\_info);*

If the two objects are in “*containers”* of the same type, i.e. created using the same VOL plugin, then the VOL layer would be able to detect that and forward those calls to the corresponding implementation using a designated VOL callback. This would allow different sorts of optimizations that a VOL plugin could choose to implement in order to copy or compare two objects. On the other hand, if the objects belong to two different containers, the VOL plugin would not be able to interpret an object created by a different plugin, and so a different approach needs to be considered.

Comparing two objects that belong to different types of HDF5 containers require processing independent of the VOL plugins. One way to accomplish this is to add utility routines to the VOL class to retrieve certain characteristics about an object for comparison. For example, datasets would need utility routines to retrieve the dataset name, dimensions, type, space, and elements. Once those values are retrieved for each object from each VOL plugin, the comparison operation could return the required result. The copy operation would require the same level of processing on the source object. Then using the fetched attributes, the same object is created using a create callback of the destination object’s VOL plugin.

# Recommendation

We recommend start working on this layer incrementally. After having the architecture in place, developers can independently pick certain API functions and route them through the VOL. As more plugins are inserted that implement certain API calls, developers would modify the top level routines to go through the VOL. This incremental approach limits the impact of this vast architecture change to the library on other developers who are working on different portions of the library.

The following is a tentative implementation plan for items addressed in this RFC:

*April/May 2011*: Have the VOL architecture mostly in place inside the library

*July 2011:* A good prototype implementation for the MDS plugin within the HDF5 library, that provides good grounds for performance analysis

Revision History

|  |  |
| --- | --- |
| *January 9, 2012:* | Draft circulated internally  |
| *March 9, 2012* | Draft circulated internally and externally |
| *April 5, 2012* | Draft circulated internally and externally |

 [References]