RFC: Breaking Free from the Collective Requirement for HDF5 Metadata Operations

Mohamad Chaarawi

Quincey Koziol

This RFC proposes a new feature in HDF5 that breaks the collective requirement for operations that modify metadata and allows independent space allocation from the file’s “end of address” (EOA) location. At the moment, the current implementation requires all operations that modify metadata or allocate space in an HDF5 file to be collective for synchronization purposes. This requirement is detrimental to the performance of some HPC applications and puts several limitations on application developers.

This work is still at the design phase, and this RFC details the several design options that have been considered. Several use cases will be outlined and used to draw out the pros and cons of the different design options.

# Introduction

Parallel HDF5 is layered on top of MPI to provide parallel I/O functionality. MPI defines a set of parallel I/O functions that can be used by processes to access data on disk independently or collectively. Independent I/O operations are similar to serial POSIX I/O, where each process individually accesses data on disk. On the other hand, collective I/O operations require that all processes participate in the I/O function call.

Some collective operations, such as file opens, require that all processes’ arguments to the function specified are the same. Other collective functions, such as writes, do not have that requirement, and the amount of data that each process accesses can be different. Processes that don’t have any data to write can specify a 0 byte selection; however they *do* have to call the function since it is collective.

When collective I/O is performed, the MPI library can provide different algorithms underneath the hood to optimize the I/O to disk. For example, when an MPI implementation detects that many processes are writing very small chunks of data to disk at once, it can have a small set of processes combine data from all the processes and write data to disk in larger blocks, improving performance on most parallel file systems.

Parallel HDF5 is positioned above MPI-I/O in the application’s software stack, and instead of treating the file as just a stream of bytes (as MPI-I/O does), gives the application a structured view of the file. The user organizes his file using different HDF5 objects (groups, datasets, etc.) and dependencies (links) between those objects. The HDF5 library is responsible for maintaining information about the file structure in the form of metadata, which is also stored in the file.

Parallel HDF5 allows all processes that opened the file to access dataset elements independently or collectively. However due to synchronization issues, all operations that need to modify the structure of the file, or in other words the file’s metadata, are required to be collective. For example, to create a subgroup in the root group of the file, a process needs to tell the other processes that it is performing that operation. This is done to notify the other processes that this group now exists and that they should not attempt to re-create it or simultaneously modify the file’s metadata.

Another functionality that requires synchronization between the processes is space allocation within the file. All processes again need to coordinate in order not to allocate overlapping space.

In this RFC, we propose several design options that will remove the collective requirement for space allocation and operations that require modifying the file metadata. Any solution proposed has to achieve the following:

1) Allow an independent (but globally visible), scalable "fetch-and-add" operation to the "end of allocated space" (EOA) value in the file (to allow new space in the file to be allocated to an individual process, without involving all the other processes).

2) Allow [the metadata for] an object to be locked, updated, unlocked, and made visible to other processes without requiring a collective operation (to allow changes to the structure of a file).

Before the design options are presented, the next section lists some basic use cases and some HPC application use cases that need to be considered when designing a solution.

# Use cases

This section presents several use cases of typical parallel applications for HDF5 and motivates the need for this RFC and at the same time outlines the challenges for a solution. All of the use cases below assume that the MPI application is writing to a single, shared file.

Some of the basic use cases include:

U.1) A parallel application with n processes, where each process creates an object in the root group for its private use. This is currently done by having all processes call the object creation function n times, despite the fact that each process needs only 1 object. But since each operation is modifying the file’s metadata, n creation calls are required.

U.2) A process within a parallel application performs independent read/write operations to a dataset while another process comes in and modifies the metadata of that dataset, making the former process’s access possibly obsolete. Currently, this use case would not happen, because all processes are required to be present at the time of metadata modification. However, this use case becomes possible when the collective requirement to metadata modification is removed and should be taken into consideration by any design option. Modifications to metadata can include, for example, extending the size of a dataset.

U.3) Collective read/write to a dataset, where all processes that opened the file are required to call the function with or without data to read/write.

U.4) 2 or more processes creating an attribute on their own datasets. In the current implementation, similar to use case U.1, the attribute creation function needs to be collective, even though they are being created on different datasets. The performance overhead is quite clear as we scale to a larger numbers of attributes and processes.

U.5) 2 or more processes creating an attribute on a shared dataset. A solution that makes this operation independent has to take into account that there are other processes that can create an attribute (or any other object) at the same time and those objects might overlap in the file if they are not serialized.

U.6) Access to compressed datasets by processes (individually or collectively).

We analyze some different applications that are considered very common HPC applications that use parallel HDF5:

A.1) Flash I/O:

Flash I/O is a benchmark that mimics the I/O behavior of the FLASH application (<http://flash.uchicago.edu>). The application behavior when it comes to I/O can be described as the following:

* Nothing process specific: All objects are shared between all processes, and the number of HDF5 objects is the same with all process counts
* Very flat hierarchy, all datasets are under the root group
* Collective and Independent I/O can be used.
* Three files are created

A.2) *Others TBD*

# Space Allocation at EOA

In this section, the solution to allow independent space allocation between all processes will be explained.

Currently, space within the HDF5 file is allocated either from a free list (of previously used and released blocks within the file), or from the end of the file’s allocated space (EOA). Recycling space within the file is complex in parallel, as well as unusual, so will be disabled for parallel applications, requiring all space allocation to be performed at the file’s EOA.

Allocating space at the file’s EOA is very simple in serial HDF5 applications: the EOA value begins at offset 0 in the file and when space is required, the EOA value is incremented by the size of the block requested. As mentioned earlier, space allocation using the EOA value in parallel HDF5 applications can result in a race condition if processes do not synchronize with each other, causing multiple processes to believe that they are the sole owner of a range of bytes within the HDF5 file.

The simplest solution to this synchronization issue is to use a globally visible atomic Fetch and Add (F&A) operation that guarantees each process can get the current EOA value and increment it by the size of the block that the process needs to allocate, all in an atomic operation. The MPI-2 standard does not define such an operation; however it does provide one-sided communication operations in the “Remote Memory Access” (RMA) section of the standard. Using MPI RMA routines, an F&A operation was implemented by Dr. William Gropp. The algorithm is detailed in his book, *Using MPI-2.*

To allocate space at the EOA using this operation, we do the following:

1. Create an MPI window on all processes with the root process exposing a memory buffer where the fetch and add will take place
2. All processes wanting to increment the EOA will execute the following code section to get the current EOA and increment by their requested allocation number.

MPI\_Win\_lock(MPI\_LOCK\_EXCLUSIVE, 0, 0, win);

MPI\_Accumulate(increment\_by\_x, nlevels, MPI\_INT, 0, 0, 1, acc\_type, MPI\_SUM, win);

MPI\_Get(get\_array, nlevels, MPI\_INT, 0, 0, 1, get\_type, win);

MPI\_Win\_unlock(0, win);

The current EOA for this process would be the increment + all the values in get\_array which corresponds to all the previous “adds” done by other processes up until this process added its own increment;

We have measured the performance of this algorithm on hopper[6], a Cray XE6 linux cluster, using a micro benchmark, where each process executes an F&A 1000 times. The performance was measured when all processes are executing F&A at the same time (column 2 below), half of them at the same time (column 3 below), and one at a time (column 4 below). Performing this test showed the effect of different workloads when all the F&A happens at one set-aside process. The results below show the performance in seconds for all 1000 F&A operations.

|  |  |  |  |
| --- | --- | --- | --- |
| Num\_procs | N | N / 2 | 1 |
| 128 | 1.137938 | 0.628627 | 0.015955 |
| 256 | 2.459199 | 1.314574 | 0.015274 |
| 512 | 5.186699 | 2.874677 | 0.030334 |
| 1024 | 13.985337 | 6.528552 | 0.027759 |
| 5120 | 249.922206 | 59.980488 | 0.039649 |
| 10240 | 1923.69371 | 215.603612 | 0.049559 |

As expected, when all processes are executing fetch and add at the same time, the performance does not scale well, because it’s still an all-to-1 communication. When the workload is lighter, the problem scales well, which is expected to be the case for HPC applications we are targeting.

Some of the disadvantages of this approach are:

* Needs to set-aside the root process (i.e. dedicate it to being available for MPI operations), or else the root process will not be able to promptly respond to the MPI RMA operations.
* There are performance issues when a large number of processes try to fetch and add at the same time repeatedly, however we don’t see that as a common case.

# Independent Metadata Operations

This section presents the three classes of designs that have been studied and selected from a larger set of options as candidates for breaking the collective requirement for metadata operations.

## Distributed Lock Manager

A distributed lock manager (DLM) is a commonly used approach to handle access to shared resources. Lustre, for example, uses that approach to handle client requests for locking, disk allocation, storage and retrieval, all through Object Storage Servers.

In our solution, we can use a DLM to serialize access to HDF5 objects. All operations needing to read or modify metadata for an object need to acquire a lock on that object from the DLM. The DLM is a set-aside process, carved off from the application processes that opened the file and does not perform file I/O itself.

The DLM implementation can be done in several ways:

* MPI point-to-point communication & RMA:

Using the algorithm described in [3], processes acquire and release locks for a certain HDF5 object using a combination of RMA and point-to-point operations. A root process will hold, for each object to be locked, an n-bit value that other processes access using RMA operations to turn their corresponding bit on, asking for the lock, and check the other bits of other processes. If the lock is available, meaning all the bits of other processes are off, the process can assume it has been granted the lock; otherwise it posts an MPI\_Recv from the process that currently has the lock. When the process owning the lock is done with it, it turns off its bit at the root node and checks for the next process asking for the lock, then sends the lock to that process using MPI\_Send. For fairness issues, the lock is granted to the process with the next highest rank than the process holding the lock. This approach eliminates polling for the lock remotely and locally at memory regions, and just blocks on an MPI\_Recv.

This algorithm handles only exclusive locks, but it can be extended to handle read locks, using the same idea as an algorithm proposed in [7]. They do use compare & swap operations, which are not available in MPI-2, but I’m pretty sure we can come up with a similar algorithm to grant shared locks.

* Using MPI point-to-point non-blocking operations, where the DLM posts n MPI\_IRecv and calls MPI\_Testsome to check all the requests that have arrived. It then queues up those requests on their corresponding objects and process the queues granting locks and posting another set of MPI\_IRecv to get the locks back.

In both approaches, there will be the need at some point to use more than one process to act as a DLM and distribute the lock management of HDF5 objects. However, doing this still does not solve the problem of hot objects being handled by one server. Another issue to note is that when a process gets a lock, it’s also getting a set of metadata entries that it can read/write, and sending the changes (in case of a write lock) back. So the interprocess communication is more complex than it appears.

The amount for messages sent and received with the DLM is fairly large. If we consider use case U.1 (or U.5, which is similar), each process must acquire a lock on the root group before being able to create its own dataset. This can get ugly once we have a very large number of processes. In use case U.4, this problem is better handled, because each process is getting a lock on its own dataset, so the attribute creations are done in parallel. There is still a large amount of messages sent between the application processes and the DLM.

## Synchronization Step

In this approach, each process modifies its metadata cache independently and logs the changes locally. Whenever a collective operation, or a flush, is issued by the user, the processes enter a sync point, and resolve any conflicts that appear in the logs. This approach is meant for applications where users guarantee that the HDF use of metadata is simple enough that the resolve step does not introduce unresolvable conflicts. An example of conflicting operations would be something similar to use case U.2, where the process that is writing to a dataset might be writing at obsolete locations if there is no sync step between the time that this process issues the writes and another process modifies the metadata of the dataset. A drawback to this approach is that the metadata operations are not entirely independent, since a sync step is still required. However, no set-aside processes are needed.

## Metadata Server

In the DLM approach, we use a lock manager to coordinate access to the metadata of an object. The lock manager would not open the file. Instead, the client will perform all the I/O to disk after coordinating with the lock manager. A different approach would be to have a metadata server (MDS) actually open the file and read/write metadata to the file.

The MDS would perform all metadata operations requested by a client and coordinate between clients’ requests locally. After making a request, a client would wait for the MDS to reply back with the results from their request. For example, if a process needs to create a dataset, it will send the request with all the parameters required to the MDS, and it will get back a dataset identifier. All synchronization issues are handled at the server. If we consider use case U.1, where n processes are creating n datasets in 1 group object, the MDS will receive n requests and reply back to each process with its own dataset. Using the DLM approach as we saw earlier, each process would need to get a lock on the group object, create the dataset, and release the lock. The interprocess communication in the DLM case will be much greater.

If we consider use case U.2, we can see that for the MDS approach to work we still need locking, because when a process is doing I/O to disk (raw data), this is not handled by the MDS. Thus, prior to executing a raw data read/write to an object a process must acquire a read lock from the MDS. The MDS is responsible for managing locks similarly to what the DLM does, but the number of locks and the amount of information communicated should be much lower. In use case U.6, where the dataset being accessed is a compressed, a client might also need a write lock when it wants to access that dataset.

When a collective dataset access is performed (use case U.3), all processes need to inform the MDS that its participation is required. This is required because the MDS is part of the communicator that opened the file and has to participate in the operation with a 0 byte contribution.

Having 1 MDS is obviously not scalable. At some point more than one MDS need to be used, where objects are spread between them. To avoid the problem of access to hot objects, we can apply a distributed hash table to spread the objects among the MDSs with replication. I believe C-MPI[2] can be useful in this case.

Clients accessing metadata for an object can still cache metadata and bypass the MDS in some cases when the file is opened read-only. The MDS could also maintain version information about the state of each object, where the client checks if the local version in its cache is the updated version or not, avoiding some communication overhead. Access to non-mutable objects would also bypass the MDS and would not require any locking.

Overall this approach looks the most attractive for the metadata problem, however given the architecture of the HDF5 library, the amount of work required to implement an MDS approach is fairly substantial.

### Enhanced MDS

Another approach to consider is using an enhanced version of the MDS. Objects in HDF5 will be marked as private or public. Public objects are those that are allowed to be accessed by all the processes that opened the file. The root group in HDF5, for example, should be a public object. A private object can be, for example, a dataset created as in use case U.1 for a process private use, where only this process will access that dataset. Access to metadata of public objects has to go through the MDS, whereas access to metadata of private objects can be done locally. If a private object is needed to be made public, the process that holds it needs to transfer the “ownership” to the MDS, and that includes pushing all the metadata to the MDS or writing the metadata to disk and informing the MDS about it. A public object that needs to be made private by a process would require the process to send the request to the MDS asking to change the status of an object to private. When the MDS clears all locks on that object, the request to make the object private can be done, and the MDS can release all ownership (metadata, cache items, lock queues, etc…) of that object.

### Data Staging

The MDS approach delegates handling metadata operations to a metadata server. A wider extension to that approach would be Data Staging, which delegates all I/O operations to server processes. ADIOS [1], another high level I/O library, uses data staging for all I/O operations, where a set of processes is set aside to handle the I/O requests for application processes. While the purpose of this RFC is to break the collective requirement of metadata operations, it is worth to explore this option that would achieve that, in addition to a much wider set of functionalities. Asynchronous independent metadata operations would be introduced, since the data staging nodes will handle all the consistency issues. Application processes would just post their I/O requests and return to computation immediately. Similar to all non-blocking operations in MPI, a mechanism to check on a current request must be in place.

In order to handle the contention on network resources between the application communication and the background data transfer, a communication protocol is used to block I/O data from being sent to staging nodes when application communication is occurring.

Some limitations to this approach include:

* set-aside process(es)
* scalability issue handled by using more data staging processes
* implementation is very complex in terms of API changes, functionality, etc…

# Progress Function

The MDS approach (and the enhanced MDS) requires set-aside process(es) to accept requests from clients to access metadata. A different approach to consider is layering, on top of HDF5, a library, where the user is responsible to call progress functions in his application to synchronize HDF5 metadata changes. This can be applied in two ways:

* The progress function mimics the collective behavior of how metadata operations are performed currently, where all processes will come together and coordinate access to metadata. This will provide the non-collective functionality for metadata operations, but suffers from the overhead that all processes have to join with the progress function, to be able to proceed.
* A process can be designated as an MDS to an object. When clients would like to access metadata to that object they send requests to its MDS (similarly to the MDS design). When a designated MDS process enters a progress function, it receives all the requests, processes them, and then goes back to the application. The enhanced MDS approach can be applied here too, where we only coordinate access to objects that are marked global.

# Recommendation

As mentioned earlier, this work is still at the design phase. While we have a tentative roadmap on how to process with this work, things might change depending on various reasons, such as new input from the HPC community, funding, etc… For now, we are looking at proceeding with the Gropp F&A approach for the space allocation problem, and the Enhanced MDS approach for the metadata operations, while moving later towards Data Staging similar to ADIOS. We would eventually like to provide both options to the user in terms of either setting aside processes or calling progress functions in the application, but will begin with the set-aside process approach.

# Revision History

|  |  |
| --- | --- |
| *September 2, 2011:* | Version 1 circulated for comment within The HDF Group. |

# [References]

[1] ADIOS: <http://www.olcf.ornl.gov/center-projects/adios/>

[2] C-MPI: <http://www.mcs.anl.gov/publications/paper_detail.php?id=1229>

[3] RMA + point-to-point locking [http://www.mcs.anl.gov/~robl/papers/ross\_atomic-mpiio.pdf](http://www.mcs.anl.gov/%7Erobl/papers/ross_atomic-mpiio.pdf)

[4] DSM 1 <http://nextmuse.cscs.ch/sites/default/files/Eurompi2010_paper_17.pdf>

[5] DSM 2 **Data Redistribution using One-sided Transfers to In-memory HDF5 Files**

**[6]** <http://www.nersc.gov/systems/hopper-cray-xe6/>

**[7]** http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.145.395&rep1&type=pdf