RFC: New HDF5 API Routines for HPC Applications Read/Write Multiple Datasets in an HDF5 file

Peter Cao
Quincey Koziol
Jonathan Kim
Neil Fortner

The HDF5 library allows a data access operation to access one dataset at a time, whether access is collective or independent. However, accessing multiple datasets will require the user to issue an I/O call for each dataset. This RFC proposes new routines to allow users to access multiple datasets with a single I/O call.
This RFC describes the new API routines, *H5Dread_multi()* and *H5Dwrite_multi()*, which perform a single access operation to multiple datasets in the file. The new routines can improve performance, especially when data accessed across several datasets from all processes can be aggregated in the HDF5 or MPI-I/O library.
The RFC was initially released in 2012. This latest revision reflects a renewed effort to officially release and support the feature in the HDF5 library as part of the ECP ExaIO project[1].

1 Introduction

Parallel HDF5 (PHDF5) supports both independent and collective dataset access. When collective I/O is used, all processes that have opened the dataset may do collective data access by calling *H5Dread()* or *H5Dwrite()* on the dataset with the transfer property set for collective access. Accessing datasets collectively using the MPIO VFD can improve I/O performance [1] since MPI can aggregate data into large contiguous accesses to disk instead of small non-contiguous ones.
However, the current HDF5 library does not support a single I/O call for accessing multiple datasets. For example, if one accesses five datasets in a file, one will need at least five I/O calls for each dataset. We plan to add two new functions to the HDF5 library: *H5Dread_multi()* and *H5Dwrite_multi()*. With the proposed new read/write functions, users will make a single function call to read or write data to multiple datasets in an HDF5 file. Note that the multiple datasets are located in the same HDF5 file for the scope of this task. The new functions can be used for both independent and collective I/O access, but this task's primary purpose is to utilize the collective I/O case.

[1] *Part of this research was supported by the Exascale Computing Project (17-SC-20-SC), a joint project of the U.S. Department of Energy's Office of Science and National Nuclear Security Administration, responsible for delivering a capable exascale ecosystem, including software, applications, and hardware technology, to support the nation's exascale computing imperative*

1.1 Purpose

The purpose of the work is to implement two new functions to the HDF5 library: *H5Dread_multi()* and *H5Dwrite_multi()*, which will often give better I/O performance when collective I/O is used.

1.2 Scope

H5Dread_multi() will take information from multiple datasets and read data from a file for the datasets requested. If collective I/O is used, a single I/O call is used for better performance. Overlapping data selections are supported.
H5Dwrite_multi() will take information from multiple datasets and write data to a file for the datasets requested. If collective I/O is used, a single I/O call is used for better performance. If data selections from multiple ranks overlap, the behavior of the H5Dwrite_multi() is not defined. Therefore, overlapping data selections should not be used.
Datasets requested in H5Dread_multi() and H5Dwrite_multi() must reside in the same file. The new functions do not support datasets across files.
Once the feature is fully productized, the benchmarks mentioned in this report will be rerun, along with other standard I/O kernels, to fully determine the performance of the feature. The final deliverables will also include a reference manual entry for each function and regression tests.

2 Use Case

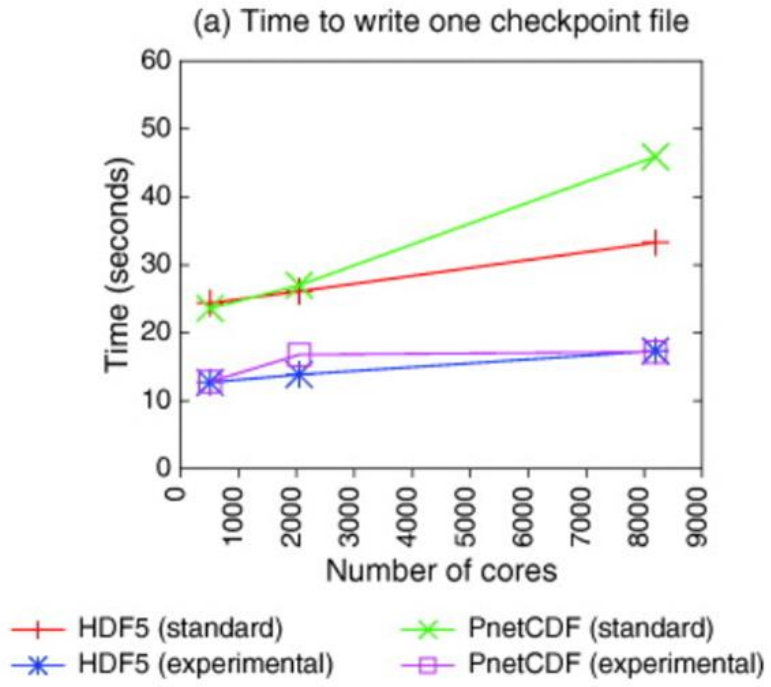
This section presents a few older (FLASH I/O and E3SM) and recent (CGNS) use cases. The primary purpose of the use cases is to show how the new library functions can be used to improve I/O.

2.1 Improving FLASH I/O for an ANL project

FLASH code was designed to simulate thermonuclear flashes on a Cartesian, structured mesh. The mesh consists of cells that contain physical quantities such as density, pressure and temperature (also known as mesh variables). Each cell is assigned to a self-contained block. In the FLASH file layout, a block is stored in an HDF5 file, and mesh variables are stored as 4D datasets in the file.
The time spent on file I/O in a FLASH simulation is a common bottleneck. Using collective I/O[1] improves I/O performance for HPC applications like FLASH. Current parallel HDF5 performs collective I/O on a single dataset and requires many I/O calls in FLASH simulations since many variables are frequently accessed during each time step. Using the proposed collective I/O on multiple datasets will reduce the number of I/O calls. In an experimental study, Rob Latham, Chris Daley, etc.[2] have shown that the average time for writing a file is reduced by half when collective I/O on multiple variables is used:

> "*The standard file layout approach (storing application data in multiple library objects), however, offers a slight performance trade-off. Each function call represents a relatively expensive I/O operation. All other factors aside, if the goal is to achieve the highest I/O performance a better approach would describe the entire application I/O pattern and then execute a single call. If the application places all mesh variables into a single I/O library object, as in the experimental file layout approach, then a single I/O library call could be issued to service all application variables instead of N separate calls. Experiments confirm that this approach does improve performance*."[2]

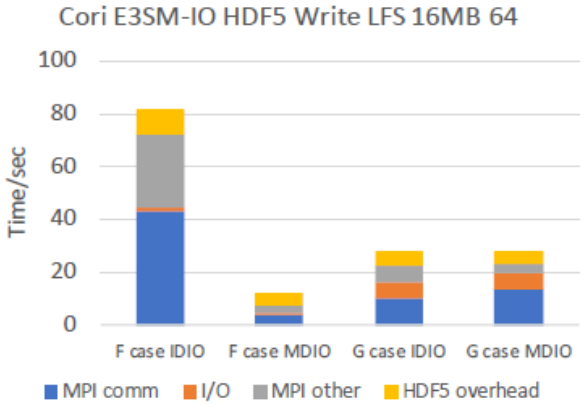
The following figure shows the I/O results for standard file layout (storing mesh variables in multiple datasets) and the experimental file layout approach (placing all mesh variables into a single I/O library object). The results show that the single I/O approach (the experimental file layout) reduces the time to write checkpoint files by half.

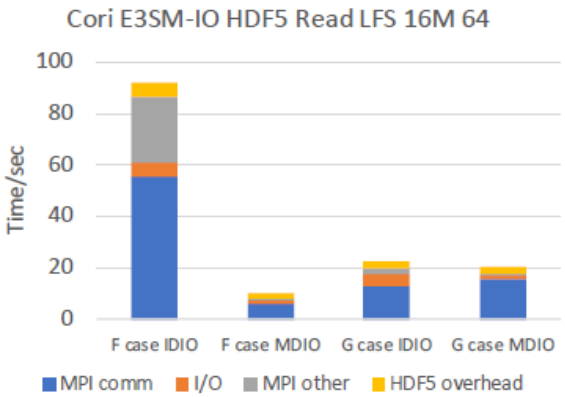


2.2 Investigating E3SM performance improvements for the Exascale Computing Project

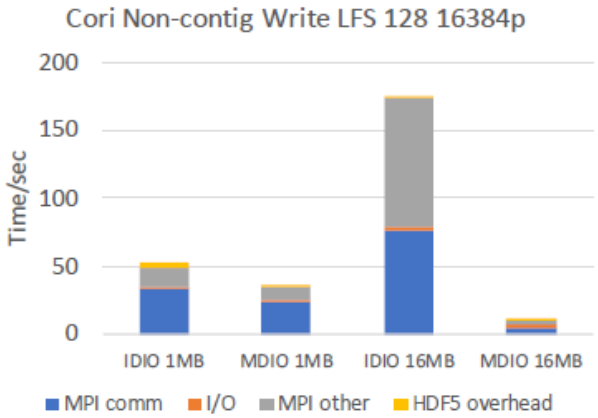
The Energy Exascale Earth System Model (E3SM) is a large-scale Earth modeling code that couples ocean, atmosphere, and ice models. As part of the Exascale Computing Project, the performance of the existing multi-dataset I/O prototype in HDF5 was evaluated for two different cases in E3SM, as well as in a synthetic I/O benchmark[3]. We briefly summarize these cases here; for full results, see the reference.

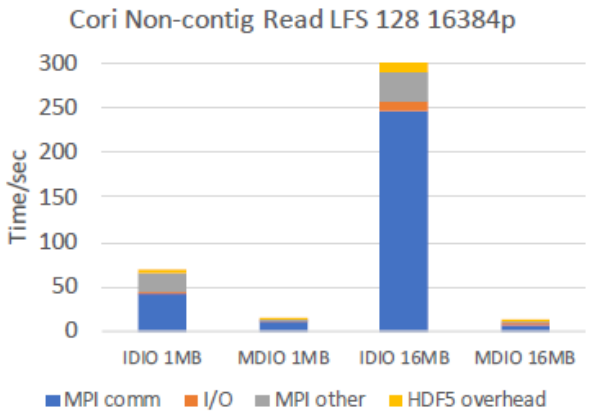
Two cases were evaluated for the E3SM benchmark: the "F" and "G" cases. The F case involved large numbers of datasets, contiguous smaller amounts of data per dataset. The G case involved smaller numbers of datasets with large amounts of data per dataset. This means the F case is more likely to benefit from multi-dataset I/O since more I/O operations can be combined, and it is not as limited by raw bandwidth. We indeed see a significant performance improvement of approximately 10x in the F case with only a minor improvement in the G case:



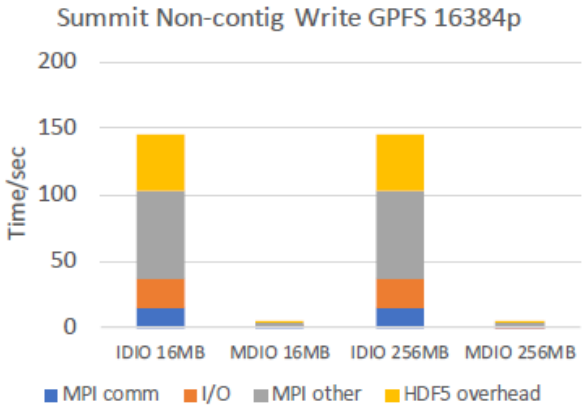


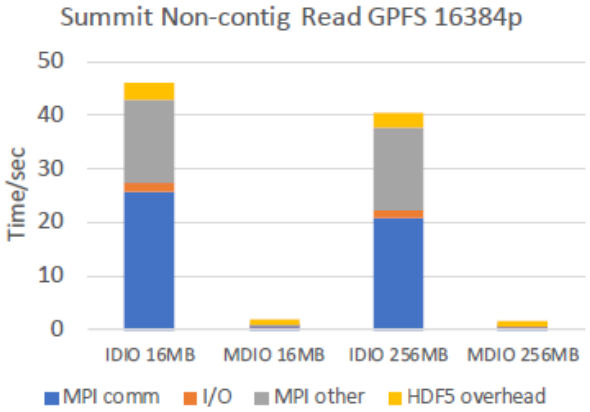
The synthetic benchmark pushes this further by increasing the number of non-contiguous blocks in the datasets, while having as many datasets as the E3SM F case and a total amount of data in between the E3SM cases. This, similarly to the E3SM F case, shows substantial performance improvement from multi-dataset I/O:





While a different system using GPFS showed even greater improvements:





2.3 Investigating CGNS performance improvements with multi datasets

For applications that typically deal with a vector of components of fields, the number of multiple datasets for each vector field may be small. For example, a CGNS parallel benchmark uses multi-datasets to write/read:

(a) x,y,z components of a 3D Vector (nodal coordinates)
(b) x,y,z components of another 3D Vector (momentum components)
(c) x,y components of a 2D Vector (General solution field components)

8 H5Dwrite and 8 H5Dread are simplified to 3 H5Dwrite_multi and 3 H5Dread_multi calls with multi dataset. The improvements for writing and reading on Summit, with collective I/O, where the problem size remains the same (strong scaling), were,

Number of ranks	Percent Improvement with multi dataset
2688	10%
10752	6%

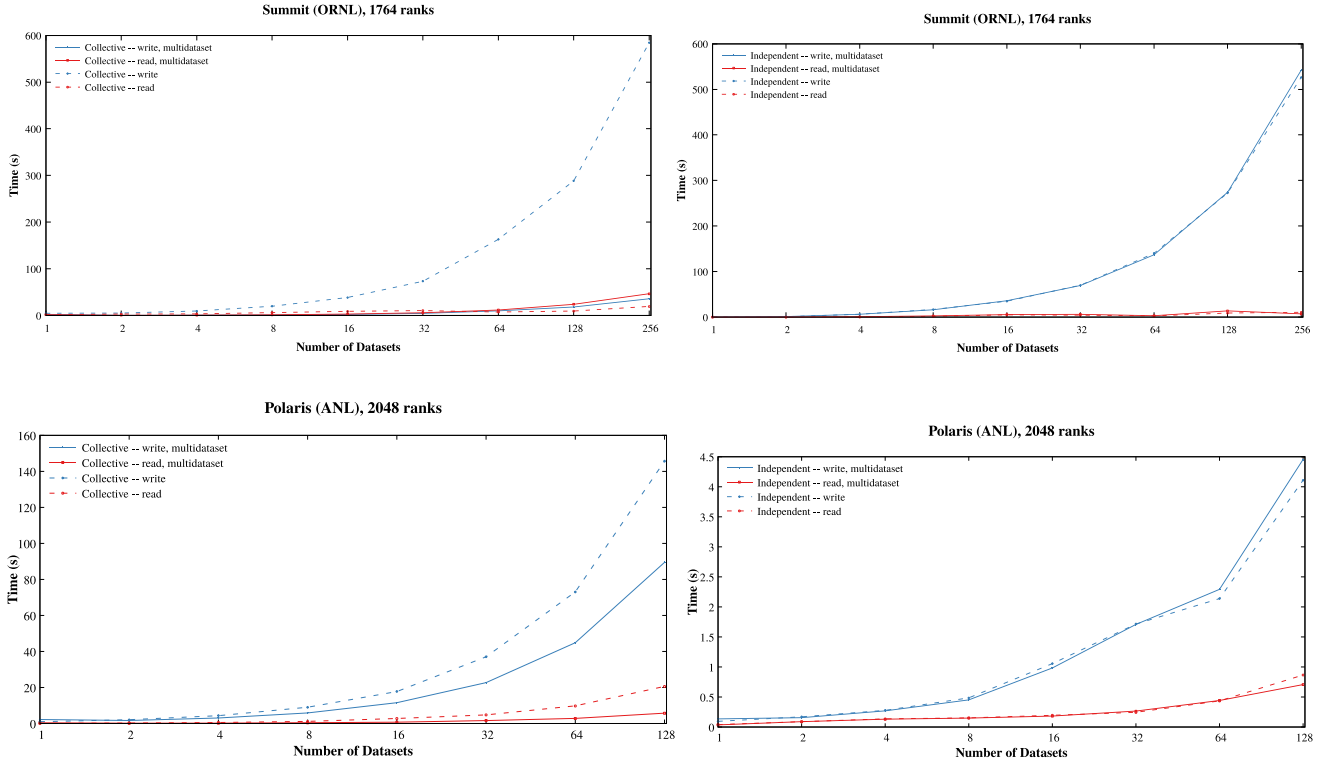
2.4 Stand-alone Benchmark

A benchmark program creates datasets which is a function of the number of MPI ranks, *mpi_size*. For this study, the number of ranks was constant, and the number of datasets was increased. The total dataset sizes are summarized in Table 1. Each MPI rank performed I/O on a column hyperslab selection of the datasets.

Table 1 The total dataset sizes as the number of datasets increases.

Number of Datasets	Total Dataset Size (GiB)
1	7
2	14
4	28
8	56
16	112
32	225
64	451
128	903
256	1806

The write performance was generally improved on both Polaris and Summit as the number of datasets increased. It was observed that read performance for multi-dataset decreased, which will be further investigated. Although multi-dataset showed improvement on Polaris for collective, neither method for collective IO could beat the overall time when using independent I/O.



3 Functional requirements

The two main purposes of the H5Dread/write_multi() functions are: better I/O performance and convenient function calls for multiple datasets. In addition, the two new functions should meet the following specific requirements other than the standard HDF5 API function requirements.

3.1 H5Dread_multi()

This function will read data from a file to memory buffers for multiple datasets. This function should attain no less I/O performance than reading data from individual datasets.

3.2 H5Dwrite_multi()

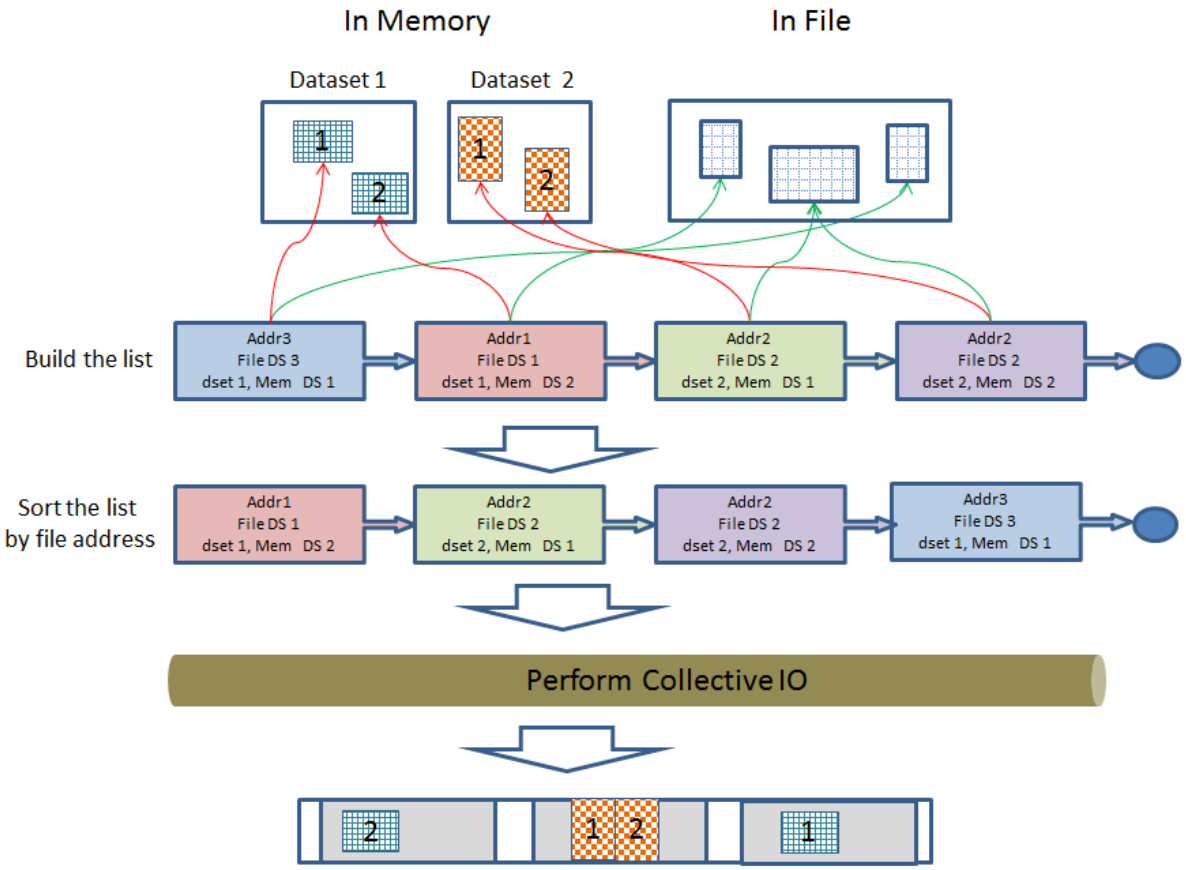
This function will write data in memory to a file for multiple datasets. The selection of data regions to be written cannot overlap. This function should attain no less I/O performance than writing data for individual datasets.

4 Implementation design

The basic approach for multi-dataset collective I/O is similar to the POSIX lio_listio() call, which takes a list of buffers, offsets and lengths to perform a series of read and write operations on a file in a single call. The primary difference from the typical HDF5 API call is that the new routines add information from multiple datasets to the I/O mapping list and construct larger MPI-derived datatypes for collective I/O operations for read and write operations in a separate manner. Internally, the multi-dataset implementation will be similar to the current implementation of collective chunk I/O on a single dataset.

4.1 Top-level design

The following example chart shows the conceptual implementation approach for the new API functions.



Note that sorting the list by file addresses is necessary because MPI requires the file type to consist of derived datatypes whose displacements are monotonically non-decreasing.

4.2 Code-level design

4.2.1 Implementation details

The current implementation achieves multi-dataset I/O by creating a new structure, *H5D_dset_io_info_t*, which contains all the information necessary for I/O on a single dataset. This struct is used in conjunction with the existing *H5D_io_info_t* struct, which now contains only information applicable to all datasets in an I/O operation. *H5VL__native_dataset_read()* and *H5VL__native_dataset_write()* then allocate an array of these structs and pass them to *H5D__read()/H5D__write()*, which iterate over all the datasets in the operation, passing the corresponding element of the array of *H5D_dset_io_info_t* structs to each dataset's layout's *io_init* operation. The io_init operations operate mostly as before, except they place info in the *H5D_dset_io_info_t* struct, and place info on I/O within a single chunk or contiguous dataset in an *H5D_piece_info_t* struct. Here, "piece" is a term for a contiguous data block, i.e. a single chunk or contiguous dataset.

After the io_init loop, a second pass is made over the datasets, making a new mdio_init layout callback for each, in which each dataset adds its piece(s) to a global unsorted list of pieces. This list of pieces is then passed to the dataset MPIO code, where it is sorted before being processed. This prevents the overhead of building a sorted skip list of pieces in cases where it is not needed, such as independent I/O or selection I/O. The algorithm for simultaneous multi-dataset I/O is contained in the

H5D__all_piece_collective_io() function which has been adapted from the previous *H5D__link_chunk_collective_io()* to be able to link pieces across all datasets. This function similarly constructs MPI datatypes that correspond to the dataspace selections for each piece, then combines these datatypes into an overall MPI datatype for the entire I/O operation.
The existing multi chunk MPIO pathway remains, and is used under the same circumstances as it was before. Multi chunk I/O refers to prcoessing the I/O request one chunk at a time. Since there is no need to handle multiple datasets at once if the operations cannot be combined at a low level, when using multi chunk I/O, the library processes one dataset at a time. Compressed dataset I/O is currently also processed one dataset at a time, though we may link compressed chunks across datasets in the future.

4.2.2 Selection I/O

Multi dataset I/O also includes full support for selection I/O. This is a new feature which allows the library to pass dataspace selections to the file driver instead of making a file driver call for each offset/length pair (or passing MPI datatypes through an undocumented side channel in the case of the MPIO driver). This allows any VFD to benefit from full knowledge of a noncontiguous I/O request. When performing selection I/O, the *H5D__read/write()* makes a layout read/write callback for each dataset, which adds to a global list of offsets, dataspaces, and buffers without performing I/O, then *H5D__read/write()* makes a single file driver selection I/O callback with these lists.

4.2.3 VOL

The VOL layer allows developers to re-implement HDF5 features by implementing custom callbacks for HDF5 API calls that access the HDF5 file. Since the multi dataset API functions do this, we have modified the VOL layer to handle multi dataset. We did this by adding the *count* parameter to the dataset *read* and *write* callbacks, and making the other parameters except *dxpl_id* arrays. In this way, the VOL callbacks are essentially always multi-dataset, with normal single dataset calls simply passing 1 for the count. This means that VOL developers will need to update their connectors. This is expected however, since we do not guarantee API compatibility between major versions of HDF5 for advanced "developer" symbols like the VOL struct.

4.3 New API functions

The two new functions, *H5Dread_multi()* and *H5Dwrite_multi()* are outlined here. Asynchronous versions of these functions, *H5Dread_multi_async()* and *H5Dwrite_multi_async()* are also being introduced.

4.3.1 H5Dread_multi()

The C API function description is shown below.

```
1.    herr_t H5Dread_multi(size_t count,
2.                         hid_t dset_id[],
3.                         hid_t mem_type_id[],
4.                         hid_t mem_space_id[],
5.                         hid_t file_space_id[],
6.                         hid_t dxpl_id,
7.                         void *buf[] /*out*/);
```

Parameters:

- count: the number of datasets being accessed (the length of the arrays).
- mem_type_id: array of memory type IDs.

- mem_space_id: array of memory dataspace IDs.
- file_space_id: array of file dataspace IDs.
- dxpl_id: dataset transfer property.
- buf: array of read buffers.

Return:

- A non-negative value if successful; otherwise returns a negative value.

The Fortran API is,

```
1.    SUBROUTINE H5Dread_multi_f(count, dset_id, mem_type_id, mem_space_id, file_space_id, buf, &
2.                               hdferr, dxpl_id)
3.      IMPLICIT NONE
4.
5.      INTEGER(SIZE_T),      INTENT(IN)                     :: count
6.      INTEGER(HID_T),       INTENT(IN), DIMENSION(1:count) :: dset_id
7.      INTEGER(HID_T),       INTENT(IN), DIMENSION(1:count) :: mem_type_id
8.      INTEGER(HID_T),       INTENT(IN), DIMENSION(1:count) :: mem_space_id
9.      INTEGER(HID_T),       INTENT(IN), DIMENSION(1:count) :: file_space_id
10.     TYPE(C_PTR),                      DIMENSION(1:count) :: buf
11.     INTEGER,              INTENT(OUT)                    :: hdferr
12.     INTEGER(HID_T),       INTENT(IN), OPTIONAL           :: dxpl_id
```

This routine performs collective or independent I/O reads from multiple datasets. In collective mode, all process members of the communicator associated with the HDF5 file must participate in the call. Each process creates the information required to perform each read in the arrays of parameters and passes the array through to *H5Dread_multi()*.

Brief description for internals after being called:

- Each process constructs an MPI-derived datatype describing the sections from multiple datasets in an HDF5 file to be read.
- All processes end up calling *MPI_File_read_at_all()* once each for collective I/O or *MPI_File_read_at()* once each for independent I/O.
- Each process tidies up and then returns the desired data into the buffer of the info[] array structure.

All processes are required to pass the same property values for the *dxpl_id*.
All datasets are required to be in the same file, and this file must be the same across all processes.
All processes must pass the same *count* and the same list of datasets. Other array parameters may differ.
All array parameters must have a length of *count*.
Refer to the example section for a better understanding of usage.
The same rule applies to *H5Dwrite_multi()*, detailed in the following section.

4.3.2 H5Dwrite_multi()

The API function description is as shown below.

```
8.    herr_t H5Dwrite_multi(size_t count,
9.                          hid_t dset_id[],
10.                         hid_t mem_type_id[],
11.                         hid_t mem_space_id[],
```

```
12.                         hid_t file_space_id[],
13.                         hid_t dxpl_id,
14.                         const void *buf[] /*out*/);
```

Parameters:

- count: the number of datasets being accessed (the length of the arrays).
- mem_type_id: array of memory type IDs.
- mem_space_id: array of memory dataspace IDs.
- file_space_id: array of file dataspace IDs.
- dxpl_id: dataset transfer property.
- buf: array of write buffers.

Returns:

- A non-negative value if successful; otherwise returns a negative value.

The Fortran API is,

```
1.    SUBROUTINE H5Dwrite_multi_f(count, dset_id, mem_type_id, mem_space_id, file_space_id, buf, &
2.                                hdferr, dxpl_id)
3.      IMPLICIT NONE
4.
5.      INTEGER(SIZE_T),      INTENT(IN)                     :: count
6.      INTEGER(HID_T),       INTENT(IN), DIMENSION(1:count) :: dset_id
7.      INTEGER(HID_T),       INTENT(IN), DIMENSION(1:count) :: mem_type_id
8.      INTEGER(HID_T),       INTENT(IN), DIMENSION(1:count) :: mem_space_id
9.      INTEGER(HID_T),       INTENT(IN), DIMENSION(1:count) :: file_space_id
10.     TYPE(C_PTR),                      DIMENSION(1:count) :: buf
11.     INTEGER,              INTENT(OUT)                    :: hdferr
12.     INTEGER(HID_T),       INTENT(IN), OPTIONAL           :: dxpl_id
```

This routine performs collective or independent I/O writes to multiple datasets. In collective mode, all process members of the communicator associated with the HDF5 file must participate in the call.
Each process creates the information required to perform each write in the arrays of parameters, and passes the array through to *H5Dwrite_multi()*.
Note that when overlapping selections are used, the data stored in the file for the overlapping regions is undefined (as is the case for H5Dwrite).

Brief description for internals after being called:

- Each process constructs an MPI derived type describing the sections from multiple datasets in an HDF5 file to be written.
- All processes ends up calling *MPI_File_write_at_all()* once each for collective I/O or *MPI_File_write_at()* once each for independent I/O.

All processes are required to pass the same property values for the *dxpl_id*.
All datasets are required to be in the same file, and this file must be the same across all processes.
All processes must pass the same *count* and the same list of datasets. Other array parameters may differ.
All array parameters must have a length of *count*.
Refer to the example section for a better understanding of usage.
The same rule applies to *H5Dwrite_multi()*, detailed in the following section.

4.4 Example cases

These examples are based on the assumption that using multi-read API on an HDF5 file with four datasets, ‘d1’, ‘d2’, ‘d3’ and ‘d4’. Using multi-write API would be practically identical. Pseudocode is used to show how the API can be used in a simplified manner focusing on this task’s scope.

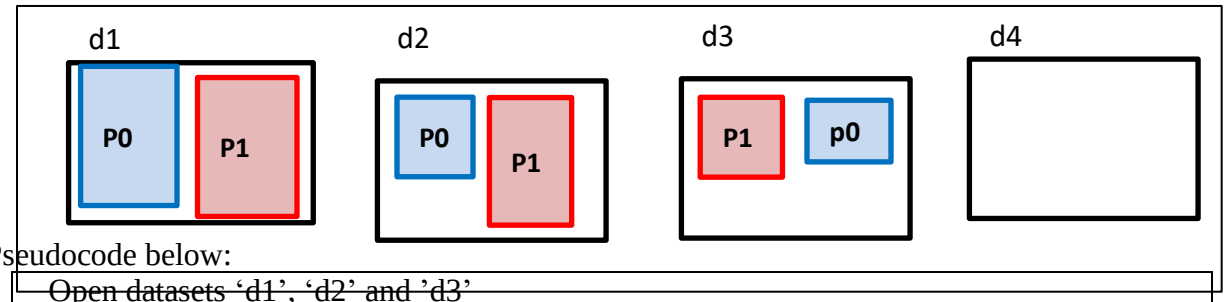
4.4.1 **Example1:** all processes read from the same datasets ‘d1’, ’d2’ and ‘d3’

Consider the following as an example running with two processes:

- Rank 0 process (P0) reads data portions from datasets ‘d1’, ‘d2’, and ‘d3’.
- Rank 1 process (P1) reads data portions from datasets ‘d1’, ‘d2’ and ‘d3’.

Chart view:

An HDF5 file



Pseudocode below:

```
Open datasets ‘d1’, ‘d2’ and ’d3’
Make selections from each dataset.
Set ‘dxpl’ for collective operation.
Set ‘mem_type_ids’, ‘mem_space_ids’, and ‘bufs’ arrays as appropriate.

size_t count = 3        /* three datasets */
If (mpi_rank == 0)     /* P0 */
   hid_t  file_space_ids[3]  =  { {d1’s P0 select}, {d2’s P0 select}, {d3’s P0 select} }

If (mpi_rank == 1)     /* P1 */
   hid_t  file_space_ids[3]  =  { {d1’s P1 select}, {d2’s P1 select}, {d3’s P1 select} }

H5Dread_multi (count, mem_type_ids, mem_space_ids, file_space_ids, dxpl, bufs)
```

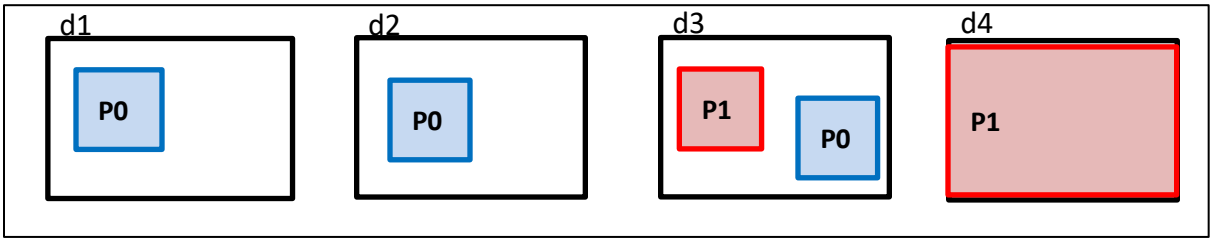
4.4.2 **Example2:** each process read from different datasets or none

Consider the following as an example running with three processes:

- Rank 0 process (P0) reads data portions from datasets ‘d1’, ‘d2’, and ‘d3’.
- Rank 1 process (P1) reads data portions from datasets ‘d3’ and ‘d4’.
- Rank 2 process (P2) does not read anything.

Chart view:

An HDF5 file



Pseudocode below:

```
Open datasets ‘d1’, ‘d2’,’d3’ and ‘d4’
Make selections from each dataset.
Set ‘dxpl’ for collective operation.
Set ‘mem_type_ids’, ‘mem_space_ids’, and ‘bufs’ arrays as appropriate.

If (mpi_rank == 0)       /* P0 */
   count = 3;             /* three datasets */
   hid_t  file_space_ids [3] = { {d1’s P0 select},  {d2’s P0 select},   {d3’s P0 select} }

If (mpi_rank == 1)      /* P1 */
    count = 2;           /* two datasets */
    hid_t  file_space_ids [2] = { {d3’s P1 select},   {d4’s P1 select} }

If (mpi_rank >= 2)      /* P2 */
    count = 0            /* no dataset access */
    hid_t  *file_space_ids = NULL

H5Dread_multi (count, mem_type_ids, mem_space_ids, file_space_ids, dxpl, bufs)
```

5 Limitations

While the API will work for any datasets and any I/O, the initial implementation will fall back to simply performing I/O for one dataset at a time in some cases. It will only perform simultaneous multi dataset I/O using MPI I/O and only in collective mode. It not initially perform simultaneous multi dataset I/O on compressed datasets. Note that any other conditions that break collective I/O (datatype conversion, not contiguous or chunked, etc.) will also break simultaneous multi-dataset I/O.

6 Future Consideration

In addition to addressing the limitations outlined above, according to some discussions, we may be able to consider developing H5Dcreate_multi(), H5Dopen_multi() and H5Dclose_multi() APIs in the future as separate tasks if necessary or requested by the user.

7 Code Repository

The latest version of the multi-dataset branch can be found at
https://github.com/HDFGroup/hdf5/tree/feature/multi_dataset

[1] Yang M and Koziol Q, 2006. Using collective IO inside a high performance IO software package—HDF5 Technical Report National Center of Supercomputing Applications
[2] Rob Latham, Chris Daley, etc., March 2012. A case study for scientific I/O: improving the FLASH astrophysics code, http://iopscience.iop.org/1749-4699/5/1/015001/article
[3] Qiao Kang, Scot Breitenfeld, Kaiyuan Hou, Wei-keng Liao, Robert Ross, and Suren Byna, December 2021. Optimizing Performance of Parallel I/O Accesses to Non-contiguous Blocks in Multiple Array Variables, 2021 IEEE International Conference on Big Data (Big Data), https://ieeexplore.ieee.org/document/9671638

Revision History

August 28, 2012:	Version 1 by Peter Cao. Circulated internally.
Sep 27, 2012:	Version 2: updated based on internal reviews.
Feb 15, 2013:	Version 3: Updated based on internal reviews. Revised APIs and related contents. The task entry is HDFFV-8313 in JIRA.
March 04, 2013:	Version 3.1: Updates based on internal reviews. More updates and add an example section.
March 07, 2013:	Version 3.2: Some minor updates. Add chart view in the example section.
March 12, 2013	Version 3.3: Some updates from an internal presentation on 03-08-2013.
March 21, 2013	Version 4: revised based on the comments from the internal presentation on 03-08-2013.
January 24, 2022	Version 5: Updated to reflect the status in early 2022.
February 15, 2022	Version 6: Updated based on internal reviews.
May 23, 2022	Version 7: Updated to reflect the latest state of development.
September 28, 2022	Version 8: Updated to reflect the latest state of development.