

Collective IO support inside HDF5

MuQun Yang

National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

Software Stack for Parallel HDF5 Application

Application

Parallel HDF5

MPI-IO(ROM-IO, etc.)

Parallel File System(GPFS, PVFS,
Lustre)

Hardware(Myrinet, infinite band etc.)

MPI-IO Basic Concepts

- **collective IO**

Contrary to independent IO, all processes must participate in doing IO.

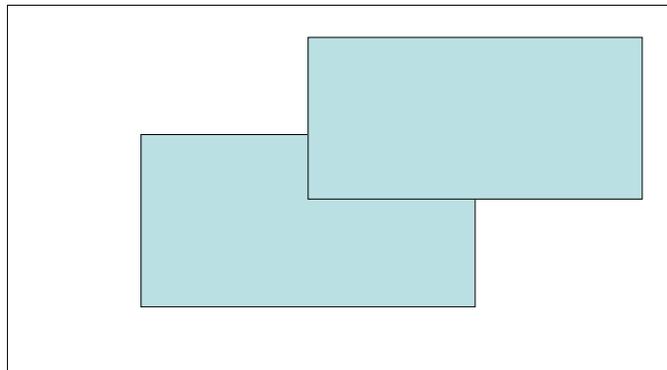
MPI-IO can do optimization to improve IO performance by using `MPI_FILE_SET_VIEW` with collective IO.

Previous collective IO support inside HDF5

- Only supports collective IO for **regular** selection with **contiguous** data storage

Non-regular selection:

- ▶ is a selection generated inside an HDF5 program with **MORE THAN** one `H5Sselect_hyperslab` routine call



Current collective IO support inside HDF5

- Both chunking and contiguous storage
- Both regular and non-regular selection
- For chunking storage, collective IO works per chunk.

MPI-IO Packages that don't support complicated MPI derived datatype

Platform	
IBM AIX	poe version: 3.2.0.20 or lower
SGI IRIX	irix older version 7.4 or lower
Linux (teragrid, altrix, tungsten mpich and mpi-lam)	mpich 1.2.5 or lower
Dec-OSF	Current version at PSC limeux

Software Stack for Parallel HDF5 Application

Application

Parallel HDF5

MPI-IO(ROM-IO, etc.)

Parallel File System(GPFS, PVFS,
Lustre)

Hardware(Myrinet, infinite band etc.)

Parallel IO performance really depends on all layers.

Future direction

- More optimization to improve the collective IO performance inside parallel HDF5
- IO performance study