Experiences with ROMIO

from a PnetCDF developer's point of view

Wei-Keng Liao, ECE Department, Northwestern University

MPICH: A High-Performance Open Source MPI Library for
Leadership-class HPC Systems
2024 CASS Community BOF Days, June 12, 2024

Northwestern | McCORMICK SCHOOL OF ENGINEERING

PnetCDF

- A parallel I/O library for accessing NetCDF classic file format
  - Relies on MPI-IO
- NetCDF classic file format
  - NetCDF files are popularly used in climate research community
  - Header and data sections
  - Data objects: dimensions, variables, attributes
- PnetCDF APIs
  - Metadata
  - Read and write subarrays of variables
  - Blocking and non-blocking reads and writes

MPI APIs used in PnetCDF

- MPI-IO
  - File open, close, seek, sync, set view
  - Collective and independent I/O APIs
- MPI communication
  - Metadata consistency check: file header
  - Mostly MPI_Allreduce, MPI_Bcast
- MPI derived datatypes to define fileview
  - Commonly used data partitioning patterns
  - Call to set fileview is collective, a problem for switch between independent and collective modes

I/O request aggregation

- User intent for I/O
  - Saving multiple variables in the file during data checkpointing.
  - Safely store all variables, not individual variables, before returning to the computation phase.
  - Such intent can be better realized by high-level libraries (through requestion aggregation feature in PnetCDF, HDF5, and PIO)
- PnetCDF
  - Use nonblocking APIs to post requests + a wait_all call later to flushed all out
- HDF5 multi-dataset APIs
  - H5Dwrite_multi — allows a single call to write multiple variables
- PIO
  - Two aggregation options: subset and box rearrangers
- Challenge for MPI-IO
  - Aggregated amount can become very large. So is the metadata describing the requests.

Challenges — I/O hints

- cb_nodes
  - Number of I/O aggregators (a subset of processes does I/O)
  - Default: one per compute node (GPFS), same as number of OSTs (Lustre)
- cb_buffer_size (default: 16 MiB)
- striping_factor
  - Too small yields poor performance, too large increases interference from other users
- striping_unit
  - Large, contiguous requests prefer large striping size
- Data sieving (romio_ds_write, romio_ds_read, romio_cb_ds_threshold)
  - I/O aggregators check holes in its file domain to determine whether data sieving is necessary, and then read-modify-write

Challenges — future ROMIO improvement

- Memory footprints
  - Fileview and user buffer datatype are flattened into offset-length pairs
  - Internal buffers allocated for storing these pairs can become significant
- Excessive number of memcpy calls
  - Communication phase in collective I/O packs data into contiguous buffers before sending
  - For large number of offset-length pairs, memcpy calls become expensive
- Communication in the two-phase I/O
  - MPI_Isend vs. MPI_Issend — MPI_Issend can prevent message queues from being overwhelmed