Advanced Message-Passing Programming

Alternative Parallel IO Libraries

Reusing this material

epcc

THE UNIVERSITY OF EDINBURGH

Overview

- Issues with MPI-IO
- HDF5
- NetCDF
- ADIOS2
- Summary

MPI-IO Issues

- Files are raw bytes
  - no header information
  - storage is architecture-specific (e.g. big / little-endian floating-point)
- Difficult to deal with in other codes downstream
  - user must write their own post-processing tools
- But ...
  - it can be very fast!

Solution

- For functionality
  - define higher-level formats
  - include metadata, e.g. “this is a 4x5x7 array of doubles”
  - enables standard data converters, browsers, viewers etc.
- For performance
  - layer on top of MPI-IO
- Many real applications use higher-level formats
  - understanding MPI-IO will enable you to get performance as well

HDF5

- "**Hierarchical Data Format** (**HDF**) is a set of file formats (**HDF4**, **HDF5**) designed to store and organize large amounts of data." (Wikipedia)
  - data arranged like a Unix file system
  - self-describing
  - hierarchical
  - can use MPI-IO

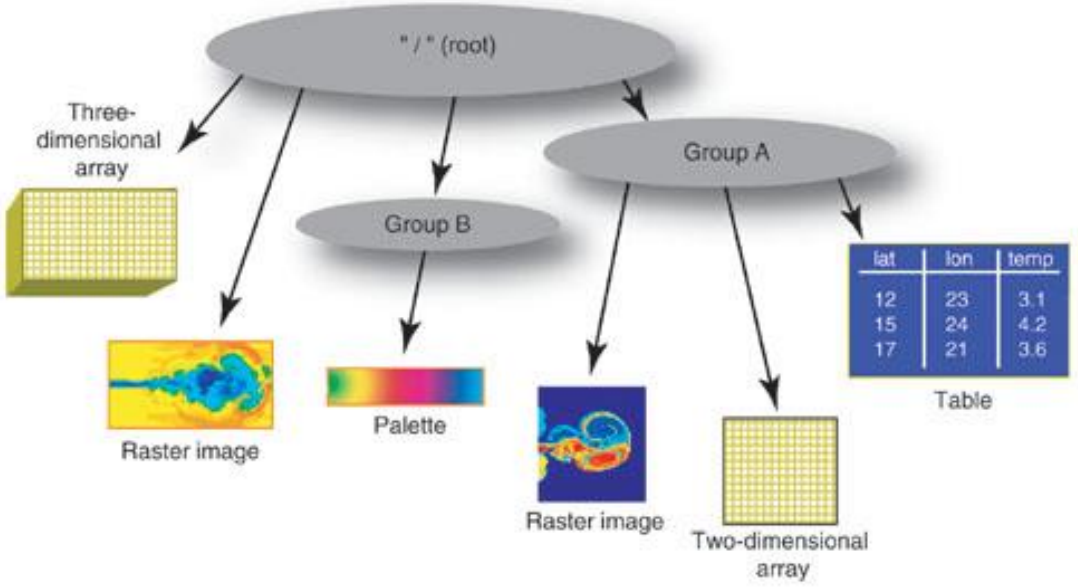


image taken from www.hdfgroup.org

Parallel HDF5 (Fortran)

- Approach much like MPI-IO
  - describe global dataset
  - s describes its local portion(s) of the g

MPI_ORDER_FORTRAN

global data, encodes sizes

```
CALL h5sselect_hyperslab_f(filespace, &
      H5S_SELECT_SET_F, offset, &
      count, error)
```

starts

subsizes

- Then call collective write
  - hyperslabs can be merged to create global file
  - actual file IO done through MPI-IO
  - important to choose collective IO

NetCDF: Network Common Data Form

- "a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data.." (Wikipedia)
  - more restricted than HDF5
  - common in certain communities
    - climate research
    - oceanography
    - GIS ...
- Rich set of tools
  - data manipulation
  - visualisation
  - ...

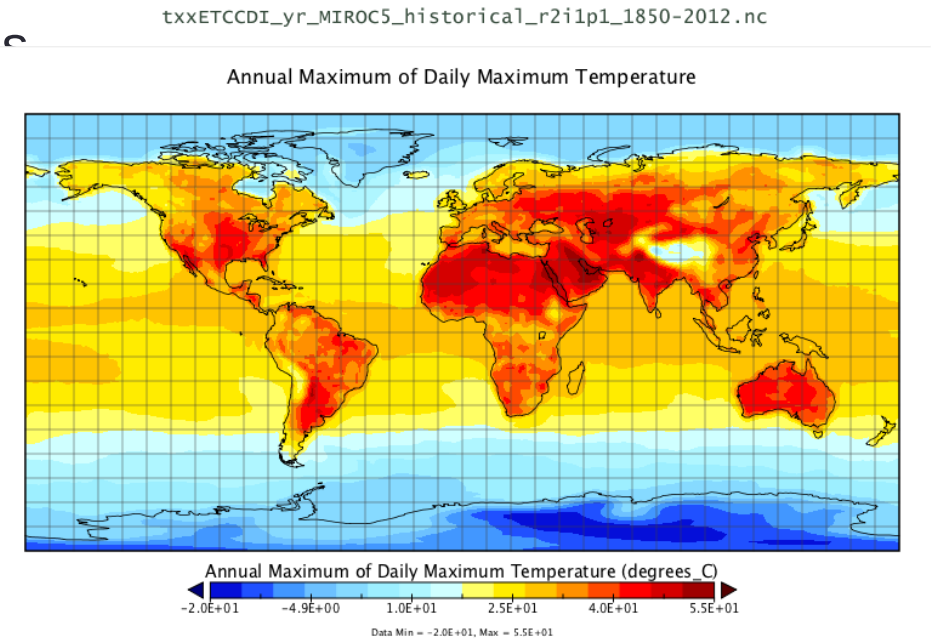
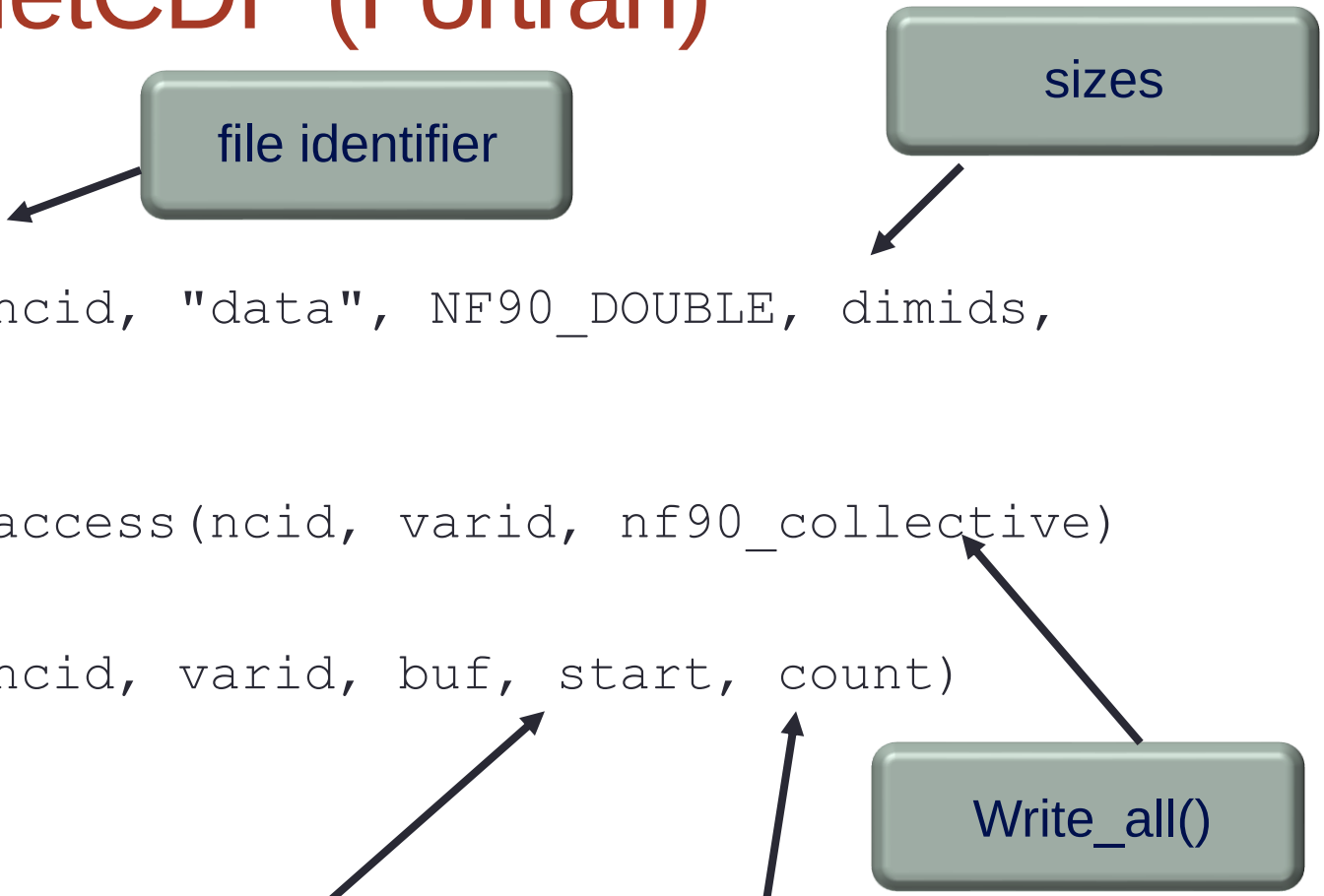


image taken from http://live.osgeo.org

Parallel NetCDF (Fortran)



ADIOS2

- A recent IO parallel library https://adios2.readthedocs.io/
  - can output using native MPI-IO or HDF5
  - also supports its own formats, e.g. BP5 (binary-pack v5)
- Same overall approach
  - each process defines what portion(s) of global data it owns
  - call read/write routines
- Much more configurable at runtime via “config.xml”
  - e.g. no need to recompile to switch MPI-IO to BP5
- Not yet part of benchio but colleagues have seen potential benefits from BP5 format in other codes
  - writes multiple files under the hood (with associated metadata)
  - may get round the limits of MPI-IO which has a single shared file

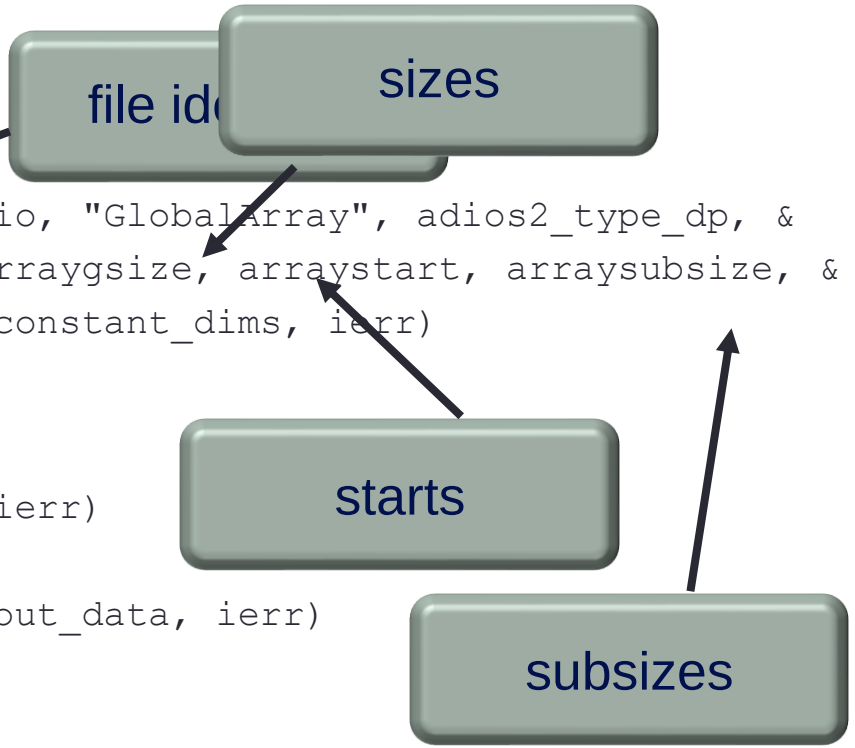
ADIOS2 (Fortran)

```
! Define the global array
  call adios2_define_variable(var_g, io, "GlobalArray", adios2_type_dp, &
                              ndim, arraygsize, arraystart, arraysubsize, &
                              adios2_constant_dims, ierr)

! Begin ouput step
  call adios2_begin_step( bp_writer, ierr)

  call adios2_put( bp_writer, var_g, out_data, ierr)

! End the output
  call adios2_end_step(bp_writer, ierr)
```



How is ADIOS2 different?

- For MPI-IO, HDF5 and NetCDF
  - output file is independent of process count or decomposition
  - requires a lot of data movement
  - locking issues as multiple aggregators write to the same shared file
- For ADIOS2 “BP5” format
  - under the hood, writes one file per aggregator
    - no data movement
    - no locking issues
  - also stores metadata to make this invisible to the user
  - most file options are stored in an XML file
    - can change output format to HDF5 without recompilation

Summary

- MPI-IO may seem a little low-level
  - but is fundamental building block of parallel IO on most systems
- Higher-level formats layer on top of MPI-IO
  - to benefit from performance work by vendors, Lustre etc.
- Common formats are HDF5 and NetCDF
  - both supported on ARCHER2
  - you might also want to look at the newer ADIOS2 library
- Understanding MPI-IO performance is key to getting good performance for HDF5 and NetCDF