Parallel IO in CESM

Jim Edwards
NCAR CGD CSEG
PIO Contributors

- John Dennis (NCAR)
- Ray Loy (ANL)
- Allison Baker (NCAR)
- Yannick Polius (NCAR)
- Steve Goldhaber (Tech-X)
The PIO library

- Rearranges data from compute decomposition to IO friendly block decomposition
- Allows the use of several backend library options through a single API
- Allows a separation of concern between model developers and IO software layer
- Provides options to tune IO performance based on application and hardware requirements
- Allows memory scalability even for single task IO
New PIO Library

Serial use of PIO
PIO Backend Libraries

- MPI-IO Binary
- NetCDF-3
  - Serial IO reads and writes from a single task
- Parallel NetCDF
  - NetCDF-3 format using MPI-IO for parallel performance
- NetCDF-4
  - NetCDF-4/HDF5 format parallel
  - Serial IO HDF5 format with data compression
PIO options in CESM

- PIO_NUMTASKS
- PIO_STRIDE
- PIO_ROOT
  - The PIO task layout is related to the component compute task layout by:
    \[ \text{iotask}(i) = \text{pio_root} + (i-1) \times \text{pio_stride} \]
    where \( i = 1, \text{pio_numtasks} \)
- PIO_TYPENAME (netcdf, pnetcdf, netcdf4p, netcdf4c)
  These variables are also available on a per component basis (eg OCN_PIO_ROOT, ATM_PIO_TYPENAME)
PIO options in CESM

- PIO_BLOCK_SIZE (default 1MB) determines the optimal contiguous blocksize in IO decompositions
PIO options in CESM

- PIO_BUFFER_SIZE_LIMIT (default 100MB)

The size of the per IO task internal data buffer for pnetcdf non-blocking write operations (pnetcdf only, not supported on Bluegene systems).
New PIO Features

- Parallel data compression with VDC2
  - Planned for use with The Geophysical High Order Suite for Turbulence (GHOST) in an upcoming ASD project
  - A wavelet-based, gridded data format supporting both progressive access and efficient data subsetting
  - Currently implementation has limited functionality, not well suited for CESM
    - A lot of metadata functionality is missing
4096^3 Homogenous turbulence simulation
Volume rendering of original enstrophy field and 800:1 compressed field

Original: 275GBs/field
800:1 compressed: 0.34GBs/field

Data provided by P.K. Yeung at Georgia Tech and Diego Donzis at Texas A&M
Future PIO features

- Async IO capability
  - IO handled on independent tasks allowing overlap of computation and write operations
  - Reduces IO performance variability as an issue for model performance
  - Requires good memory buffer management on IO tasks
Other PIO users

- ESMF
- DART
- MPAS
- GHOST
Questions?