CMMAP Data Services Update

Karen Schuchardt
UC Berkeley
January 2011

Pacific Northwest
NATIONAL LABORATORY

PNetCDF

- defines cdf5 format based on cdf3 with fixes for large variables
- New psuedo non-blocking IO that improves performance
- Support for setting the header padding size (important for basic metadata editing)
- Integers are 4 bytes; no longs
- In theory, can be specified as the mechanism to use via the NetCDF4 interface
- Up to 10GB/s franklin (of possible 16GB)

- Pretty robust and good performance (version 1.2)
- Data can only be processed with pnetcdf-based analysis tools
  - Future support model not clear

Pacific Northwest
NATIONAL LABORATORY

NetCDF4

- Based on widely used HDF5 data model
  - HDF5 format is a rich data format with filesystem-like constructs
  - Lots of tunable features like compression, chunking
  - Can be restricted to netcdf data model constructs

- Fortran interface (still) has 32bit restriction
- Not stress tested
- Performance currently lags PNetCDF

- New DOE Exascale project to optimize HDF5
  - LBNL (Prabhat – PI), HDF5 Group, PNNL
  - Being done in context of real applications
    - GCRM
    - Pore scale simulators (groundwater, physics)

Pacific Northwest
NATIONAL LABORATORY

Runs

- Z anelastic model (run to date)
  - 15km, 26 interfaces,18 days, 3 hourly
  - Jablonowski test case, 10000 processors on franklin
  - About 8 variables ~= 44GB
- Z anelastic model (by spring/summer)
  - 4 or 7 km (or both), 36? interfaces,18 days, 3? hourly
  - Initial conditions TBD, ~10000/40000 processors on hopp2
  - Will have physics added (Don Dazlich)
    - 8? variables ~= **TBD**

NOTE: time free only until April

Pacific Northwest
NATIONAL LABORATORY

Analysis - ParCal

- New DOE Project for Parallel Data Analysis Tools
- ANL (Rob Jacob PI), NCAR, PNNL, Sandia
- Major outcomes
  - Version of NCL which transparently turns current NCL data arrays and operations into parallel operations
  - Climate specific compression to be tested in PNetCDF