Parallel Analysis Tools and New Visualization Techniques for Ultra-Large Climate Data Sets

ParVis is building a new, big-data, parallel, open source analysis library and DAV application that will solve several problems at once…

ParGAL: The Parallel Gridded Analysis Library

ParGAL provides data-parallel and multi-grid capable versions of many typical analysis functions on gridded data sets. ParGAL functions will also be able to operate on the native grids of the output. ParGAL is built from several existing libraries:

- Mesh-Oriented DataBase (MOAB)
- Intrepid: Interoperable Tools for Rapid Development of compatible Discretizations

ParGAL uses ParGAL, MOAB and NCL graphics to create a parallel version of NCL.

ParNCL: Parallel version of NCL

NCL (NCAR Command Language) is a widely used scripting language tailored for the analysis and visualization of geoscientific data.

ParNCL uses ParGAL, MOAB and NCL graphics to create a parallel version of NCL.

Additional ParVis projects

Task-parallel diagnostic packages

ParGAL and ParNCL are long-term projects. ParVis is providing more immediate speed-up to CESM Working Group Diagnostic packages by implementing them in the task-parallel language Swift.

- Swift is a parallel scripting system for grids and clusters
- Swift is easy to write: simple high-level C-like functional language
- Swift is easy to run: a Java application. Just need a Java interpreter installed.
- Swift is fast: Karajan provides Swift a powerful, efficient, scalable and flexible execution engine.

AMWG Diagnostic Package Results:

(a) Calculate the climatological mean files for 5 years of 0.10-degree (up-sampled data from a 0.25-degree) CAM-SE cubed sphere simulation. This was run on Fusion, a cluster at Argonne, on 4 nodes using one core on each.
(b) Compare two data sets: each 30 years, 1-degree monthly average CAM files. This was run on one data analysis cluster node on mirage at NCAR.
(c) Compare 10 years of 0.5-degree resolution CAM monthly output files to observational data. This comparison was also ran on one node on mirage.

OMWG Diagnostic Package Results:

In the above timings, the OMWG Diagnostic Package was used to compute the averages of 10 years of 1 degree POP data (clim files). Then NCL was used to create the plots. This was done on 4 compute nodes on lens running a maximum of 8 tasks per node. The original was ran on 1 lens compute node.

Efficient Compress of High-Resolution Climate Data

Climate model output could overwhelm available disk space and, when moving, network bandwidth.

2-phase Lossless compression: First phase predicts next value based on previous. Second phase encodes next value with entropy-based encoding.

Experiments with various chunk sizes because chunk size can affect compression ratios and degree of parallelism.

Small chunk size: low compression ratio, high degree of parallelism
Big chunk size: high compression ratio, low degree of parallelism

Further Work:

- Experiment with multi-layer compression
- Increase throughput with pipeline parallelism between read of compressed data and decompression
- Incorporate compression in to ParNCL.

Results:

ParGAL: Parallel versions of vorticity and divergence calculations implemented. New general versions of streamfunction and velocity potential implemented. All applicable to regional and global domains. Parallel version of advantage, min, median, median along any dimension. Extensive test suite and nightly build/test. ParNCL: parallel addfsses working (NetCDF only). Simple math operations (sin, cos, tan, etc.) working. Addition and subtraction of parallel multi-dimensional arrays supported. Beta release available late August. Watch website.