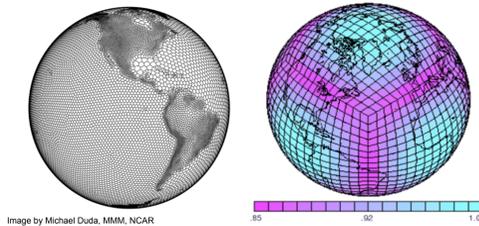


# Parallel Analysis Tools and New Visualization Techniques for Ultra-Large Climate Data Sets

Robert Jacob, Jayesh Krishna, Xiabing Xu, Sheri Mickelson, Tim Tautges, Mike Wilde, Rob Latham, Ian Foster, Rob Ross, Mark Hereld, Jay Larson (Argonne National Laboratory); Pavel Bochen, Kara Peterson, Mark Taylor (Sandia National Lab.) Karen Schuchardt, Jian Yin (Pacific Northwest National Lab.); Don Middleton, Mary Haley, David Brown, Richard Brownrigg, Wei Huang, Dennis Shea, Mariana Vertenstein (NCAR), Kwan-Liu Ma, Jinrong Xie (UC-Davis)

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

Use data-parallelism to allow larger files to be analyzed

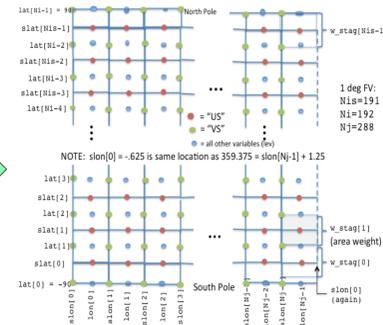


The growing number of grids used in climate modeling. Spherical Centroidal Voronoi Tessellation (left), atmosphere cubed sphere grid (right)

## Growing Size of Climate Model Output

Model	Resolution	Single 3D variable	Single 2D variable	Single history file	1 year of monthly output	100 years of monthly
CAM	HOMME at 0.125 degrees	616 MB	25 MB	24 GB	288 GB	28.8 TB
CSU	GCRM 4km horizontal, 100 levels	16 GB	50.3 GB	571 GB	6 TB	.6 PB

Reproduce discretizations used to generate the output for more accurate analysis



Enable analysis directly on new unstructured and semi-structured grids.

**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):

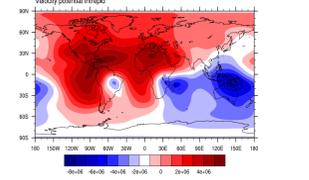
MOAB is a library for representing structured, unstructured, and polyhedral meshes, and field data on those meshes

### Intrepid: Interoperable Tools for Rapid development of compatible Discretizations

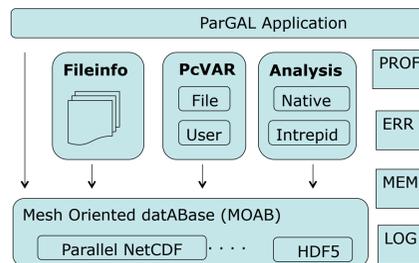
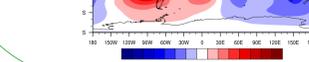
An extensible library for computing operators on discretized fields  
Will compute div, grad, curl on the unstructured or structured grids maintained by MOAB.

### PNetCDF: NetCDF output with MPI-IO

Velocity Potential from ParGAL (Using Intrepid's Finite Element Method)

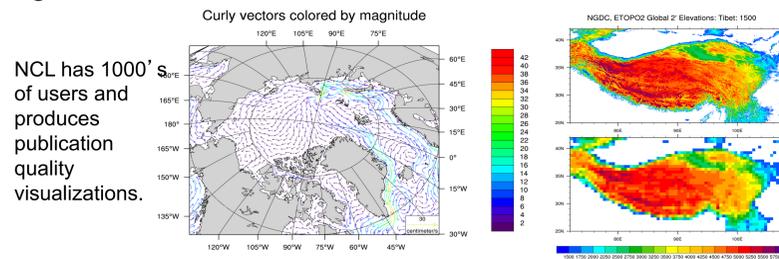


Velocity Potential in NCL (Using Spherical Harmonic-based function uv2sfvpg)

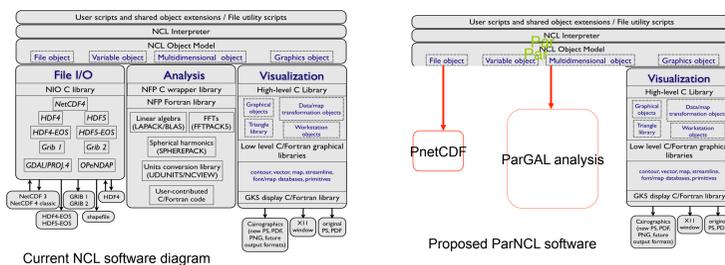


## ParNCL: Parallel version of NCL

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.



Using ParNCL requires a parallel environment:

Now: `> ncl myscript.ncl`

With ParNCL: `> mpirun -np 32 parcn1 myscript.ncl`

## Additional ParVis projects

### Task-parallel diagnostic packages

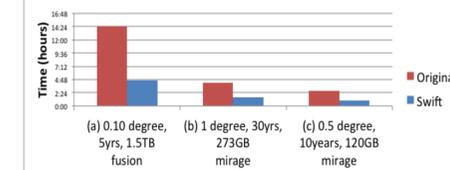
ParGAL and ParNCL are long-term projects. ParVis is providing more immediate speed-up to CESM Working Group Diagnostic packages by implementing them in the task-parallel language Swift.



- 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.

**RESULTS:** We have written a Swift version of the AMWG diagnostic package and is included in the amwg\_diag5.3 release. We have also re-written the OMWG diagnostic package (with Swift AND in all-NCL) which will be released soon.

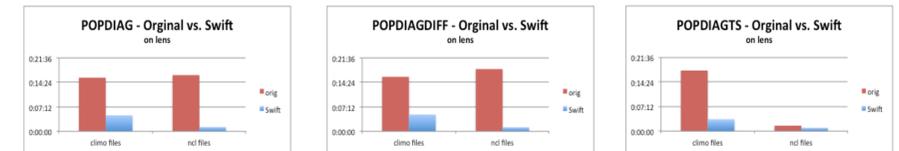
### Original vs. Swift Timings for Various Datasets



### AMWG Diagnostic Package Results:

- 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.
- 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.
- 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:

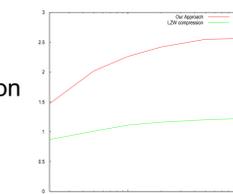


In the above timings, the OMWG Diagnostic Package was used to compute the averages of 10 years of 1 degree POP data (climo files). Then NCL was used to create the plots. This was done on 4 compute nodes on lens running a maximum of 8 tasks per node. The original was ran on 1 lens compute node.

## Efficient Compression 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.



**Lossy compression:** Error for each value is bounded. We can achieve a compression of around 10 when error bound is 0.1%

### Further Work:

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

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

### 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 versions of average, min, max, median along any dimension. Extensive test suite and nightly build/test. **ParNCL:** parallel addfiles 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.**