Efficient Extraction of Regional Subsets from Massive Climate Datasets

Jeff Daily*, Karen Schuchardt* and Bruce Palmer*
* Pacific Northwest National Laboratory
jeff.daily,karen.schuchardt,bruce.palmer@pnl.gov

***Abstract*—**
Parallel programming is needed to analyze the vast amount of data produced by today's climate models, however the majority of climate analysis software remain at the scale of workstations. Many climate analysis tools adequately process regularly gridded data but lack sufficient features when handling unstructured grids. This paper presents a data-parallel subsetter capable of correctly handling unstructured grids while scaling to the size of modest analysis clusters. The approach is based on the partitioned global address space (PGAS) parallel programming model and one-sided communications. The paper demonstrates that IO remains the single greatest bottleneck for this domain of applications and that parallel analysis of climate data succeeds in practice.

I. Introduction

Parallel programming is needed to analyze the size of output data produced by today's climate models. [1] A single snapshot of Randall's Global Cloud Resolving Model will produce terabytes of data [2]; the analysis of even a modest time series of this data will quickly overwhelm today's software and traditional climate analysis systems. For these data sizes, I/O bandwidth represents the single greatest bottleneck for analysis tools. Parallel software leveraging parallel file systems must be used to process this data, however current climate analysis tools are at most task parallel and rely on a single data reader. [3] [4] [5]

Many climate analysis tools robustly handle the manipulation and display of regularly gridded data. However, these same applications lack sufficient features when handling unstructured or irregular grids such as the geodesic or cubed sphere [6]. Unstructured grids are gaining popularity, further widening the gap between current software and these types of models. For unstructured grids it is necessary to provide more information about the topology of the grid and maintain the integrity of this topology information in the face of data culling.

Regular grids allow for the topology to be implicitly defined by how the data is stored; coordinate variables are generally monotonic and cell neighbors are adjacent both logically and in memory. These assumptions allow for operations over regular grids which are otherwise more difficult to perform over unstructured grids. In the case of partitioning these grids for data parallel processing, unstructured grids will often have more of the logically adjacent cells scattered across memory partitions than in the regular case.

Subsetting is a fundamental capability for any analysis tool and allows users to operate over the regions of the data with which they are interested. The subsetting operation is useful as part of a larger operation over the data, such as for regional averages, but is also useful to post-process data into a new dataset such that the cost of subsetting can be amortized across future operations over the same region. Further, as the size of datasets grow subsetting is important to reduce the data to a size that traditional analysis tools are capable of handling.

In this paper, we present a parallel tool for subsetting very large geodesic climate data in parallel while preserving the explicit topology. The code is built using the Global Arrays (GA) toolkit which provides an efficient and portable "shared-memory" programming interface for distributed-memory computers and features truly one-sided communications. [7] GA traditionally deals with dense arrays, however its sparse matrix operations as well as its one-sided operations allow for efficient subsetting over unstructured grids.

The primary contributions of the paper are:

- A parallel subsetter of geodesic data based on the partitioned global address space programming model and one-sided communications
- A novel algorithm for the maintenance of unstructured grid data
- A novel algorithm for the subset and even distribution of unstructured grid data
- Evaluation showing IO to be the greatest bottleneck in scaling these types of applications

The paper is organized as follows. Section II describes the requirements while III describes the design of the subsetter, its algorithms, and how GA's unique features were leveraged. Section IV briefly describes how the subsetter was implemented. Section V presents our experimental design and the performance characteristics of the subsetter on nearly full-scale data set sizes of model data up to a resolution of 4Km. We present the capabilities under development as well as the capabilities we would like to see in section VI. Finally, section VII presents our conclusions.

II. Requirements

The requirements for our software stem from the growing need for parallel analysis in the domain of climate science [1]

Wei-keng Liao

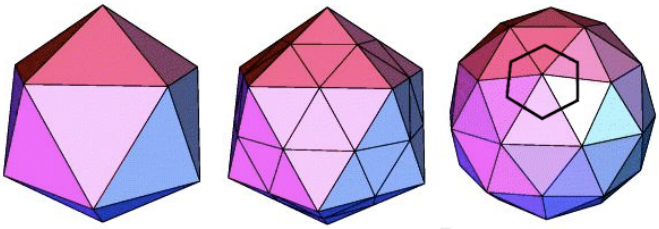

Fig. 1. First stage generation of the geodesic grid

but also from the use of the geodesic grid.

A. Geodesic Grid

Until recently, climate and weather models have primarily been simulated on structured grids that divide the latitude and longitude axes in even increments, resulting in logically structured simulation grids. Standard conventions for describing this data in the NetCDF data model have been formalized by the Climate and Forecast (CF) conventions. CF defines conventions and metadata standards that enable both human and computer interpretation of the data. Human interpretation is supported through the use of standard names while definition of spatial and temporal properties of the data have enabled an extensive set of tools for data manipulation and display such as [5], [8], and [9]. The CF Conventions have been evolving to support many variations of structured grids including Orthographic, Polar stereographic, Transverse Mercator, and many others .

The GCRM uses a geodesic grid. The geodesic grid is created by recursively bisecting an icosahedron of 20 triangular faces and twelve vertices and projecting the resulting faces onto a unit sphere. The resulting vertices represent the centers of hexagonal grid cells with the exception of twelve pentagons (the centers of the original twelve vertices.) See 1 for an example of the first stage of bisection and projection, followed by the definition of one of the hexagons. Further details can be found in [10].

From the previous description, it can be seen that the geodesic grid used by the GCRM is fairly regular. However, the horizontal dimension has some important properties in common with unstructured grids: the grid coordinates are not monotonic and simple conventions are not available for identifying the neighbors of all cells. As a consequence, it is necessary provide more information about the topology of the grid. Other unstructured grids such as triangular, cubed sphere [6], and arbitrary unstructured polygons are also being applied to various models. There is a recognized need to extend the CF conventions to unstructured grids so that general data analysis, regridding, and display tools can be developed. As yet, no such standard exists. Balaji has cataloged a number of unstructured grids [11] and proposes a tiling approach to describing grids. The tiling approach supports embedded and nested grids but has been slow to gain momentum, in part due to its complexity. In the ocean modeling community, workshops to discuss standards have taken place [12], however, these efforts are still in the early stages of development and acceptance. Lacking a standard, development of general purpose tools has been slowed. However some preliminary tools and approaches have progressed by focusing on the in-memory operations and abstracting the data loading from the processing [13].

Although the geodesic grid is fairly structured, we choose to represent it in an unstructured way. Each of the grid's cells, corners, and edges are uniquely indexed from zero. For a given positive integer R, there are $N = 10 \times 2^{2R} + 2$ cells, $C = (N - 2) \times 2$ corners, and $E = (N - 2) \times 3$ edges. Increasing the value of R increases the resolution of the model. For example, a value of $R = 10$ is approximately 8Km while $R = 11$ is approximately 4Km. The cells, corners, and edges are represented as dimensions within a NetCDF file.

The horizontal topology describes the connectivity relationships between cells, nodes, and edges, all of which may have associated 3D data. The topology consists of three primary arrays: a mapping between cells and cell corners, a mapping between cells and cell edges, and a mapping between edges and corners. Because neighbor lists are important for visualization programs but difficult to generate in the general case, they are included as part of the topology as the `cell_neighbors(cells,neighbors=6)` variable. The vast majority of cells are hexagons, so we have `cell_corners(cells,cellcorners=6)`, `cell_edges(cells,celledges=6)`, and `edge_corners(edges,edgecorners=2)`. For the twelve pentagons, the last values in these arrays are repeated but could just have easily been set to a negative index.

The horizontal geometry describes the longitude and latitude location of each object. That is, there is one latitude and one longitude array for each of the topology objects (cell centers, corners, and edges).

The vertical grid consists of layers sandwiched between interfaces with the number of layers equal to the number of interfaces minus 1. Because grid variables are associated with both locations, layers and interfaces are defined as dimensions. They can be thought of as two distinct vertical grids from a representation standpoint.

The data model is designed to support efficient model output, fully describe the grid topology, and provide the sufficient information for tessellation to triangles for 3D visualization. The approach taken with the model is to adopt the CF conventions to the extent possible and adopt early ideas circulating within the community. However, as many of the details have not been decided, custom data analysis tools are currently required.

B. Data Parallelism

Larson, Ong, and Tokarz note that the current popular climate data analysis packages remain single-processor applications which lack the memory required to handle the large data volumes as well as the processing power to analyse the data in a timely fashion. [1] Although they emphasize using OpenMP as a first step toward parallism, we instead emphasize

using the message-passing model first. Well designed message-passing libraries such as MPI may already take advantage of shared-memory parallelism within a compute node or multi-core desktop computer.

The data parallelism offered by libraries such as MPI or Global Arrays is absolutely necessary to handle the size of data of modern climate models. An edge data variable of the geodesic grid at an approximate resolution of 2Km and 100 levels is nearly $10 \times 2^{2R} \times 3 \times 100 \times 4\,\text{bytes} \approx 50\,\text{gigabytes}$ in size. Even a modest number of these variables will surpass the memory available in most desktop and even some small clusters.

C. Fast IO

For data of this size, efficiently reading from and writing to disk requires the use of parallel IO libraries such as the Parallel NetCDF [14] or the HDF5/NetCDF4 [15] [16] libraries, both of which are in turn built on top of the MPI-IO libraries [17]. Model output is stored in netCDF [16] files, a format for storing array-oriented machine-independent data.

D. Dataset Abstraction

Model output is often distributed across many files for a given model run. There are any number of schemes for organizing so many files e.g. one variable per file with multiple timesteps per file, separating out the grid into a separate file, one timestep per file with multiple variables. The reconstitution of these files into a logical set of variables and metadata is an established practice. [18] [19] We emphasize that the aggregation of files into an abstract dataset is required in order to operate on the data itself. Operations on a dataset are more intuitive than needing to know the addling details of which files hold which variables.

E. Maintenance of Topology Variables

Regular grids such as the cartesian, rectilinear, or curvilinear lend themselves to representations as multidimensional arrays such that logically adjacent cells are either adjacent in memory or can be located via a shape-based index calculation. Although some attempt is made to keep logically adjacent cells nearby in memory, geodesic grids do not have the luxury of using relatively simple shape-based index arithmetic to locate neighbors. The topology variables mentioned in II-A are unique to our grid. When a subset occurs, the indices must be updated to reflect the remaining corners, edges, and cells.

F. Maintain Integrity of Entire Grid Cells

Howe and Maier detail the properties of well-formed grids in [13]. Proper subsets should maintain the same well-formed properties of the original in order to remain useful to further analysis. Therefore, the cell and its surrounding corners and edges must remain intact during a subset.

III. Design

In this section we describe the design of our classes and algorithms based on the requirements established in II.

A. How to Run the Subsetter

The success of the NetCDF Operators and similar tools demonstrate the need for user-ready applications for the analysis of their data while the success of tools such as CDAT validate the need for a scriptable interface and customization of basic and advanced operators. We plan to provide both the scriptable interface as well as a set of predefined tools.

The subsetter is the first in a series of parallel command-line tools based on unstructured grids and the PGAS programming model. It takes arguments specifying specific variables (-v) or dimension ranges (-d) to extract, or at a higher level a latitude and longitude bounding box (-b). In this way it is most akin to the NetCDF Operators' "kitchen sink" application. Example usage looks like:

- mpiexec -np 128 subsetter -b 20,-20,160,90 -v vorticity january.nc february.nc MJO_vorticity_janfeb.nc
- mpiexec -np 64 subsetter -b 90,0,180,-180 -d levels,1,5 geopotential.nc

B. Dataset Abstraction

The subsetter minimally supports two forms of input file aggregation, either across a specified dimension e.g. time or by taking the union of all input files such that duplicate dimensions and variables within later files are ignored. These forms of aggregation are modeled after what is available when using NetCDF Markup Language. [18] NcML input is not directly supported at this time but is planned for a future release.

TODO

C. Parallel IO Abstraction

The use of Parallel NetCDF was selected because work on the HDF5/NetCDF4 was incomplete at the time of development.

TODO

D. The Global Arrays Library

TODO PGAS

TODO ONE-SIDED COMMUNICATION

TODO GA DOES IT ALL

The subsetter was built using the GA library for the wealth of features it provides which are tailored to our problem domain. GA provides a distributed dense multidimensional array programming abstraction and the data we will be operating over is stored as dense arrays within NetCDF files. It should be noted that dense distributed arrays would also work well for regularly gridded data. However, due to the use of unstructured grid data, the algorithm for subsetting the data will look quite different than for the structured case. Recall that for unstructured grids, logically adjacent cells are not necessarily adjacent in memory. In order to evenly distribute a subset, a single process will need to send a varying amount of data to any number of other processes. Certainly a collective operation could be considered, but GA provides the necessary functionality without needing any explicit cooperation from any other process. Any given process will simply put the

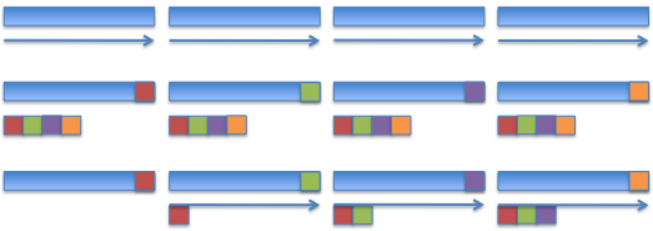

Fig. 2. Distributed partial sum on a 1-D array

section of the subset into the remote process's memory. It is unclear how point-to-point communications could mimic the ease of use of the interface provided by GA's one-sided operations.

Each dimension of the data has two arrays associated with it, a bitmask and an integer array representing the new indices of the dimension in case of a subset. For instance, if any of the bits are turned off, the corresponding index array will have negative values. The remaining values of the index array will increase monotonically, skipping the negative or masked indices. The bitmasks are generated based on a rectangular latitude and longitude region specified on the command-line, or by specifying one or more indices of a dimension to select. Although a rectangular region is currently used for simplicity, once translated the bitmasks allow for arbitrary subsets to be defined. These bitmasks are then used to evenly distribute the resultant subset across all processes. Note that these bitmask and associated index arrays are one-dimensional and distributed.

REMOVE There are certain GA operations which are tailored for use on one-dimensional arrays such as the bitmasks. These operations include `GA_Patch_enum`, `GA_Scan_add`, `GA_Scan_copy`, `GA_Pack`, and `GA_Unpack`. The remaining GA operations are N-dimensional and one-sided and include `NGA_Scatter`, `NGA_Gather`, `GA_Put`... REMOVE

The vast majority of functionality within the subsetter is provided by either PnetCDF or GA. GA allocates and evenly distributes the arrays which are then filled with data by PnetCDF. GA operations are then used to prepare the data for packing at which point a custom n-dimensional packing routine is used. The packed, evenly-distributed data is then written back to disk using PnetCDF. Of these algorithms, the novel ones include reindexing the masks, reindexing the connectivity variables, and the n-dimensional pack routine.

E. The Algorithms

The one-sided communications and PGAS model afforded by GA allowed us to develop some novel algorithms for the manipulation of unstructured grids. In this section we diagram and describe the algorithms we developed.

1) Partial Sum:

2) REINDEXING OF DISTRIBUTED MASKS: Creating the index array associated with a mask is the easiest of the algorithms. It only requires three specific GA operations, `GA_Fill`, `GA_Patch_enum` and `GA_Unpack`. `GA_Fill` fills the index array with a value of -1. `GA_Patch_enum` enumerates the values of a second array starting from zero with an increment of 1. `GA_Unpack` expands the enumerated array values into the filled array based on the associated mask array.

3) REINDEXING OF CONNECTIVITY VARIABLES: Recall that the connectivity variables are those which map from one index to one or more other indices such as from a cell index to each of its corner indices. A typical subset operation reduces the number of cells, corners, and edges within the data, so it is important to maintain the integrity of these mapping arrays such that they map to real indices.

The reindexing of the connectivity variables relies on the recalculated index array of the associated domain. For example, when reindexing the mapping from cells to corners, the recalculated corners index array is required. The mapping values represent indices into the recalculated index array, so the values are organized into set of indices for the GA routine `NGA_Gather` to query. The `NGA_Gather` routine gathers array elements from a global array into a local array and in this case gathers the new values for the mapping. The gathered values then appropriately replace the old mapping values.

4) N-DIMENSIONAL PACK ROUTINE: TODO – Consider pseudo code? `GA_Scan_add` routine was used to perform partial sums on all masks. That helps determine where to `NGA_Put` the subset data.

IV. Implementation

V. Evaluation

VI. Future Work

We are currently developing a general C++ API for climate data analysis in a data parallel fashion based on the PGAS model and one-sided communications. The API will be leveraged to produce additional command-line tools, however it is intended primarily to be used by climate scientists to produce the kinds of tailored analyses which they require. Time permitting, the API will be exposed to the Python language in order to facilitate ease of use in a scripted environment. Larson, Ong, and Tokarz also point out certain capabilities missing from popular climate data analysis packages such as probability density function (PDF) estimation as well as the sorting and ordering of data.

The evaluation performed in V revealed that IO for our application remains the greatest bottleneck. This fact is exacerbated when small regions representing a mere fraction of the entire grid are subset. Our current strategy is terribly inefficient because it reads in entire variables only to immediately cull the majority of them. If a masked read were available within the Parallel NetCDF library we would likely see performance gains. The algorithms presented in this paper which evenly distribute the subset data will hopefully encourage work in this area.

VII. Conclusion

We developed a novel data parallel subsetter of climate data based on unstructured grids and the PGAS model. The experimental evaluation showed scalability to hundreds of processors and acceptable IO bandwidth.

Acknowledgment

TODO

References

[1] J. W. Larson, E. T. Ong, and C. Tokarz, "The spheroidal analysis library and toolkit: Tools for climate model output analysis," in *MODSIM 2007 International Congress on Modelling and Simulation*. Modelling and Simulation Society of Australia and New Zealand Inc., December 2007, pp. 2974–2980.

[2] D. A. Randall, "Design and testing of a global could-resolving model," February 2009, midterm Progress Report to the U.S. Department of Energy's Program for Scientific Discovery through Advanced Computing (SciDAC).

[3] TODO, "Climate data analysis tools," TODO TODO, tODO.

[4] ——, "Climate data operators," TODO TODO, tODO.

[5] C. Zender, "The netcdf operators (nco)," http://nco.sourceforge.net/.

[6] A. Adcroft, J.-M. Campin, C. Hill, and J. Marshall, "Implementation of an atmosphere-ocean general circulation model on the expanded spherical cube," *Monthly Weather Review*, vol. 132, no. 12, pp. 2845–2863, December 2004.

[7] J. Nieplocha, B. Palmer, V. Tipparaju, M. Krishnan, H. Trease, and E. Apra, "Advances, Applications and Performance of the Global Arrays Shared Memory Programming Toolkit," *International Journal of High Performance Computing Applications*, vol. 20, no. 2, pp. 203–231, 2006. [Online]. Available: http://hpc.sagepub.com/cgi/content/abstract/20/2/203

[8] TODO, "Ferret," TODO.

[9] ——, "Ferret," TODO.

[10] D. A. Randall and R. P. Heikes, "Spherical geodesic grids: A new approach to modeling the climate," http://kiwi.atmos.colostate.edu/BUGS/geodesic/.

[11] V. Balaji, "A standard description of grids used in earth system models," http://www.unidata.ucar.edu/events/2006GridWorkshop/balaji.pdf, October 2006.

[12] F. Aikman, C. Beegle-Karuse, S. Hankin, T. Gross, and J. Caron, "Community standards for unstructured grids," https://cf-pcmdi.llnl.gov/trac/wiki/UnstructuredGrids, Boulder, CO, USA, October 2006.

[13] B. Howe and D. Maier, "Algebraic manipulation of scientific datasets," in *In VLDB*, 2004, pp. 924–935.

[14] J. Li, W.-k. Liao, A. Choudhary, R. Ross, R. Thakur, W. Gropp, R. Latham, A. Siegel, B. Gallagher, and M. Zingale, "Parallel netcdf: A high-performance scientific i/o interface," in *SC '03: Proceedings of the 2003 ACM/IEEE conference on Supercomputing*. Washington, DC, USA: IEEE Computer Society, 2003, p. 39.

[15] TODO, "Todo," TODO.

[16] R. K. Rew and G. P. Davic, "Netcdf: An interface for scientific data access," *IEEE Computer Graphics and Applications*, vol. 10, no. 4, pp. 76–82, July 1990.

[17] R. Thakur, W. Gropp, and E. Lusk, "Data sieving and collective i/o in romio," in *Proceedings of the 7th Symposium on Frontiers of Massively Parallel Computation*. Washington, DC, USA: IEEE Computer Society, February 1999, pp. 82–189.

[18] S. Nativi, J. Caron, E. Davis, and B. Domenico, "Design and implementation of netcdf markup language (ncml) and its gml-based extension (ncml-gml)," *Computers and Geosciences*, vol. 31, no. 9, pp. 1104–1118, November 2005.

[19] TODO, "Todo," TODO.