IO strategies and data services for petascale data sets from a global cloud resolving model

KL Schuchardt, BJ Palmer, JA Daily, TO Elsethagen and AS Koontz

Pacific Northwest National Laboratory, PO Box 999 Richland WA 99352

karen.schuchardt@pnl.gov

Abstract. Global cloud resolving models at resolutions of 4km or less create significant challenges for simulation output, data storage, data management, and post-simulation analysis and visualization. To support efficient model output as well as data analysis, new methods for IO and data organization must be evaluated. The model we are supporting, the Global Cloud Resolving Model being developed at Colorado State University, uses a geodesic grid. The non-monotonic nature of the grid's coordinate variables requires enhancements to existing data processing tools and community standards for describing and manipulating grids. The resolution, size and extent of the data suggest the need for parallel analysis tools and allow for the possibility of new techniques in data mining, filtering and comparison to observations. We describe the challenges posed by various aspects of data generation, management, and analysis, our work exploring IO strategies for the model, and a preliminary architecture, web portal, and tool enhancements which, when complete, will enable broad community access to the data sets in familiar ways to the community.

1. Introduction

Climate simulations at resolutions of 4 km or less will create enormous quantities of data severely straining existing models for IO, data storage, data management, and post-simulation analysis. A Global Cloud Resolving Model (GCRM) simulation, at this resolution, which exports results to disk every simulated hour will generate on the order of 1-10 petabytes of data for a simulated year. This is more data than will fit on existing disk systems as well as disk systems considered in the near future. It is also more data than can be easily transported, either electronically or by physically moving the storage media, thus breaking traditional data sharing methods. For collaborators to access the data, it must first be reduced in size or the analysis must be co-located with the data.

The GCRM code we are supporting, currently under development at Colorado State University, operates on a geodesic grid, creating challenges for analysis. Geodesic grids do not conform to the regular latitude-longitude grids around which an extensive array of analysis and visualization tools have been developed. The non-monotonic nature of the grid requires extensions to existing tools and leads to increased processing time for some operations, such as the creation of hyperslabs. Furthermore, the size of individual files produced by these simulations is potentially very large. For a 4 km resolution simulation that exports all data for just one time step to a single file, that file is estimated to be on the order of 1-10 terabytes, with each individual variable approximately 10 gigabytes. This substantially exceeds the 4GByte variable size currently handled with NetCDF four-byte offsets. Strategies for decomposing the data across files are necessary. Under most configurations under consideration, it will be necessary to work with both large numbers of files and files of

substantial size. The grid itself is of substantial size (4 gigabytes) and should not be duplicated in each file, introducing another challenge for data analysis.

Because the total volume of data is so large, most of it is expected to reside on off-line storage. Accessing the data is then a two stage process, each of which can incur significant latency: first the desired data must be located and moved from archive to disk, and second, reduction and/or analysis must be performed on the data before making it available to the user. If the user is requesting an analysis on the complete set of data or a large fraction of it, then the analysis must be performed "in situ". The challenge is to create an environment that makes this data available to a large community, either by allowing extraction of suitable small, transportable subsets and delivering the subsets to researchers' computers, or by providing tools that will allow researchers to run analyses of the complete data set remotely. Strategies for structuring the data to minimize movement from archive to disk cache based on usage patterns must also be examined.

Finally, the large data volumes involved will require parallel IO strategies, both in generation of the data from the GCRM simulation and for subsequent data reduction, reformatting and analysis. The original calculation is expected to run on tens of thousands of processors. The traditional method of funneling results to the head node and writing them out from there obviously breaks down in this case, but even methods in which every node dumps its data to individual files will likely overwhelm the file system. New strategies that use only some fraction of the total nodes for IO need to be considered. The programs that perform post simulation analyses themselves will need, in many instances, to be parallel. The fact that a single timestep for one variable represents on the order of 10's of gigabytes of data implies that it will need to be distributed just on the basis of memory alone. Furthermore, producing analyses and reduced data sets in a timely manner also suggests that parallel strategies need to be considered.

2. IO Benchmarking

The large volumes of data produced by the GCRM will severely strain resources, including in-core memory, available disk, and IO bandwidth. The desired data formats may also complicate IO. For example, the easiest way to move data from processors to disk is to unload the spatial region allocated to a processor or a group of processors to a single file. This implies that individual files will contain many variables but represent a localized geographical region. However, such a data organization severely complicates data analysis, which is typically performed on a per-variable basis. This means that either the data must be post-processed into the desired structure or data must first be moved internally between processors before being written to disk using an alternative strategy.

To date, this project has examined three IO strategies. The first is to have each processor individually write its local data to a file. The number of files written to disk per snapshot is equal to the number of processors. This approach, while not considered a viable strategy for large processor counts, was investigated to establish baseline performance characteristics. It is straightforward to implement and does not exceed maximum variable size limits but severely strains available IO bandwidth if the number of processors is very large (the expected case for GCRM production runs) and requires extensive post-processing. Contention for IO resources that will occur when all processors try to write data simultaneously will likely also lead to serious performance degradation.

To minimize the effects of IO contention, a second strategy is to have only a subset of processors perform write operations. The data from several processors is collected by an IO processor and written to disk. This reduces the number of processors doing IO but comes at the cost of increased program complexity and higher interprocessor communication. An important consideration for this strategy is that data must be moved to the IO processor and written to disk in relatively small chunks so that the IO buffer does not create unreasonable demands for additional memory.

Both of the IO strategies outlined above must interrupt the main computation while IO is being performed. If the time required to download a system snapshot is a significant fraction of the computation time between snapshots, then IO may represent an unacceptable overhead for the computation. A third approach is to create a dedicated group of processors that perform IO while the

calculation is proceeding concurrently on the remaining processors. This can be implemented using processor groups. This approach also represents significantly more program complexity and will essentially double the memory requirements of the calculation since two copies of the data will be required. One copy represents the current calculation and a second, static copy represents the snapshot that is being written to disk. This second copy is necessary because the data used by the compute nodes will be continuously updated during IO.

These strategies manage disk IO from the processor side but some tools under development can also help manage IO by exploiting properties of the file system. These include parallel file libraries such as Parallel NetCDF[1] and NetCDF4. Parallel toolkits such as these may allow creation of single variable per file formats without having to move as much data between processors as would be required for non-parallel tools such as the standard NetCDF library.

Some preliminary results for the first two strategies at resolutions up to 8 km are shown in the figures below: effective bandwidth decreases for large numbers of processors; very large numbers of processors appear to overwhelm the file system; and internal communication is scalable.

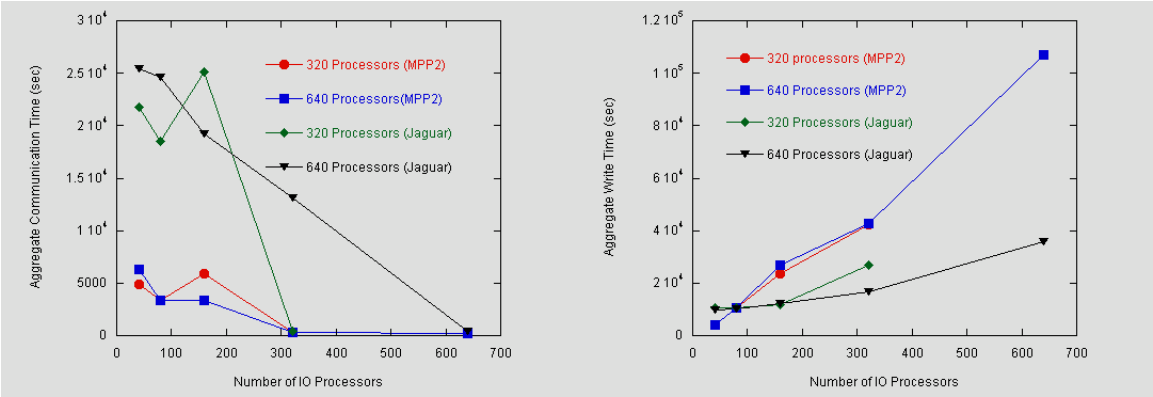

Figure 1. Preliminary results: a) total communication time b) total write time

3. Post Simulation Data Services

To provide community access to the GCRM data, we are developing a web portal that hosts data reduction, data analysis, and visualization services. The web portal will provide ubiquitous community access to the data services with support for restricted access when appropriate. Users will locate datasets via traditional methods such as searching and data set navigation. In addition, a semantic model for the metadata will enable users to discover data related to their original searches or to browse based on selected criteria. Because of the size and detail of the data sets, a mechanism to visually browse pre-constructed views of the data will be developed, enabling users to locate "interesting" data. Subsetting and analysis interfaces will strive to provide users the same capabilities traditionally performed by existing tools after the data has been transferred to the client systems. Registered users will have a separate web space detailing their completed and pending data requests. Once a request completes, users have the option of further manipulating the data *in situ*, or if the data has been sufficiently reduced, transferring the data to their client systems where they can apply one of the many popular climate analysis tools.

The earth systems community has expended significant effort to create web interfaces, tools, protocols, and data guidelines with great success [2-4]. We will leverage this work where possible and extend capabilities where needed. Figure 2 shows a high level view of our preliminary architecture and some of the core technologies we plan to adopt. In the following sections, we focus on areas in which we will be creating extensions or adding new capabilities.

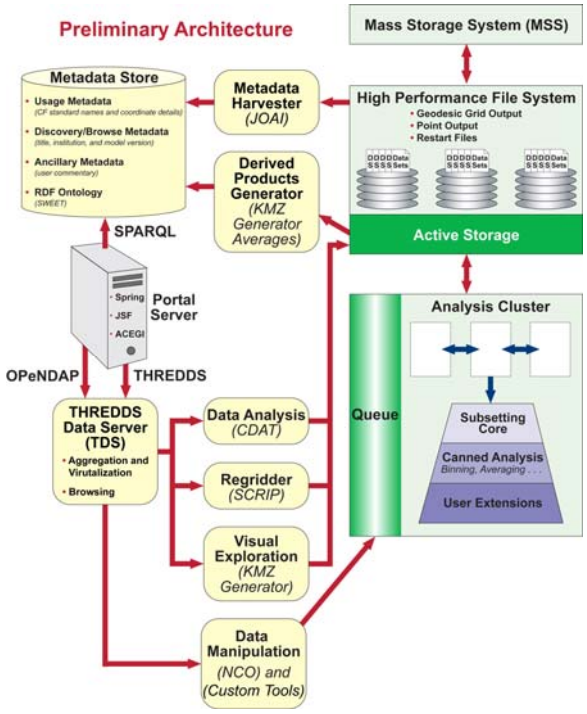

Figure 2. Preliminary data services architecture. We will build upon tools such as OPeNDAP, THREDDS, CDAT, and NCO to meet the unique processing requirements of the GCRM model output.

3.1 Metadata Services

Metadata services are crucial to support browsing, discovery, and delivery of datasets. Metadata resulting from datasets produced by a GCRM simulation, as well as derived data sets, will be harvested and stored in a metadata repository. For the metadata repository we are currently investigating the use of a Resource Description Framework (RDF) based triple store as our storage and query platform of choice. An RDF store can support the layering of ontologies such as the Semantic Web for Earth and Environmental Terminology (SWEET) (*http://sweet.jpl.nasa.gov/ontology/*) ontology being developed by NASA, to enhance searching and categorization of the metadata.

A principal component of our metadata services architecture will be a THREDDS (Thematic Real-time Environmental Distributed Data Services) Data Server (TDS) (*http://www.unidata.ucar.edu/projects/THREDDS/tech/TDS.html*). The TDS will serve as middleware connecting users to their datasets. A fundamental concept of the TDS is THREDDS catalogs, which are XML documents describing the available datasets. Metadata will be extracted from simulation output files and stored in TDS catalogs, where it will later be harvested for storage in our metadata repository. TDS also supports file aggregation. A GCRM simulation will be composed of thousands of files and the ability to aggregate these files into a single logical dataset is an important requirement. We have found that current TDS aggregation methods already work for some of the data organization strategies currently under consideration.

An objective of the metadata services work is to make use of existing metadata standards and conventions where possible to promote the processing, comparison, and sharing of data. GCRM outputs will adhere to the NetCDF Climate and Forecast (CF) Metadata Conventions (*http://cf-pcmdi.llnl.gov*). A conformance document, based on outputs from a precursor model has been developed. We are also investigating the use of an emerging standard for the description of grids used in Earth system models titled gridspec (*http://www.gfdl.noaa.gov/~vb/gridstd/gridstd.html*). It is expected that gridspec will be used to describe the grid layout (neighbour information) and provide a standardized mechanism for describing the grid external to each output file, thus avoiding duplication.

3.2 Data Request Services

As shown in the figure, data services will be provided via the OPeNDAP protocol. Though there are many OPeNDAP servers in use, substantial changes will be required to accommodate our large data sets. For example, because individual variables are large (10GB per step), we expect to build parallel tools for standard mathematical operations such as averaging, perhaps expanding on the capabilities of the NetCDF Operators[5]. We have already enhanced ncks to support subsetting of geodesic grids based on lattitude/longitude values. Because geodesic coordinate variables are not monotonic, performance compared to standard grids can be substantially slower depending on the request. For example, while a request for a 10 degreee slice of data along the equator in traditional lat/lon grids can be optimized to minimal IO operations, the same request against the geodesic grid at the target resolution will result in 13,926 non-contiguous blocks of data. Additional new tools unique to the geodesic grid such as geodesic upscaling will also be needed. Such operations require knowledge about cell neighbours which may be available via gridspec extensions or alternatively computed as needed. We plan to support interpolation to latitude/longitude grids using the SCRIP software (http://climate.lanl.gov/Software/SCRIP/).

Because data sets may be off-line, subsetting and analysis tasks may require substantial time to complete. Our intent is to provide a parallel framework for custom analysis. We also plan to work with the OPeNDAP community to support asynchronous requests, which will return immediately but data may not be available until a later time.

3.3 Portal Capabilities

The unique scale and resolution of the GCRM data suggests some enhancements to current earth systems portals. For example, we expect to support polar projections for region selection in addition to the standard cylindrical earth projection. The resolution of the model makes it possible to validate against observed data such as satellite imagery which requires, at a minimum, unique slicing patterns through the data. It should also be possible to follow major weather patterns, such as the movement of a hurricane, suggesting selection of subsets moving through time and space. Finally, we are looking into derived products, which could be held outside of archival storage, that can provide an overview of the data set and enable users to visually locate interesting data. One approach under consideration is to provide GoogleEarth imagery that zooms to the full resolution (~ 3km).

4. Summary

The GCRM model creates many new challenges to both generating the data and providing community access to it for analysis. Significant effort is required to develop an IO strategy that will enable both the code and analysis to perform efficiently. Extensions are required to many of the existing analysis tools to support the geodesic grid. Special portal capabilities will be required to maximize the usefulness and accessibility of the data. We are extending the extensive base of existing tools and technologies to provide an efficient IO strategy for the GCRM and make model outputs available to the community for full analysis.

5. Acknowledgements

This work was conducted at the Pacific Northwest National Laboratory, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy under contract DE-AC05-76RLO 1830.

References

[1] Jianwei Li , Wei-keng Liao, A. Choudhary, R. Ross, R. Thakur, W. Gropp, R. Latham, A. Siegel, B. Gallagher, M. Zingale. Parallel netCDF: A High -Performance Scientific I/O Interface. *Supercomputing 2003*, pages 39-49.

[2] Live Access Server. http://ferret.wrc.noaa.gov/Ferret/LAS/index.html. (June 6, 2007)

[3] P. Cornillon, J. Gallagher, T. Sgouros. OPeNDAP: Accessing data in a distributed, heterogeneous environment. *Data Science Journal 2003*, Vol. 2, pages 164-174.

[4] D. Bernholdt, S. Bharathi, D. Brown, K. Chanchio, M. Chen, A. Chervenak, L. Cinquini, B. Drach, I. Foster, P. Fox, J. Garcia, C. Kesselman, R. Markel, D. Middleton, V. Nefedova, L. Pouchard, A. Shoshani, A. Sim, G. Strand, D. Williams. The Earth system grid: supporting the next generation of climate modeling research. *Proceedings of the IEEE, 2005*. Vol. 93, Issue: 3, pages 485-495. .

[5] C.S. Zender, D.L. Wang. High performance distributed data reduction and analysis with the netCDF Operators (NCO). *AMS Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, January 2007.