J9.15 Breaking the Billion Zone Barrier –
Simulation, Data Handling and Visualization: An Example

Robert Wilhelmson**, Paul Woodward*, Sarah Anderson*, David Porter*,
Steven Peckham, and Crystal Shaw
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

*Laboratory for Computational Science and Engineering
University of Minnesota

1. INTRODUCTION

The geoscience community in the first decade of the 21st century has been challenged to advance the understanding of geophysical phenomena and to translate this understanding into meaningful actions in service to the public. The document NSF Geosciences Beyond 2000[1] states that improved understanding that results will "enable reliable predictions of significant changes to Earth's current state. Earthquakes, severe storms, solar storms, and biological invasions represent threats, but we have the opportunity to mitigate these threats for society. Predictions of extreme planetary events can help save lives and/or lessen property damage."

Technological advances continue to provide new opportunities to advance our understanding through enhanced observational studies/monitoring and through the modeling of complex geophysical flows/physics. These technological innovations include the advent of cluster computing, the availability of the national computational Grid, and the possibility of Grid computing. Here cluster computing refers to the use of commodity processors organized in small, shared memory configurations that are coupled using high-speed local networking. "A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities.[2]" Grid computing refers to the use of computational resources on the Grid to carry out simulations that utilize multiple hardware resources at different physical sites with the aid of high speed networks.

These technology advances coupled with increased computer processor speed now provide the backdrop for carrying out simulations of atmospheric fluid flow where the integration domain consists of billions of gird zones. This has already been done for turbulent flow by the co-authors at LSCE (Laboratory for Computational Science and Engineering)[3]. The use of a billion zones per field is a full order of magnitude beyond the largest mesoscale meteorology computations of which we are aware and poses new challenges in the handling and display of huge volumes of model results. Currently an atmospheric model, COMMAS, is being adapted to carry out such computations on one or more parallel cluster computers composed of commodity processors and high speed interconnects. These computers will be a part of the TeraGrid[4]

The TeraGrid[5] is a new NSF funded effort to advance the cyberinfrastructure for 21st century science and engineering. It is a response to the pressing need for greater computational power to enable experimentation, modeling, data analysis, and visualization activities that often involve large volumes of data and use of an expanding national computational grid. It is a vision to develop and deploy a comprehensive computational, data management, and networking infrastructure of unprecedented scale and capability that couples distributed scientific instruments, terascale and petascale computing facilities, multiple petabyte data archives, and gigabit and beyond networks, all widely accessible by researchers including those in the atmospheric sciences, oceanography, and hydrology.

2. ENABLING WORK AT LCSE

Work at LCSE has centered on the development and use of PPM[6] codes to study turbulent, compressible fluid flows including 3D simulations of luminous red giant stars, stellar convection in the sun, convectively unstable layers, and homogeneous compressible turbulence.[7,8,9] The current strategy for many of these

**Corresponding author address: Robert B. Wilhelmson, National Center for Supercomputing Applications, U. of Illinois, 605 E. Springfield Ave., Champaign, IL 61820

[1] NSF Geosciences and Beyond 2000: Understanding and Predicting Earth's Environment and Habitability: Summary, National Science Foundation, NSF 00-28.

[2] Foster, I., and C. Kesselman, Ed., 1999: *The Grid Blueprint for a New Computing Infrastructure*, Chapter 2, Computational Grids, p. 18 (Ian Foster, Carl Kesselman), Morgan Kaufmann Pub., 677 pp.

[3] http://www.lcse.umn.edu

[4] See preprint J3.5 in this volume entitled A Report on Plans for a TeraGrid

[5] http://archive.ncsa.uiuc.edu/MEDIA/vidlib/new.html (streaming video introduction)

[6] Piecewise Parabolic Method

[7] e.g, P. R. Woodward, D. H. Porter, I. Sytine, S. E. Anderson, A. A. Mirin, B. C. Curtis, R. H. Cohen, W. P. Dannevik, A. M. Dimits, D. E. Eliason, K.-H.

very large calculations can be characterized as cluster programming with shared memory on disk. Coupled with the multiple level caches on current machines, this hierarchical model can be utilized to carry out simulations on distributed memory architectures (e.g., the Itanium cluster at NCSA) where high-performance processors are coupled with matching high-performance network interconnects[10]. Cluster implementations of hydrodynamics codes typically have left domain decomposed independent data contexts in each cluster member's memory, sending only an updated "halo" of domain boundary information to neighboring nodes. On these new network and processor balanced clusters, there is enough separation between data production and reuse to completely overlap the read or write of two complete contexts to remote node disk with the update of a third context. Any node may update any context, resulting in a simple coarse-grained load balancing capability. In addition to essentially continuous check pointing on local node disk, it is possible to stop and restart a computation on any subset of the cluster, or run "out of core", solving problems larger than would fit in even the whole cluster's memory. The implementation of this programming model is based on a small Fortran or C callable library implementing asynchronous remote I/O, transparently reading and writing named data objects residing anywhere on the cluster. In addition, library routines are available to simplify master/worker task queue management. To the application programmer, the library offers the basic interlock-free read/write capabilities of a shared file system without giving up the performance provided by a high-speed network interconnect.

The focus at LCSA is not only on programming for large and possibly distributed clusters but also on the processing of the massive amounts of data/information produced. Analysis and visualization software have been written for this purpose including a powerful volume renderer that for modest amounts of data can be run on newer PC's with high performance graphics cards and that for large amounts of data can be run in a distributed mode. This software allows users located at PCs anywhere on the Grid to interactively specify keyframes for volume rendered movie visualizations of terascale data sets, to preview at low image quality these movies as rendered on their local PC, and, if the preview is satisfactory, to generate high image quality movies on a remote system and have the resulting images sent to them over the Grid. This software, the first of its kind to our knowledge, runs the remote movie generation on the powerful visualization cluster at the LCSE.

To use this software, a researcher will move the data to be visualized from the location on the Grid where it is archived (or streaming it from where it is being generated) to the LCSE's 4 terabyte, high-performance disk cache. Recent experiments in collaboration with NCSA and the University of Minnesota networking specialists have produced sustained transfers at bandwidths ranging from 6 to 20 MB/s between NCSA staging disks and the LCSE disk cache. Since this Internet-2 connection operates at OC-12 bandwidth, up to 40 MB/s sustained data transfers should be achievable.

Once the simulation data arrives at the LCSE, the software will process it into a hierarchical format for efficient volume rendering at user-specified quality and resolution. The program that the researcher will run from a PC on the Grid, which we will refer to as the LCSEmovieGUI, will pull over the Grid the coarsest representation of the data from the hierarchical format stored on the LCSE disk cache. Using this coarse representation of the data, and using the graphics hardware of the user's PC, the researcher can generate previews of images that can be selected as key frames for the finished movie. The user can interactively modify color maps and opacity maps as well as the viewing position, viewing angle, and location of front and/or back clipping planes. Such image parameters can be set independently for any elements of a sequence of movie data files, or for successive views of the same data file. Once the parameters of the image are set, the user may request that a high quality image be rendered on the LCSE equipment and automatically sent to the viewing window of the LCSEmovieGUI for inspection.

If all the key frames so specified by the user are acceptable, any subset of them or all of them can be designated in any sequence as key frames for a movie. The number of frames to be generated from parameters interpolated between those of any two successive key frames in this sequence can then be specified by the user. Common operations, such as

---

Winkler, and S. W. Hodson, "Very High Resolution Simulations of Compressible, Turbulent Flows," to appear in the Proc. of the Fourth UNAM Supercomputing Conference on Computational Fluid Dynamics, Mexico City, June, 2000, edited by G. Cisneros, World Scientific; available at www.lcse.umn.edu/mexico.

[8] I. V. Sytine, D. H. Porter, P. R. Woodward, S. H. Hodson, and Karl-Heinz Winkler 2000, "Convergence Tests for Piecewise Parabolic Method and Navier-Stokes Solutions for Homogeneous Compressible Turbulence," J. Computational Physics, 158, 225-238. http://www.lcse.umn.edu/parcfd/

[9] I. V. Sytine, D. H. Porter, P. R. Woodward, S. H. Hodson, and Karl-Heinz Winkler 2000, "Convergence Tests for Piecewise Parabolic Method and Navier-Stokes Solutions for Homogeneous Compressible Turbulence," J. Computational Physics, 158, 225-238. http://www.lcse.umn.edu/parcfd/parcfd99.pdf

[10] Sarah E. Anderson, B. Kevin Edgar, David H. Porter and P. R. Woodward, Cluster Programming with Shared Memory on Disk. http://www.lcse.umn.edu/publications

view point rotation about a specified axis, can also be specified as methods for interpolating frames between members of the key frame sequence. Once the entire movie has been specified in this manner, the user can request that a preview be generated interactively on the local PC. Such a preview will have low image quality, but today's PCs allow for it to be rendered at interactive speed from the coarse data that was brought to the local PC from the hierarchical data decomposition on the LCSE disk cache. If a better preview is desired, the researcher can designate image quality and resolution parameters that allow the movie to be rendered on the LCSE equipment very rapidly. This preview movie will be sent to the user's PC automatically, if the user so requests, and this movie can provide a much better impression of the desired finished product. If this preview is acceptable, the researcher can revise the image quality and/or resolution settings for the movie for a final rendering at the LCSE. The movie images can be generated for any desired display, either mono or stereo, either single or multiple panel, and, if multi-panel, for any geometrical arrangement of image panels and preferred viewer eye position. Experience to date indicates that movie images can be sent over the Grid to a researcher faster than they can be rendered.

This Grid-based movie generation capability is a natural application for the emerging very fast national research networks. The computing systems that can perform the now very arduous and time-consuming rendering tasks are very different in character from those designed to generate the data from which the visualization is made. Volume rendering, like other forms of computer graphics, is highly amenable to encapsulation in specialized hardware. Such hardware, driven by the movie and PC gaming industries, tends to deliver a roughly 100-to-1 price/performance advantage over general purpose scientific computing hardware. Not only is the rendering equipment specialized for the movie generation application, but also a specialized disk subsystem, such as the LCSE 4 TB disk cache, is required. The movie generation process requires multiple terabytes of raw data to be on-line simultaneously. In order to use this data efficiently, a shared file system, such as the one LCSE, is needed together with a large Fibre Channel storage area network.

3. ATMOSPHERIC APPLICATIONS

The storm group at the University of Illinois led by Dr. Wilhelmson is collaborating with the LCSE co-authors to create a new version of COMMAS (Collaborative Model for Multiscale Atmospheric Simulation) to study tornadogenesis and evolution in supercells and in hurricanes. The new version will use cache more effectively through changes in numerical algorithms and through code optimization. For example, the small time step used in most convective scale models today has been replaced by a forward-backward scheme that lends itself to better cache reusage. The recoding goal is to create a faster and more accurate model code that can be blended with the shared memory on disk model outlined above. This approach differs from the one being undertaken by the WRF community in that the effort is not constrained by the desire to create an easily extensible and readable community code for general usage. Further, the shared memory model on disk is not currently implemented in the WRF framework. However, it is expected that the results of the COMMAS recoding effort will impact WRF implementation, particularly for efficient and optimized Linux cluster usage.

An example of one planned simulation with the new PCOMMAS model is the study of tornadogenesis within a supercell where both the supercell and tornado are resolved with the same resolution. The growth of a supercell will be modeled using modest resolution. As it matures, the model data will be interpolated to a grid that is 20 m in the horizontal and whose vertical resolution varies from 5 m near the surface to 250 m near the top of the integration domain. The grid size then becomes 3072×3072×256 grid (2.4 billion zones) in a domain approximately 60 x 60 x 20 km domain. It will take about twelve days on a 1024 processor Itanium cluster to complete a one hour simulation based on current estimates. These numbers may sound very large but other scientific communities are also beginning (if not already doing such calculations) to consider simulations of this size and duration.

4. ACKNOWLEDGEMENTS

The authors wish to thank NCSA and NSF for its support in helping us build the infrastructure necessary to carry out our "extreme" calculations.