NSF-ITR Gleaning Insights in Large Time-Varying Scientific and Engineering Data

Annual Report – 2006-07
NCAR

VAPOR

Fueled by years of exponential improvements in microprocessor technology, computational scientists are able to run numerical simulations at unprecedented scales and generate volumes of data as never before. Yet our ability to analyze complex, time-varying, multi-variate, multi-dimensional data sets has not kept pace with our ability to generate them. In practice, scientists simply do not have access to analysis resources that are on par with those used for simulation. The aim of the VAPOR project is to improve this situation for earth scientists employing very high-resolution numerical fluid flow models. A single turbulence simulation today may require weeks or even months to compute, and generate terabytes of multi-variate, time-evolving data.

VAPOR leverages two key technologies, operating hand-in-hand, to allow a scientific end-user to penetrate vast data sets. Advanced, hardware accelerated visualization enables a researcher to rapidly identify a spatial and/or temporal region of interest (ROI). Quantitative, more computationally expensive tools may then be used to further explore the reduced domain ROI. Visualization and the rapid identification of ROIs alone, however, is not sufficient to mitigate the challenges of terabyte size data sets. Also integral to VAPOR's large data handling capabilities is wavelet-enabled progressive data access, permitting the user to make speed/quality tradeoffs. Simulation outputs undergo a wavelet transform, allowing them to be accessed at progressively finer resolutions, and offering order-of-magnitude reductions in data set size. The combination of visualization and progressive access make possible highly effective, interactive exploration of terascale data sets using only a lowly desktop platform. In this way VAPOR differentiates itself from other packages that employ brute force methods and demand substantial computing resources for large data investigation.

A foremost design goal of VAPOR is to provide a comprehensive environment for data analysis, not simply a tool for generating imagery for publication and presentation purposes. In this way, VAPOR further distinguishes itself from the multitude of general-purpose visualization applications already available. Fundamental to this objective is VAPOR's robust support of mathematical operators for data manipulation and quantitative interrogation.

Development of VAPOR is closely guided by a steering committee comprised of turbulence researchers from around the world. This panel of experts sets development priorities, dictates software requirements, and serves as friendly users for testing and evaluating new software features.

2006-07 Accomplishments

VAPOR was first released to the scientific community in March of 2006. A second major release followed a year later in March of 2007. Over 1500 copies of the software have been downloaded from the VAPOR web site (www.vapor.ucar.edu), and the package is rapidly establishing itself as the tool of choice for the high resolution, earth sciences CFD community. Research groups from institutions around the world including: NCAR, CU, UNH, UCSB, SDSC, JHU, Ecole Normale Superiéure, Neils Bohr Institute, and the University of Stockholm, to name a few, are now actively collaborating and using VAPOR in their work. The strongest evidence of VAPOR's success are the number of scientific publications citing VAPOR – six currently with more on the way – and the number of invited VAPOR talks given (see complete list below). An excellent example of VAPOR's use in aiding scientific exploration is the work of Mininni et al [1] in the investigation of a 1536^3 Magnetohydrodynamics (MHD) simulation - the largest MHD simulation published to date - as seen in the image below (Figure 0).

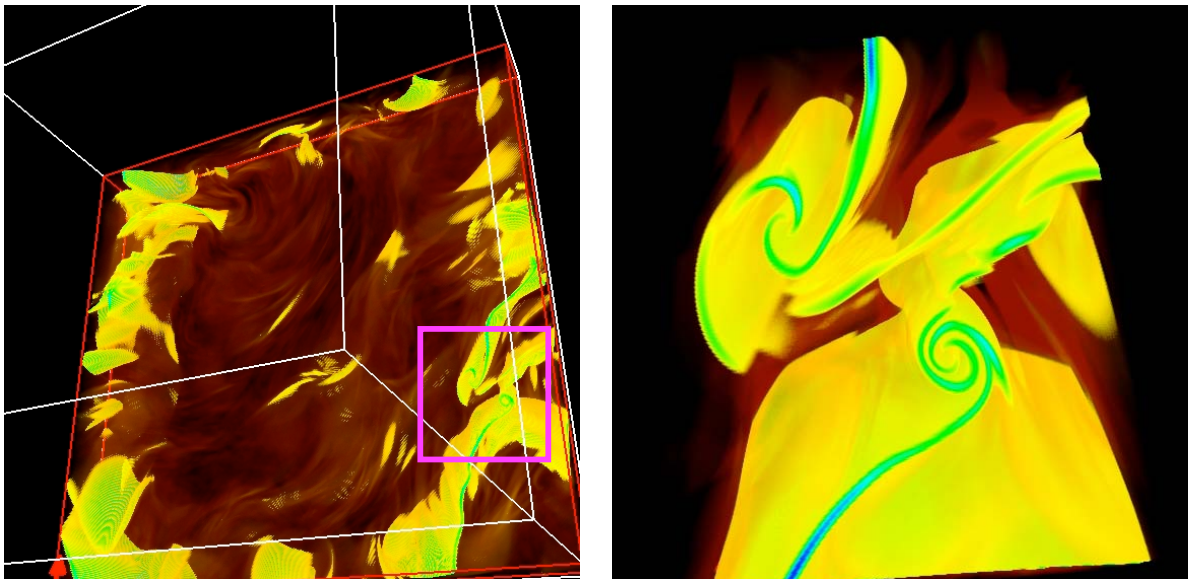

Figure 0 Parallel current sheets exhibiting "roll up" in a 1536^3 MHD simulation. Shown on the left is the full spatial domain at $1/64^{th}$ resolution; on the right a full resolution close up of the area indicated in magenta on the left.

Outreach

A focus area for the VAPOR team in recent months has been promotion of the tool to new users and new scientific domains. Towards that end we have been working with the experimental weather community to better understand their needs and identify gaps in VAPOR's current capabilities. A summer intern has been helping prototype new features to further assess VAPOR's application to these new communities. We've also visited numerous current customer sites to better understand their use of the package and to look for future development areas.

Student Internships

We continue to offer intern opportunities for undergraduate and graduate students. This summer two students joined the VAPOR development effort: Kenny Gruchalla, a PhD candidate from the University of Colorado, is working on advanced GPU volume rendering techniques and prototype spherical grid rendering; and Victor Snyder, an undergraduate from the Colorado School of Mines, is helping explore more generalized data grids to attract users from outside the current turbulence-dominated community.

Partners

We actively seek out collaborations with computer scientists and computational scientists alike. Working with astrophysicists from CU and computer scientists at SDSC and NCAR we have deployed a prototype VAPOR-powered, TeraGrid visualization node. VAPOR's multiresolution capabilities make it an excellent choice for working with remote data sets (as seen in the figure below) where bandwidth is limited, even on the TeraGrid. We have also entered into an agreement to assist Johns Hopkins University with the deployment of a community turbulence database front-ended by VAPOR.

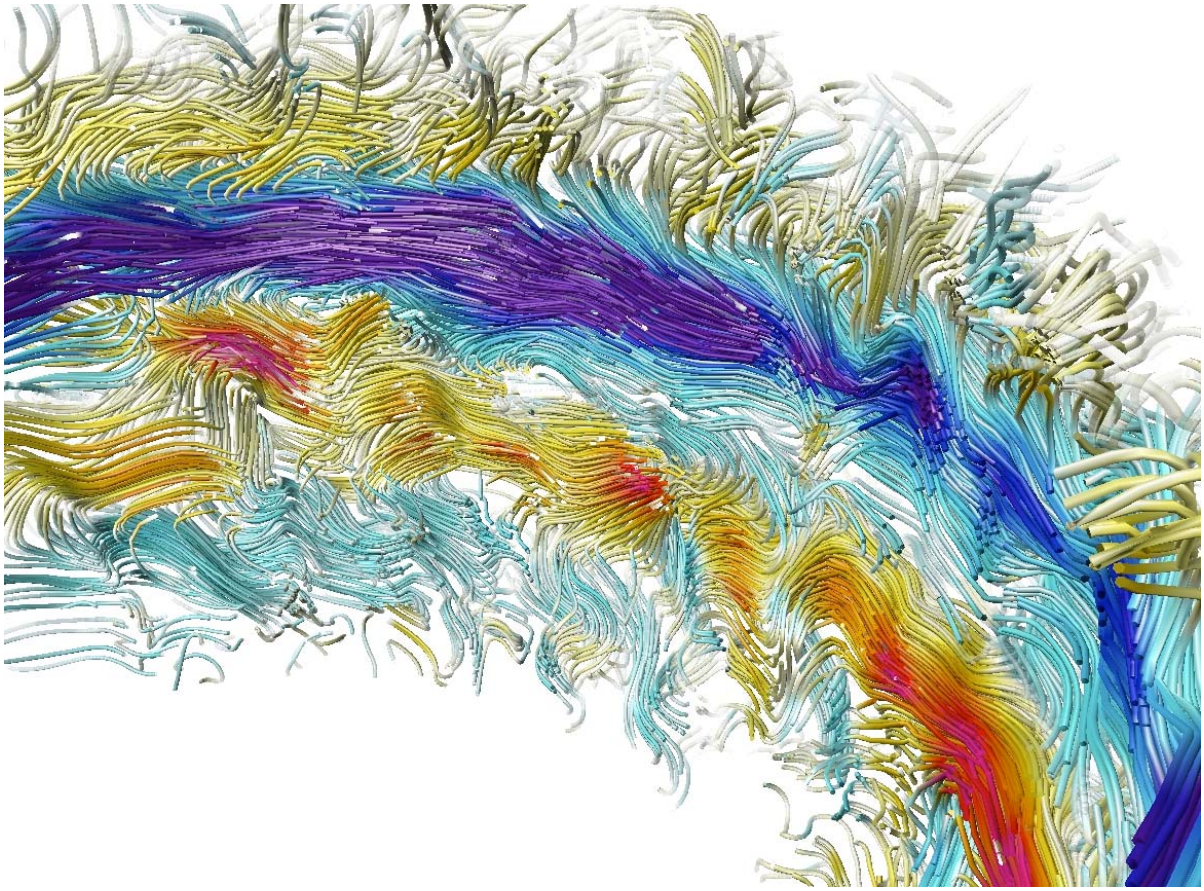

Figure 2: Magnetic field lines near the equatorial region of a stellar convection simulation. The data were computed, and currently stored, at SDSC, while the analysis with VAPOR was performed from CU, Boulder.

Development highlights

Key new features added since the March, 2006 release include:

Mac OS X Support
The Mac is becoming an increasingly popular desktop among the scientific community. Over 50 Mac binaries have been distributed since support for the platform became available in March of this year.

Volume Rendering
With the aid of student interns from the University of Colorado, we continue to improve VAPOR's GPU based volume rendering engine in response to input from VAPOR's steering committee. Rendering quality has been improved through the addition of pre-integrated volume rendering and a lighting model. A prototype spherical grid renderer has also been implemented and will be available with the next release. Spherical grids are of great importance to both the solar physicists and geodynamo scientists we support.

Data Probe
An interactive data probe and arbitrarily oriented planar sampling tool (cutting plane) have been added. The probe can be used to obtain quantitative information such as min, max, mean values over a region; display pseudo colored contours; and may be used to interactively place seeding points for flow advection.

Flow Visualization
VAPOR offers a highly advanced flow visualization system, supporting both steady and unsteady flows. Flow integration may occur over periodic boundaries, forwards and backwards in time, and with non-uniform time sampling. A variety of user-directed methods are available for seed placement including: random with or without field strength bias, placement via a 3D data probe, imported seed coordinate lists, and uniform distribution.

Usability
Much effort has been invested in improving the usability of the GUI, making it more intuitive and better suited to the needs of data analysis. Context sensitive help is now available to provide in-depth technical explanations of features. Step-by-step instructions are integrated into the GUI for performing highly complex operations such has unsteady flow visualization.

Publications

J. Clyne, P. Mininni, A. Norton, and M. Rast. Interactive desktop analysis of high resolution simulations: application to turbulent plume dynamics and current sheet formation, New Journal of Physics (to appear), 2007

M. Rast, J. Mendoza, and J. Clyne, Compressible thermal starting plumes, Journal of Visualization (to appear) 2007

Posters

A. Boggs, B. Brown, J. Clyne, P. Gillman, J. Mendoza, A. Norton, and G. Vasil. Exploring Astrophysical Flow Simulations Using the TeraGrid, TeraGrid '07, Madison, WI, June 4-8, 2007.

Invited talks

J. Clyne and M. Rast. Analysis and Visualization of High Resolution Astrophysical Flows, Astronum 2007, Paris, June 13, 2007

J. Clyne. The peril of the petascale: emerging challenges in large scale computational science, International Workshop on Visualization of High Resolution 3D Turbulent Flows, Paris, June 8, 2007

J. Clyne. The peril of the petascale: challenges to scientific computing. IMAGe Theme-of-the-year workshop, Boulder, CO, May 21-23, 2007

A. Norton. Visualiation and analysis of massive turbulence data sets, International Workshop on Visualization of High Resolution 3D Turbulent Flows, Paris, June 8, 2007

A. Norton. Interactive Analysis and Visualization of Massive Scientific Datasets, Seminar, University of Oregon, April 25, 2007

2007-08 Plans

We will continue to expand and refine VAPOR's core capabilities in the coming year. We anticipate a new minor release near the end of summer that should round out the basic feature set for our current user base as well as provide a number of minor usability enhancements. We will then target a new major release in the 9 to 12 month time frame that will offer support for more generalized data grids. The focus of the new major release will be to broaden the current user community beyond numerically simulated turbulence.

We will also further our petascale preparation efforts, investigating ways to scale VAPOR to support forthcoming petascale application data sets such as NSF's planned, Track 1, $12{,}000^3$ Navier-Stokes turbulence simulation. In particular we will look at *in situ* data transformation and more aggressive compression techniques. An example of our current research work in this area is shown in the image below comparing present hierarchical approximate methods with the new progressive refinement techniques that we have been investigating.

We will continue to look for partnerships with other research groups and, funding permitting, will continue to offer student internship opportunities.

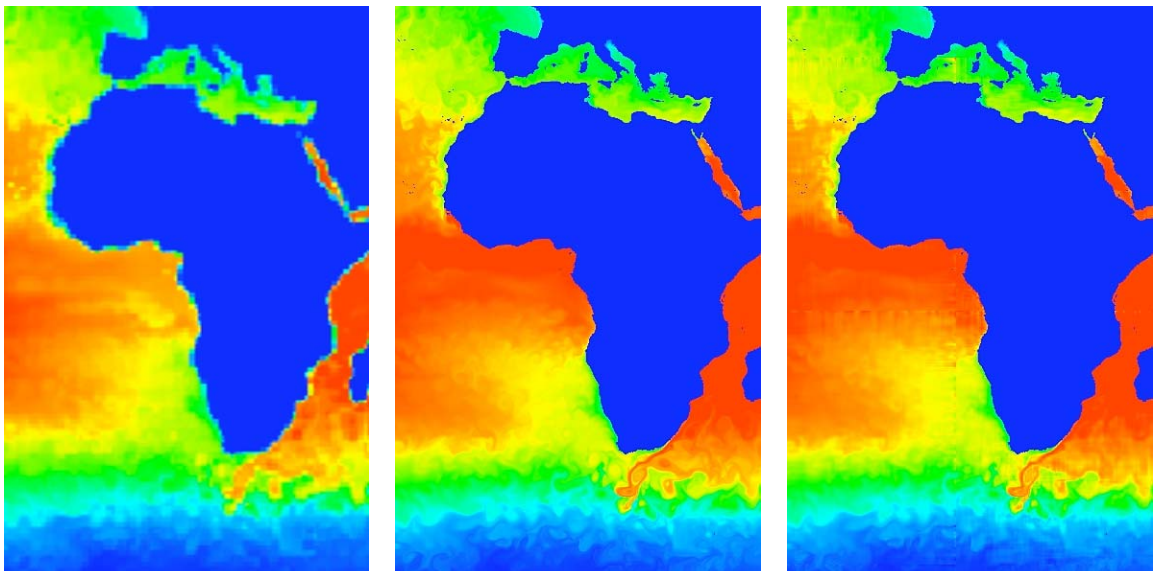

Figure 3 POP 1/10th degree ocean model data compressed 512:1. Shown from left to right are: data compressed using current frequency truncation methods; original, uncompressed data; and data compressed using new wavelet coefficient prioritization methods.

Scientific Data Compression

We have been investigating wavelet-based methods for compressing scientific data sets and are currentl c y working with three different groups: the Southern California Earthquake Center (SCEC), the DOE-funded Earth System Grid (ESG), and ocean modelers at NCAR.

SCEC

Researchers at the SCEC have a need to maintain more seismic simulation data sets on-line than the capacity of their current storage system permits. We have been working with the SCEC to compress their simulation data and compare seismograms derived from the compressed data with original simulation data. Current compressions, which exploit coherence in the temporal domain, appear promising using compression rates up to 20:1.

ESG

The Earth Systems Grid (ESG) provides a database for publishing and distributing outputs from community climate models. ESG users are numbered in the thousands and are located around the globe. Two constraints are presently limiting: secondary storage space and bandwidth to sites with poor connectivity. We have developed climate data compressions tools similar to the ubiquitous gzip command, albeit employing lossy compression and restricting operation to netCDF data sets. We will begin offering compressed climate data to ESG users shortly.

POP Ocean Data

With a grant from IBM climate scientists at NCAR are running the POP ocean model at a global $1/10^{th}$ degree resolution, generating terabytes of data. The data, computed at IBM TJ Watson, must be brought back to NCAR for visualization and analysis. Bandwidth between NCAR and TJ Watson is limited. Using compression tools developed for the ESG we are compressing POP data to facilitate its return to NCAR. A sample of a visualization of the temperature field compressed 512:1 can be seen in Figure 3.

2007-08 Plans

We will continue our promising work with each of the groups discussed above. We will be capturing download metrics from the ESG that will allow us to assess the level of interest from this community and may also provide contacts with groups or individuals to help improve our methods. We will be conducting another compression experiment with researchers at SCEC; this time with an aim to support a 400 TB simulation. And we will continue efforts with NCAR's ocean modelers and look for other opportunities as the arise.

References

[1] Mininni, P., Pouquet, A., and Montogomery, D., 2006: Small-Scale Structures in Three-Dimensional Magnetohydrodynamic Turbulence. *Physics Review Letters*, **97**, 244503