Ultra-Scale Visualization: Research and Education

Kwan-Liu Ma, Robert Ross, Jian Huang, Greg Humphreys, Nelson Max, Kenneth Moreland, John D. Owens, and Han-Wei Shen

SciDAC Institute for Ultra-Scale Visualization

E-mail: ma@cs.ucdavis.edu

Abstract. Understanding the science behind large-scale simulations and high-throughput experiments requires extracting meaning from data sets of hundreds of terabytes or more. Visualization is the most intuitive means for scientists to understand data at this scale, and the most effective way to communicate their findings with others. Even though visualization technology has matured over the past twenty years, it is still limited by the extent and scale of the data that it can be applied to, and also by the functionalities that were mostly designed for single-user, single-variable, and single-space investigation. The Institute for Ultra-Scale Visualization (IUSV), funded by the DOE SciDAC-2 program, has the mission to advance visualization technologies to enable knowledge discovery and dissemination for peta-scale applications. By working with the SciDAC application projects, Centers for Enabling Technology, and other Institutes, IUSV aims to lead the research innovation that can create new visualization capabilities needed for gleaning insights from data at petascale and beyond to solve forefront scientific problems. This paper outlines what we see as some of the biggest research challenges facing the visualization community, and how we can approach education and outreach to put successful research in the hands of scientists.

1. Introduction

Scientists increasingly use parallel supercomputers and state-of-the-art experimental facilities in effort to answer some of the most important and difficult questions in science. The output from these simulations and experiments is so voluminous and complex that advanced visualization technologies are necessary to interpret the data. Even though visualization technology has progressed significantly over the past twenty years, we are barely capable of analyzing terascale data to its full extent, and petascale and exascale datasets are on the horizon. The U.S. Department of Energy's Scientific Discovery through Advanced Computing (SciDAC) program has recognized this urgent need and invested in basic and applied visualization research by funding the Institute for Ultra-Scale Visualization (IUSV). The mission of IUSV is to address this upcoming ultra-scale challenge by leading and coordinating efforts in the visualization community, which will enable data understanding at this scale. IUSV is bringing together a critical mass of experts from visualization, high-performance computing, and science application areas to study new approaches to data visualization, create a suite of comprehensive, scalable visualization tools, and instruct application scientists on how to best use these tools.

Complementing the research program is a diverse outreach and education program, demonstrating new visualization technologies at major conferences, organizing tutorials to educate potential users of these new technologies, benchmarking software and hardware, and making recommendations to industry for revising hardware and software architectures and

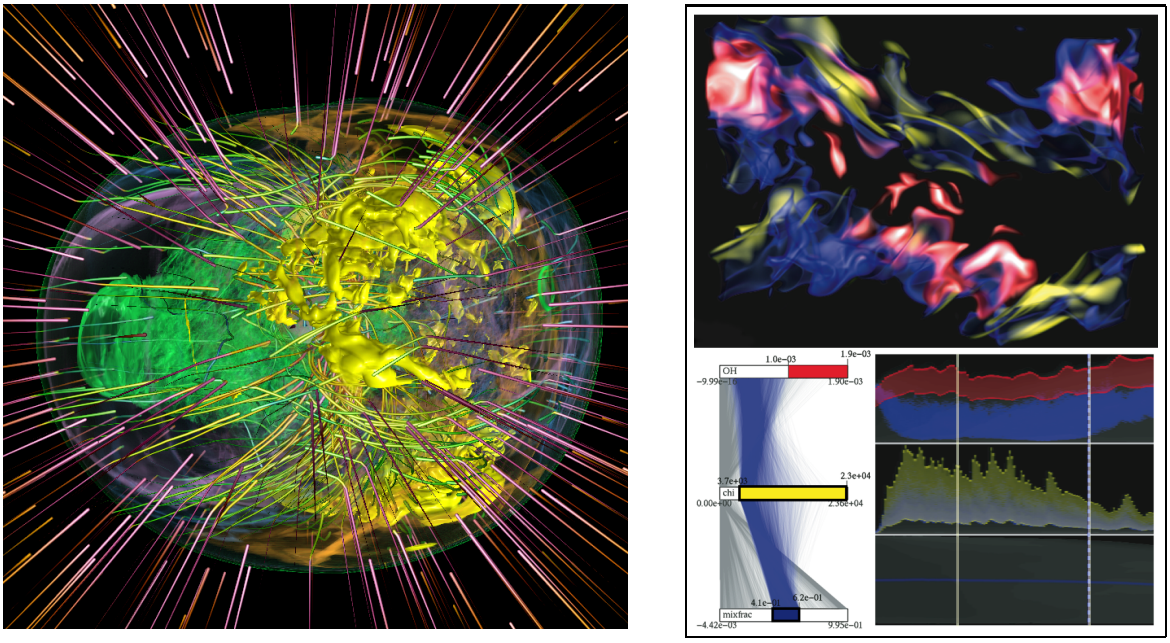

Figure 1. Left: Visualization of pathlines and angular momentum transport from a supernova simulation. The data set was provided by John Blondin. Right: A Time-varying, multivariate data visualization interface.

protocols to support large-scale visualization calculations. The outreach activities, fostering collaborations both within the visualization community and crossing disciplines, aim at expanding existing ties with the Institute by bringing in additional expertise on various topics. These collaborations also serve as an important conduit for distributing the tools, benchmarks, and expertise of the Institute members. For example, the Ultra-Scale Visualization Workshop held at the annual Supercomputing Conferences provides a forum for focused, intensive exchange between researchers and practitioners. This integrated set of activities is strengthened by connections to SciDAC science applications, which serve as early adopters of technology, and by connections to computer science partners, such as VACET, whose tools combine with those developed in the Institute to further aid in the scientific process. The rest of this paper describes more of IUSV's research and education perspectives.

2. Visualization and Analysis Research

While the visualization community has made great strides in providing useful tools to the scientific community, additional work is necessary in order to meet the short-term and long-term needs of SciDAC application teams. Through discussions within the Institute and with application scientists in the community, we have identified three critical focus areas for research:

- Parallel and distributed visualization
- Support for complex datasets
- Knowledge-enabling visualization and analysis

These focus areas match well with the findings of the recent DOE visualization and analytics workshop [2], and they are the basis on which we have formed our research plan.

2.1. Parallel and Distributed Visualization

SciDAC application teams are using some of the largest compute resources in the world to tackle their science challenges. Their codes generate terabytes to petabytes of data during the simulation process, and the size of these datasets continues to grow as more capable computing facilities come on-line. Parallel visualization is the most plausible approach to extracting knowledge from these ultra-scale datasets, and in fact parallel rendering techniques are routinely applied in clusters and on small numbers of nodes on high-end computing systems. Unfortunately, current tools have critical gaps in areas such as parallel I/O and data pre-processing and lack the ability to perform certain types of visualization, such as full-range time-varying visualization, which are extremely helpful to scientists. More research is necessary to determine the right architectures and capabilities for visualization systems that will succeed at petascale and beyond.

Alternatives to traditional CPUs provide a possible avenue for accelerated visualization of scientific datasets. Graphics processors (GPUs) are playing an ever-larger role in the visualization process, but these processors are extremely difficult to use in a coordinated manner. Tools that help map visualization algorithms onto multiple GPUs will enable a much larger community to leverage the power of these components [8], taking over where vendor-supported programming environments such as CUDA and CTM fall short.

Another area with great promise is the use of in-situ data processing and visualization [7]. Current tools almost exclusively operate in a post-processing mode, and this means that data must be stored at run-time, then pulled back onto the system later for visualization. This will not be feasible as dataset sizes continue to grow. Some scientists have moved to a co-processing approach where data is shipped to visualization servers at run-time, but this ties up additional resources and requires a great deal of communication during run-time. In-situ processing and visualization side-steps the deficiencies of both of these approaches by performing some analysis functions on the data during simulation run-time while it still resides on the system. These operations reduce the data size before the data is transferred to storage. Additional research is necessary to understand both what work should be performed in-situ and how these operations are best integrated into applications.

Finally, SciDAC science projects are multidisciplinary and collaborative in nature. Project members are geographically distributed. There is surprisingly very little support for remote, collaborative data analysis and visualization that would allow scientists to effectively share their experience and findings. In addition to providing the means for effective visualization at these scales, new interface designs and technologies must be invented to enable remote collaborative work.

2.1.1. Support for Complex Datasets Most of SciDAC applications involve the modeling and analysis of evolving physical phenomena, and many use adaptive, irregular grids to manage computing requirements and to better align with feature boundaries. Adaptive, irregular grids present great challenges to the subsequent visualization calculations. To visualize adaptive mesh data, the usual approach is to flatten the mesh into a uniform one, which could increase the storage and rendering requirements by one to two orders of magnitude. New, possibly hardware-accelerated, algorithms that render adaptive mesh data directly are critical to avoid the overhead of this flattening step.

The manner in which simulation datasets are organized also has an immense impact not only on the performance of subsequent visualization and analysis of the data, but also on the productivity of visualization experts, who spend a great deal of time porting novel algorithms to particular dataset organizations. A scalable, storage-independent interface [3] for multivariate data exploration and multidimensional feature extraction and analysis [1, 10] would allow visualization experts to focus on the algorithmic challenges while providing effective tools to

the largest audience.

Finally, because SciDAC applications study evolving physical phenomena, time-varying data visualization is a high-priority. The fundamental challenge of visualizing time-varying data lies in the need to fetch each time step of the data from the disk, transfer it over network, and load it into the memory of a visualization server before *any* visualization calculation can be done [6]. One viable approach is to tightly couple data reduction and rendering so that both the data transfer and rendering costs are minimized. This suggests the development of a spatio-temporal multi-resolution data management framework [9] as well as a new approach to feature extraction and compaction.

2.2. Knowledge-enabling Visualization and Analysis

Existing visualization techniques and systems were not designed to utilize information derived during the process of data visualization and analysis. As visual exploration is an inherently iterative process, information such as the visualization algorithm, chosen parameters, visualizations, and findings is knowledge critical to deeper understanding of the data under study. This knowledge can effectively aid data visualization if it is stored and organized in a structured fashion. We begin to see growing efforts to collect and use such information, especially when the cost of visualization is high or when the visualization work is collaborative in nature. Such knowledge-enabling visualization will be a key aspect of future visualization systems [5].

Tracking and visualizing the evolution of features in time-varying data is another way that extracting knowledge from datasets could revolutionize how scientists interact with very large datasets. Mathematically robust algorithms for feature detection and tracking should be developed. Currently, this development requires working closely with scientists to understand what features are important to them and how those features may be mathematically defined. In cases where features are not mathematically definable, such as vortex cores in CFD data sets, an appropriate user interface must be developed that allows scientists to construct queries and verify their hypothesis on the fly. In the long run, a framework is needed to facilitate the categorization, characterization, and creation of the semantics, taxonomies, and ontologies of features extracted from domain-specific problems. More research is necessary to understand how such a framework might be best constructed to cater to the needs of computational science.

There are numerous sources of error and uncertainty in data, especially when we have to work with reduced versions of ultra-scale data. These errors can be introduced at different stages in the process of data acquisition and analysis. New visualization methods should be designed to enable scientists to work more effectively in the presence of uncertainty by helping them "see" uncertainty and understand its source. New techniques are necessary to effectively extract this information from steps in the visualization pipeline and convey this information in the final product.

3. Future Visualization Tools

The early part of this decade saw an exodus of visualization tools from specialized monolithic SMP graphics workstations to distributed clusters comprising COTS hardware to take advantage of the latter's better scalability and economy of scale. Our current generation of parallel visualization tools has scaled well to larger clusters and terascale simulations, and the performance remains solid as we move into peta-scale. Although peta-scale visualization tools may need to adapt to exploit new parallel paradigms in hardware (such as multiple cores, multiple GPUs, and cell processors), the underlying framework of our visualization tools will remain the same. That said, our march into petascale and exascale simulations presents some fundamental changes in some of the tasks visualization tools will have to perform. First, file I/O is a more critical concern than ever before due to the size of the data. To maintain scalability

in our file reads and writes, parallel coordinated disk access must be an integrated part of our visualization tools.

Second, petascale machines will be installed in only selected locations and accessed by scientists through wide-area networks. As a consequence, remote visualization between organizations becomes a more critical component. These remote visualizations may be constrained by network connections with poor latency and bandwidth properties. The delivery of visualization must take into account the variance in network capacity for different users to maintain a reasonable level of interactivity.

Third, as the amount of data a visualization user must sift through increases, the visualization tool has an increased burden to help the user find and extract information relevant to the problem being analyzed. Users must be able to "drill-down" to make connections between a qualitative overview and a quantitative focus. Traditional views of data no longer suffice. A clear understanding of the data can only be achieved through multiple and linked representations of the data. New representations for the data, as well as those borrowed from information visualization [4] and other domains, can be used to organize data and provide correlations. Many scientific disciplines will require targeted visualization algorithms to most effectively present the data.

Finally, peta-scale computing will also provide the facility for several new modes of analysis. With higher fidelity simulations will come the ability to quantify uncertainty. Comparative analysis will become even more important for both ensuring the validity of a simulation as well as verifying the results which correspond to physical phenomenon. Ensembles of runs will have much higher fidelity or larger dimensions than ever before on peta-scale supercomputers. All three of these analyses are underrepresented in our current visualization tools, and this is a problem we must correct for effective analysis of data at peta-scale.

4. Education

One of the challenges of our research agenda is to bring together elements from many research fields—visualization, I/O, parallel computation, data management, networking, benchmarking, and so on—into results that incorporate many research thrusts into a greater whole. Equally important is our communication of these lessons to the larger research community through education and outreach.

As educators, we incorporate the lessons of our research into our coursework. But peta-scale visualization offers two important greater opportunities for education and outreach beyond the confines of our current research limited to our own research field.

First, at a high level, peta-scale visualization is in some degree a time machine. It is a computationally intense application with a large appetite for parallel computation, memory bandwidth, low-latency operation, high-performance I/O, and efficient data management. Few of today's applications have such demanding performance needs. As we build systems for today's peta-scale visualization needs, we learn about how to build more general systems for tomorrow's general-purpose application needs. Our education and outreach is designed to bring those lessons to the broader computation community. We thus place special emphasis in our outreach toward understanding and teaching next-generation techniques in our tutorials.

Second, visualization provides a unique opportunity to bridge the gap between applications and computing. Whether because of an incompatible vocabulary or a lack of contact in professional venues, too often these two groups do not effectively communicate. Visualization is a vital tool for allowing application developers to see and understand the output of computation, and for computation developers to better understand the needs and goals of the application developers. Consequently, we target a broad range of tutorial venues, particularly venues with a wide spectrum of attendees, such as Supercomputing and SciDAC, to ensure that we evangelize large-scale visualization as a key component in bringing computing and applications

closer together.

5. Conclusion

As the computational science community moves towards petascale and eventually exascale computing platforms, scientific visualization faces two challenges. First, new algorithms and frameworks must be developed to address the size, complexity, and variety of data that must be analyzed in order to extract knowledge for scientific discovery. Second, these technologies must make it into the hands of scientists, despite the inevitable complexity of the tools. The IUSV is uniquely positioned to lead these efforts and to work with the VACET team to ensure that novel and successful technologies make it into production toolsets.

Acknowledgments

We would like to thank all our DOE and SciDAC application scientist collaborators. You are all critical in our understanding of the needs of the community and provide invaluable feedback during the research and development process.

This work was supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, Office of Science, U.S. Dept. of Energy, under Contract DE-AC02-06CH11357.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

[1] Akiba, H., and Ma, K.-L. A tri-space visualization interface for analyzing time-varying multivariate volume data. In *Proceedings of Eurographics/IEEE VGTC Symposium on Visualization* (May 2007).

[2] DOE/ASCR Visualization and Analytics Workshop, Salt Lake City, Utah, June 7-8, 2007. (http://www.sci.utah.edu/vaw2007).

[3] Glatter, M., Mollenhour, C., Huang, J., and Gao, J. Scalable data servers for large multivariate volume visualization. *IEEE Transactions on Visualization and Computer Graphics 12*, 5 (2006), 1291–1299.

[4] Jones, C., Ma, K.-L., Sanderson, A., and Myers Jr., L. R. Visual interrogation of gyrokinetic particle simulations. *Journal of Physics (also Proceedings of SciDAC 2007 Conference)* (2007).

[5] Ma, K.-L. Visuaizing visualizations: User interfaces for managing and exploring scientific visualization data. *IEEE Computer Graphics and Applications 20*, 5 (2000), 16–19.

[6] Ma, K.-L. Visuaizing time-varying volume data. *IEEE Computing in Science and Engineering 5*, 2 (2003), 34–42.

[7] Ma, K.-L., Wang, C., Tikhonova, A., and Yu, H. In-situ visualization. *Journal of Physics (also Proceedings of SciDAC 2007 Conference)* (2007).

[8] Owens, J. D. Towards multi-gpu support for ultra-scale visualization. *Journal of Physics (also Proceedings of SciDAC 2007 Conference)* (2007).

[9] Wang, C., Gao, J., Li, L., and Shen, H.-W. A multiresolution volume rendering framework for large-scale time-varying data visualization. In *Proceedings of International Workshop on Volume Graphics* (2005), pp. 11–19.

[10] Wang, C., and Shen, H.-W. LoD map, a visual interface for navigating multiresolution volume visualization. *IEEE Transactions on Visualization and Computer Graphics 12*, 5 (2006), 1029–1036.