

Terascale Visualization: Approaches, Pitfalls and Issues

Organizers

Carol Hunter, Lawrence Livermore National Laboratory
Roger Crawfis, The Ohio State University

Panelists

Michael Cox, NASA Ames Research Center
Roger Crawfis, The Ohio State University
Bernd Hamann, University of California, Davis
Chuck Hansen, University of Utah
Mark Miller, Lawrence Livermore National Laboratory

INTRODUCTION

Massively Parallel Supercomputers are once again quickly outpacing our ability to organize, manage and understand the prodigious amounts of data they generate. Graphics technology and algorithms have greatly aided in analyzing the modest datasets of years past, but rarely with enough interactivity to squelch the end-user's exploratory questions. Will computer graphics and scientific visualization or even computational science proceed as status quo or are new paradigm shifts needed? What is the architecture of tomorrow's high-end visualization systems? How much data can we even expect to pull off of these massively parallel machines? What are the new computer graphics technologies that can aid in terascale visualization? This panel, leveraging the panelists past experience and their current knowledge of the field, will provide visions (or dilemma's) for what the next stage or stages of scientific visualization and data management will look like.

STATEMENTS

Michael Cox

In scientific visualization, input data sets are generally quite large. Simulation results **today** can surpass 100 Gbytes, and these are expected to scale with the ability of supercomputers to generate them. Observational data (satellite, experimental, remote sensor) can be larger; the Earth Observing System project is expected to generate several Tbytes of data per day that must be reduced, archived, and made available for general dissemination. The focus in scientific visualization has been almost exclusively on the algorithms required for interactive or real-time traversal and rendering; there has been a dearth of attention paid to the systems issues required to handle these data sizes. For single large data sets, the problem is one of memory hierarchy, and we need research and algorithms specifically targeted for visualization when any data set 1) doesn't fit in main memory, 2) doesn't fit on local disk, 3) doesn't fit on the largest remote disk available at your installation! In addition, in some areas of scientific visualization, the number of data sets that must be searched and processed presents more of a problem than the size of any particular data set. (The Earth Observing System will archive on the order of 50,000, 50 Mbyte files a day that must be available for later queries and visualization). To deal with both types of "big data" challenges, it is clear that significantly more work must be done on big data management; it is clear that significantly better systems must be built and deployed.

I will discuss the work that has been done, the possible technologies that may be applicable, and suggest research and development directions that may bear fruit.

0-8186-8262-0/97 \$10.00 Copyright 1997 IEEE.

Roger Crawfis

Large-scale computational science problems being studied on today's massively parallel supercomputers generate terabytes of raw computational data. This data must typically be analyzed to either ensure the computational simulation codes are functioning properly, or to gain insight in the problem being studied. Much research and development has gone into designing computer graphics and scientific visualization algorithms for processing this raw data, extracting meaningful fragments and representing it visually to the end user or code developer. The pace of supercomputing, and now massively parallel computing, continues to outpace our ability to post-process the raw data and generate the resulting scientific imagery for all but the smallest simulations or the simplest visualizations.

Rather than focusing on trying to remedy this endless battle, let's approach the problem from a more novel direction. We propose to accomplish this by intelligently replacing the terabyte of very coarsely sampled time dumps with perhaps an equal amount of finely sampled and highly compressed multi-resolution time animations. The win here, is the transformation of the cumbersome and bulky raw data into a format that can be easily streamed to the user's desktop (or alternative interaction environment), and immediately presented. When compared with the 20 minutes to calculate and display a single 3D iso-contour surface, this low-latency approach greatly aids in the user's ability to form mental models of their simulation, resulting in greater comprehension. A further advantage of this approach is its alignment with current market trends towards digital video and high-definition television.

Several questions immediately rise to the forefront:

- How do we guarantee the low latency?
- How can we be assured that the chosen set of video's will address all (or a substantial fraction) of the user's inquiries?
- What extensions to current imaging technologies are required to support image-based scientific visualization?
- What system-level mechanisms are required for efficient retrieval of the compressed data?

In order to support this new paradigm shift to Image-based Approach to Scientific Visualization, advances are required in three key areas: scientific visualization, multi-resolution video compression technologies, and delivery mechanisms for hi-resolution, multi-dimensional data. I will address my vision for each of these areas and encourage a lively debate on the costs and benefits of this proposed path.

One of the most challenging and important problems that the science and engineering communities are facing today – and even more so in the future – are representing, visualizing, and interpreting very large data sets. Very large data sets result from computer simulations of complex physical phenomena (e.g., climate modeling, ocean modeling) or from high-resolution imaging (e.g., satellite imaging, medical imaging). The technology currently being used to represent extremely large data sets is inappropriate for interactive, efficient, and detailed data analysis and visualization. It is impossible for a user of a visualization system to *navigate* through a data set consisting of millions of points and analyze it entirely. We will have to develop new ideas and a vision to overcome some of the problems associated with the representation of very large data sets. A crucial aspect towards the solution of massive data set analysis and visualization problems will be the merging of ideas from approximation theory and geometric modeling (e.g., splines), computational geometry (e.g., tessellations), pattern recognition, statistics, and other related fields. One of the most important issues will be the representation of very large data sets by different types of hierarchies that will facilitate efficient data storage at visualization at various levels of detail.

Chuck Hansen

Although massively parallel computation and terabyte/petabyte datasets are definitely not common place (nor will they be in the near future), they still exist in two predominate application areas: large government labs and information databases. The problem of attempting to comprehend the data which resides on or is generated by these diverse systems completely overwhelms current visualization technology. Attempts to address this problem have focused on individual problems such as hierarchical methods which require all of the data to be accessed in order to build the hierarchy. A more holistic view of the visualization process is required for manipulation of these humungus amounts of data. This includes factoring in system level feedback, such as network bandwidth and latency, and utilizing this self-monitoring system to provide an adaptive solution. This includes approaching the visualization process from the top-down rather than the bottom up (as is the case with current hierarchical methods) as well as exploiting the local viewpoint updates provided by image based rendering techniques.

Chuck Hansen (Alternative Position)

Massively parallel supercomputers are a thing of the past and are quickly being replaced by cheap and functional PCs. Besides, the problem with large data is that it completely ignores the current computational trends. That is, why should we run such large problems when it is so cost effective to run on PCs. For that matter, we shouldn't bother writing software since we can purchase what we need from Microsoft. Of course, it may not provide the functionality but then we can upgrade to the next release thereby solving all of our problems ... including the large data problem.

Mark Miller

The human visual system is extremely well suited and can be easily trained to detect anomalies of various kinds in images and image sequences. However, the volume of data that can be generated in a terascale computing environment exceeds by several orders of magnitude that which can be viewed in any single image or reasonably short sequence of images on conventional display devices. Using visualization to detect and assess calculational anomalies of a simulation in such an environment is a lot like using a microscope to find the needles, or proving that none exist, in a haystack.

One approach is to employ multi-resolution visualization techniques which provide views of the data in a scale space commensurate with output image resolution and/or human perception. These techniques show promise in helping to assess large scale anomalies and in helping to speed up visualization operations to the point of providing interactive feedback. However, when the aggregation rules for building lower resolution representations minimize the mean squared error, the same approaches also help to obscure small scale anomalies.

A second approach is to employ **direct** automated feature extraction techniques which sift through the data and return the existence and location of specific types of anomalies for which detection algorithms can be written. Following this step, visualization can then be confined to only those portions of the data that feature extraction methods have flagged as anomalous. The shortcomings of this approach lie in the feasibility of designing and implementing feature extraction algorithms and it does not leverage the capabilities of the human visual system in the detection process.

A third approach is to employ **indirect** feature extraction techniques which compute a derived metric on the data set known to amplify certain class of anomalies. This derived metric is a new data variable which is large where the likelihood of the anomaly is high and small where it is low. Following this step, a multi-resolution approach where the aggregation rules involve minimizing the maximum error, will reveal even small scale regions of the data set where the derived variable is large. Upon visualization of this derived data, the human visual system can be leveraged in performing the anomaly detection.

In order to proceed in the aforementioned directions, work is required on two fronts. Multi-resolution representation, storage, processing and rendering of three dimensional, perhaps unstructured gridded data and derivation of useful metrics for indirect feature extraction.

BIOGRAPHIES

Michael Cox

Michael Cox is a senior research scientist at MRJ/NASA Ames Research Center, where he leads the Data Exploitation group working on "big data" in scientific-visualization. Prior to NASA, he has worked at Intel, S3 Inc., Sun Microsystems, Advanced Computer Communications, and Altas Corporation in graphics hardware architectures, networking, distributed systems, and solar energy. He holds a Ph.D. in Computer Science from Princeton University, and a B.A. in Biology from the University of California at Santa Cruz.

Roger Crawfis

Roger A. Crawfis is an Assistant Professor of Computer Science at the Ohio State University in Columbus, OH. From 1984 to 1996 he worked at the Lawrence Livermore National Laboratory. His primary interests are visualization, flow visualization, image manipulation, and computer graphics. Crawfis received his B.S. in computer science and applied mathematics from Purdue University, and an M.S. and Ph.D. in computer science from the University of California, Davis. Crawfis is a member of the Institute of Electrical and Electronics Engineers (IEEE).

Bernd Hamann

Bernd Hamann is an Associate Professor of Computer Science and Co-Director of the Center for Image Processing and Integrated Computing (CIPIC) at the University of California, Davis, and an Adjunct Professor of Computer Science at Mississippi State University. From 1991 to 1995 he was a faculty member in the Department of Computer Science at Mississippi State University and a research faculty member at the NSF Engineering Research Center for Computational Field Simulation. His primary interests are visualization, computer-aided geometric design (CAGD), and computer graphics. He is the author of numerous publications and has presented his research at leading conferences in the U.S. and in Europe.

Hamann received a B.S. in computer science, a B.S. in mathematics, and an M.S. in computer science from the Technical University of Braunschweig, Germany. He received his Ph.D. in computer science from Arizona State University in 1991.

Hamann was awarded a 1992 Research Initiation Award by Mississippi State University, a 1992 Research Initiation Award by the National Science Foundation, and a 1996 CAREER Award by the National Science Foundation. In 1995, he received a Hearin-Hess Distinguished Professorship in Engineering by the College of Engineering, Mississippi State University.

Hamann is a member of the Association for Computing Machinery (ACM), the Institute of Electrical and Electronics Engineers (IEEE), and the Society for Industrial and Applied Mathematics (SIAM).

Chuck Hansen

Chuck Hansen is a Research Associate Professor of Computer Science at the University of Utah. From 1989 to 1997, he was a technical staff member in the Advanced Computing Laboratory located at Los Alamos National Laboratory where he was project leader for the scientific visualization environment for the DOE High Performance Computing Research Center. During 1996, he was the principle investigator for scientific visualization for the DOE ASCI program. His research interests include scientific visualization, massively parallel processing, parallel computer graphics algorithms, 3D shape representation, and computer vision. He received his BS in Computer Science from Memphis State University in 1981 and a PhD in Computer Science from the University of Utah in 1987. He was a Bourse de Chateaubriand PostDoc Fellow at INRIA, Rocquencourt France, in 1987 and 1988. He is a member of IEEE-CS and ACM-SIGGRAPH.

Mark Miller

Dr. Mark C. Miller recently completed work as Lawrence Livermore's Principle Investigator for Visualization in the Accelerated Strategic Computing Initiative program. Mark has worked for the past two years in B support group which provides computer science support to B Division, a nuclear weapons design division, at Livermore Labs. Mark is currently developing multi-resolution visualization techniques aimed at providing interactive visualization of terascale data sets. Mark is also involved in scientific data management to support visualization. Mark completed is doctoral and master's degrees in the department of Electrical Engineering at the University of California, Davis. His doctoral work focuses on multi-resolution representations and compression of surface data to maintain real time rendering requirements. His master's thesis develops theory for recovery of under-sampled periodic signals.