DOE's SciDAC Visualization and Analytics Center for Enabling Technologies -- Strategy for Petascale Visual Data Analysis Success

CTWatch Quarterly

November 2007

E. Wes Bethel[1], Chris Johnson[5]
Cecilia Aragon[1], Prabhat[1], Oliver Rübel[1], Gunther Weber[1]
Valerio Pascucci[2], Hank Childs[2], Peer-Timo Bremer[2], Brad Whitlock[2]
Sean Ahern[3], Jeremey Meredith[3], George Ostrouchov[3]
Ken Joy[4], Bernd Hamann[4], Christoph Garth[4]
Martin Cole[5], Charles Hansen[5], Steven Parker[5], Allen Sanderson[5], Claudio Silva[5], Xavier Tricoche[5]

[1]Lawrence Berkeley National Laboratory
[2]Lawrence Livermore National Laboratory
[3]Oak Ridge National Laboratory
[4]University of California, Davis
[5]University of Utah

Introduction

Galileo Galilei (15 February 1564 -- 8 January 1642) has been credited with fundamental improvements to early telescope designs that resulted in the first practically usable instrument for observing the heavens. With his "invention," Galileo went on to many notable astronomical discoveries: the satellites of Jupiter, sunspots and the rotation of the sun, and proved the Copernican heliocentric model of the solar system (where the sun, rather than the earth, is the center of the solar system). These discoveries, and their subsequent impact on science and society, would not have been possible without the aid of the telescope – a device that serves to transform the unseeable into the seeable.

Modern scientific visualization, or just visualization for the sake of brevity in this article, plays a similarly significant role in contemporary science. Visualization is the transformation of abstract data, whether it be observed, simulated, or both, into readily comprehensible images. Like the telescope and other modern instruments, visualization has proven to be an indispensable part of the scientific discovery process in virtually all fields of study. It is largely accepted that the term "scientific visualization" was coined in the landmark 1987 report [1] that offered a glimpse into the important role visualization could play in scientific discovery.

Visualization produces a rich and diverse set of output – from the x/y plot to photorealistic renderings of complex multidimensional phenomena. It is most typically "reduced to practice" in the form of software. There is a strong, vibrant, and productive worldwide visualization community that is inclusive of commercial, government and academic and academic interests.

The field of visualization is as diverse as the number of different scientific domains to which it can be applied. Visualization software design and engineering studies and solves what are essentially computer science problems. Much of visualization algorithm conception and design shares space with applied mathematics. Application of visualization concepts (and software) to specific scientific problems to produce insightful and useful images overlaps with cognitive psychology, art, and often the scientific domain itself.

Figure 1.   Visualization offers the ability to "see the unseeable."' This image shows visualization of coherent flow structures in a large scale delta wing dataset: Volume rendering of regions of high forward (red) and backward (blue) Finite Time Lyapunov Exponent. Coherent structures appear as surfaces corresponding to the major vortices developing over the wing along the leading edge [2]. Occlusion is a limitation that can be addressed with cropping or clipping. (Image courtesy of X. Tricoche, University of Utah and C. Garth, University of California – Davis)

In the present day, the U.S. Department of Energy has a significant investment in many science programs. Some of these programs, carried out under the Scientific Discovery through Advanced Computing (SciDAC) program [3], aim to study via simulation scientific phenomena on the world's largest computer systems. These new scientific simulations, which are being carried out on fractional-petascale sized machines today, generate vast amounts of output data. Managing and gaining insight from such data is widely accepted as one of the bottlenecks in contemporary science [4]. As a result, DOE's SciDAC program includes efforts aimed at addressing data management and knowledge discovery to complement the computational science efforts.

The focus of this article is on how one group of researchers – the DOE SciDAC Visualization and Analytics Center for Enabling Technologies (VACET) – is tackling the daunting task of enabling knowledge discovery through visualization and analytics on some of the world's largest and most complex datasets and on some of the world's largest computational platforms. As a Center for Enabling Technology, VACET's mission is the creation of usable, production-quality

visualization and knowledge discovery software infrastructure that runs on large, parallel computer systems at DOE's Open Computing facilities and that provides solutions to challenging visual data exploration and knowledge discovery needs of modern science, particularly the DOE science community.

Why Visualization Works So Well

One of the reasons that scientific visualization, and visual data analysis, has proven to be highly effective in knowledge discovery is because it leverages the human cognitive system. Pseudocoloring, a staple visualization technique, performs a mapping of data values to colors in images to take advantage of this very ability. Figure 2 is a good example, where high data values are mapped to a specific color that attracts the eye. Additionally, a very clear 3D structure becomes apparent in this image; it would be virtually impossible to "see" such structure by looking at a large table of numbers. While Figure 2 shows a 3D example, we are all familiar with 2D versions of this technique: the weather report on the evening news often shows psuedocolored representations of temperature or levels of precipitation overlaid on a map.
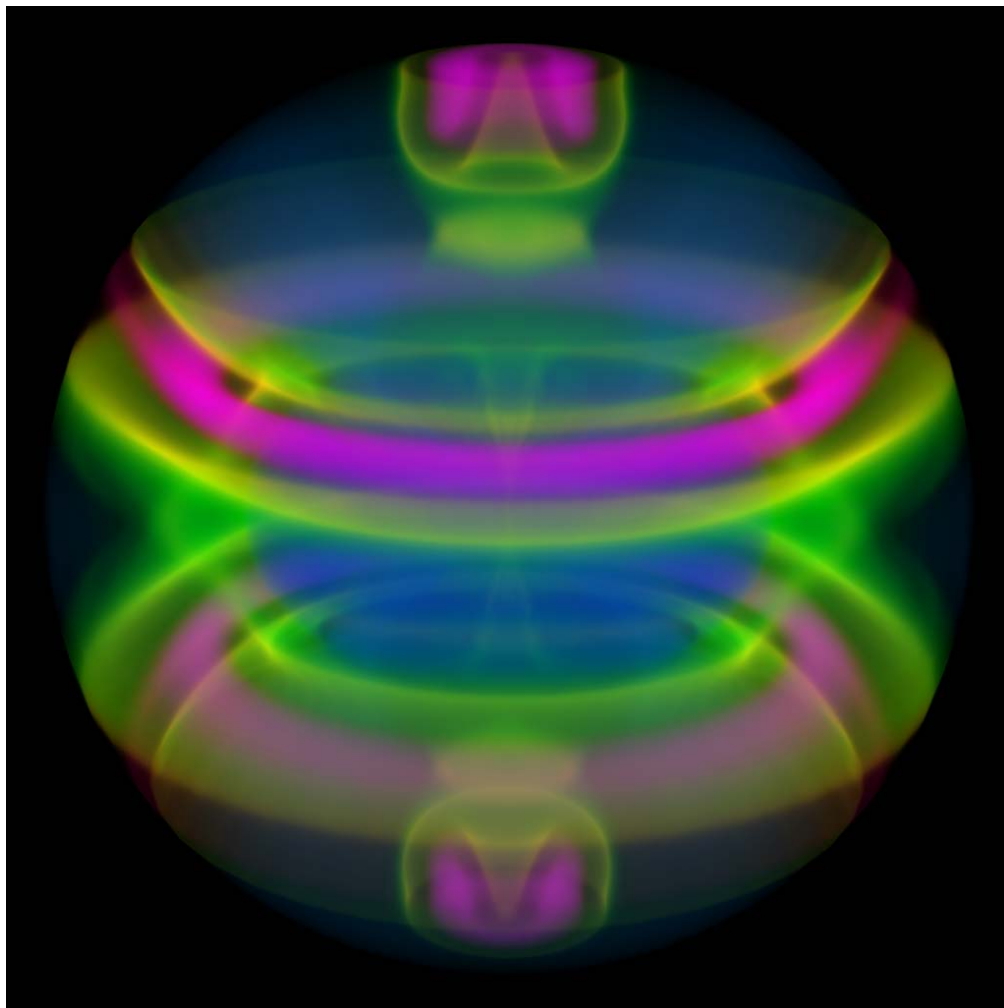

Figure 2.   Two types of "features" are immediately visible in this image showing the entropy field of a radiation/hydrodynamic simulation that models the accretion-induced collapse of a star, a phenomena that produces supernovae. One "feature" is the "sandwiching" of high values of entropy between lower values. The other is an overall sense of 3D structure. (Simulation data courtesy of Adam Burrows, University of Arizona, SciDAC

Surviving the Data Tsunami

Many "tried and true" visualization techniques – like using psuedocoloring to map scalar data values to color – do a great job of leveraging the human cognitive system to accelerate discovery and understanding of complex phenomena. However, we are faced with some difficult challenges when considering the notion of using visualization as a knowledge discovery vehicle on very large datasets. One of many challenges is limited human cognitive bandwidth, which is conveyed in the notional chart shown in Figure 3. This chart conveys that while our ability to generate, collect, store and analyze data grows at a rate tracking the increase in processor speed and storage density, we as humans have fixed cognitive capacity to absorb information. Given that our ability to generate data far exceeds what we can possibly understand, one major challenge for "petascale visual data exploration and analysis" is how to effectively "impedance match" between "limitless data" and a fixed human cognitive capacity.
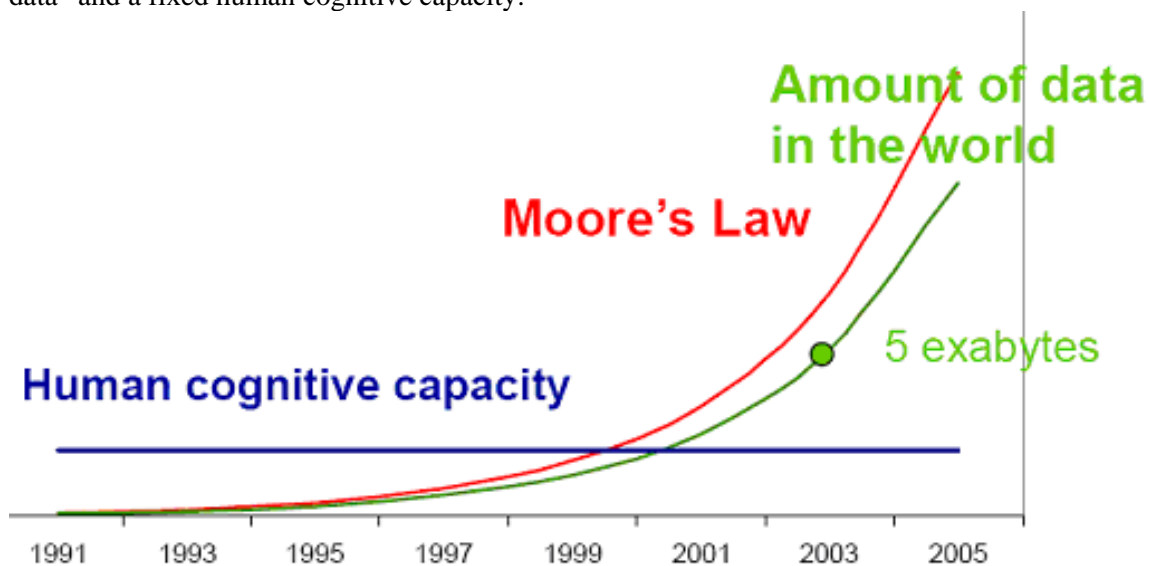
Figure 3. While our ability to generate, collect, store, and analyze data grows at a rate that tracks the increase in processor speed and storage density, our ability as humans to absorb information remains fixed (Illustration adapted from a slide by J. Heer, PARC User Interface Research Group).

"What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it." *Herb Simon*, as quoted by Hal Varian [18].

In the context of our work – namely, petascale visual data analysis – we are faced with several dilemmas. First, even if we could simply scale up our existing tools and algorithms so they would operate at the petascale rather than the terascale, would the results be useful for knowledge discovery? Second, if the answer to the first question is "no," then how can we help to "allocate attention efficiently among the overabundance of information"?

Let's examine the first question a bit more closely. First, let's assume that we're operating in a gigabyte-sized dataset (10^9 data points), and we're displaying the results in a monitor that has,

say, 2 million pixels ($2*10^6$ pixels). For the sake of discussion, let's assume we're going to create and display an isosurface of this dataset. Studies have shown that on the order of about $N^{2/3}$ grid cells in a dataset of size N^3 will contain any given isosurface [13]. In our own work, we have found this estimate to be somewhat low – our results have shown the number to be closer to $N^{0.8}$ for N^3 data. Also, we have found an average of about 2.4 triangles per grid cell will result from the isocontouring algorithm [14]. If we use these two figures as lower and upper bounds, then for our gigabyte-sized dataset, we can reasonably expect on the order of between about 2.1 and 40 million triangles for many isocontouring levels. At a display resolution of about 2 million pixels, the result is a depth complexity – the number of objects at each pixel along all depths – of between one and twenty.

With increasing depth complexity come at least two types of problems. First, more information is "hidden from view." In other words, the nearest object at each pixel hides all the other objects that are further away. Second, if we do use a form of visualization and rendering that supports transparency – so that we can, in principle, see all the objects along all depths at each pixel – we are assuming that a human observer will be capable of distinguishing among the objects in depth. At best, this latter assumption does not always hold true, and at worst, we are virtually guaranteed the viewer will not be able to gain any meaningful information from the visual information overload.

If we scale up our dataset from gigabyte (10^9) to terabyte (10^{12}), then we can expect on the order of between 199 million and 9.5 billion triangles representing a depth complexity ranging between about 80 and 4700, respectively. Regardless of which estimate of the number of triangles we use, we end up drawing the same conclusion: depth complexity, and correspondingly scene complexity and human workload, grows linearly with the size of the source data. Even if we are able to somehow display all those triangles, we would be placing an incredibly difficult burden on the user. They will be facing the impossible task of visually trying to locate "smaller needles in a larger haystack."

The multi-faceted approach we're adopting takes square aim at the fundamental objective: help the scientific researchers more quickly and efficiently do science. In one view, one primary tactical approach that seems promising is to help focus user attention on easily consumable images from the large data collection. We do not have enough space in this brief article to cover all aspects of our team's effort in this regard. Instead, we provide a few details about a couple of especially interesting challenge areas.

Query-Driven Visualization

The term "query-driven visualization" (QDV) refers to the process of limiting visual data analysis processing only to "data of interest [7]." In brief, QDV is about using software machinery combined with flexible and highly useful interfaces to help reduce the amount of information that needs to be analyzed. The basis for the reduction varies from domain to domain, but boils down to "what subset of the large dataset is really of interest for the problem being studied." This notion is closely related to that of "feature detection and analysis," where "features" can be thought of as subsets of the larger population that exhibit some characteristics that are either intrinsic to individuals within the population (e.g., data points where there is high pressure and high velocity) or that are defined as relations between individuals within the population (e.g, the temperature gradient changes sign at a given data point).

For the purposes of our discussion here, we will focus on the first category of features. The second category is also of great interest to our team, where we have developed new technologies

for topological data analysis [15] that have proven very useful as the basis for enabling scientific knowledge discovery.

Broadly speaking, QDV consists of three broad conceptual elements. One is how one goes about "specifying interesting." Another is how one displays and analyzes that subset of data. Yet another is the process of storing, indexing, querying and retrieving data subsets from large data archives.

Specifying Queries

In many scientific data analysis applications, "interesting" data can be defined by compound boolean range queries of the form "*(temperature > 1000) AND (0.8 <= density <= 1.0)*". Obviously, one could manually enter such a SQL-like query, but doing so is somewhat clumsy from an interface perspective, but also requires that the user know something about the data characteristics. In many instances, the users are quite familiar with their data, so the expectation of a priori knowledge is not unreasonable. Rather than typing in queries, we propose that a visual interface for specifying queries will result in greater scientific productivity and better serve our mission of enabling data exploration and knowledge discovery.

We have implemented several different types of visual interfaces for specifying queries. The general theme in these implementations is that the visual interface helps the user to formulate queries while at the same time gaining an overall sense of data characteristics. This type of interaction is a variation on a well-known usability design principle called "context and focus," where a given presentation affords the opportunity to see overviews of data (the context) as well as details about specific data of interest (the focus). Numerous works have applied this principle to the effective navigation of complex dataspaces, e.g., application to browsing of hierarchical filesystems [16].

One example for formulating queries along these lines is an application for exploration of large collections of particle-based datasets produced by the Gyrokinetic Turbulence Code (GTC), which is used to model microturbulence in magnetically confined fusion plasmas [17]. Output from GTC consists of on the order of tens of millions of particles per timestep on present-day computational platforms; this figure is expected to rise at a rate commensurate with growth in computational capacity. From this output, fusion researchers are interested in studying various types of phenomena: formation, evolution and analysis of turbulent structures (eddies, vortices, etc.); how particle "trapping" and "untrapping" in magnetic fields through microturbulence leads to an erosion of energy efficiency.

We clearly don't want to present an image of the entire dataset at each timestep – the result would be a very cluttered and unintelligible display. Instead, we want to offer the ability for a fusion scientist to focus visual analysis on subsets of data. The result, which is shown below in Figure 4, is an effective context-and-focus interface for rapidly selecting subsets of particles for display.
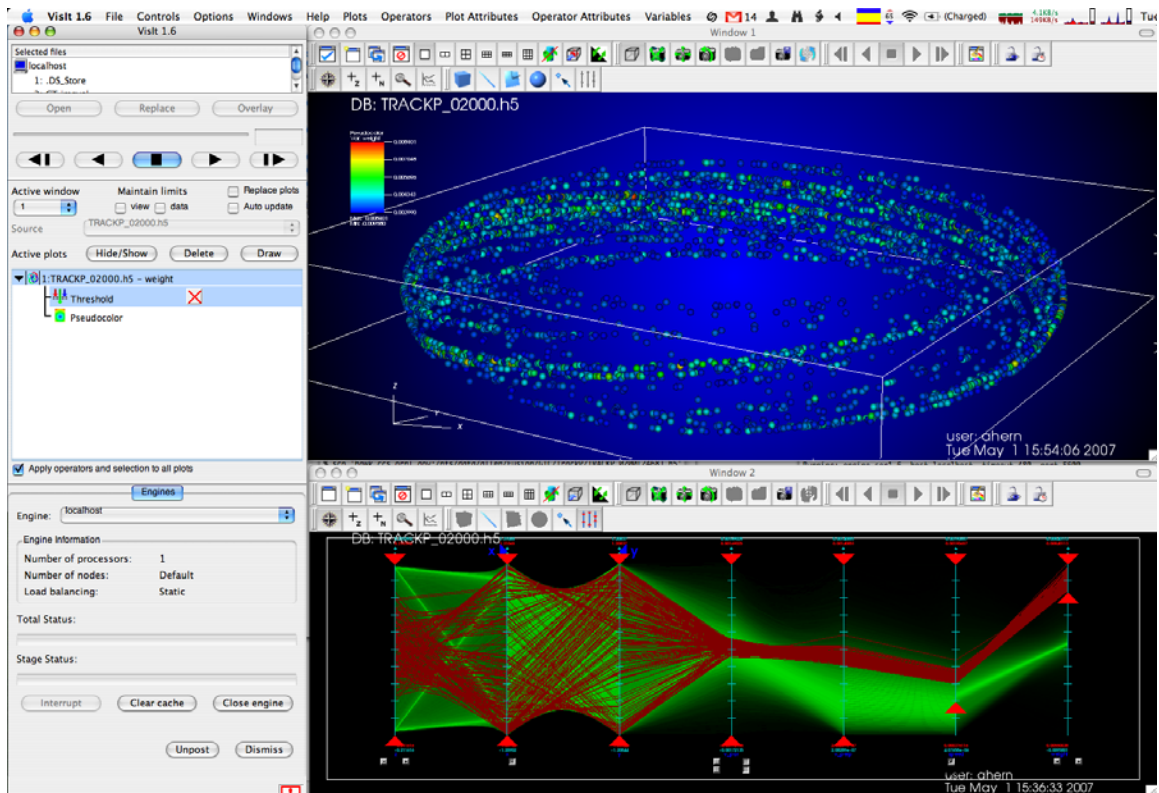
Figure 4. For visual exploration and analysis of GTC data, our implementation provides a visual interface for selecting subsets of particles that meet a set of user-defined criteria. Here, "interesting" is defined as those data points that satisfy a set of multivariate range combinations via the parallel coordinates interface (lower image). The subset satisfying these range conditions then appears in the physical view (top image), where the view may be manipulated, the color transfer changed to draw attention to specific particles based upon other user-defined criteria, or subject to other types of visual or traditional analysis. (Image courtesy S. Ahern, Oak Ridge National Laboratory)

These concepts can be applied to other types of data in other scientific domains, such as exploring the relationships between gene expression levels in cells of a developing organism as shown in Figure 5. These ideas, when combined with multiple linked views where updates in one display are then propagated to other views of the same dataset, offer an extremely powerful framework for rapid exploration of complex data [8].
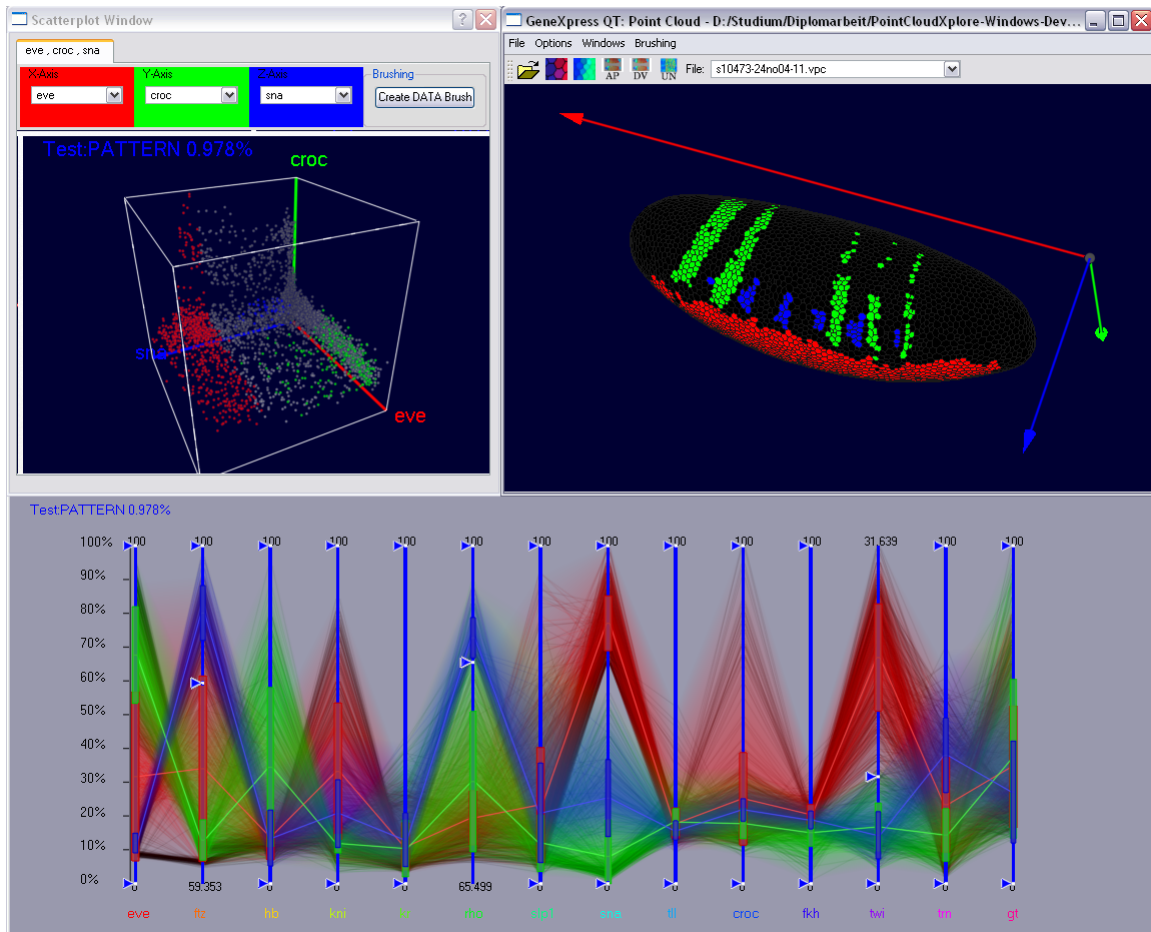
Figure 5. Here, we define three groups of "interesting" data – colored red, green, and blue – with a parallel coordinates interface (bottom pane), and those data points satisfying the multivariate range condition appear in the physical view (upper right). Here, we introduce a third linked view (upper left) that shows a 3D scatterplot – each of the data points from the three "groups of interesting" are colored red, green or blue (according to which "group of interesting" they belong). The three axes of the scatterplot are the expression levels for three specific genes; there are on the order of about 20 different gene expression levels per cell in this dataset – we picked three genes from the group of 20 for the purposes of display. This type of linked view presentation is very helpful in conveying different types of relationships in complex data. (Image courtesy of Oliver Rübel, Lawrence Berkeley National Laboratory)

High Performance Implementation

So far, we've discussed how one might go about specifying queries, or "defining interesting," and have also shown a couple of different ways to present the results that show only "the interesting data." Here, we want to turn our attention to the underlying machinery that makes this kind of approach feasible in high performance implementations suitable for use with very large datasets.

All Computer Science undergraduates are introduced to the idea of binary trees and their use as an indexing data structure. Briefly, if you have a sorted array of data of *N* items, you can construct a binary tree that will have *N-1* nodes and *N* leaves where each interior node partitions the data in deeper nodes and leaves into two groups – "greater than" and "less than or equal to" the value of a key. Once you have constructed this data structure, the search for the data record having the value of some key is performed in log_2N search steps assuming an optimal, or

balanced tree. This basic idea – called tree-based indexing – is widely used in many types of relational and object-oriented database systems. One obvious limitation of this type of approach when considering very large data is that the size of the indexing structure – the tree – is linear with respect to the size of the dataset being indexed. As this size grows larger, we clearly don't want to incur a commensurately larger storage cost for our search indices. Another problem, which may not be quite as obvious, is that these tree-based approaches require the original data to be sorted. For scientific data, where you typically write the data once then examine it over and over again, this may not be a serious limitation. In some instances, it may simply be impractical to sort the data.

Of more concern is the so-called "Curse of Dimensionality [9]" The previous paragraph calls out that the storage complexity for a tree-based structure is $O(N)$ when there are N data points. If these data points, or records, have two variables, and we want to create a two-dimensional tree that spans both variables, we end up with a storage complexity of $O(N^2)$. If there are three variables, the storage requirements are of $O(N^3)$. The basic premise is that storage requirements for tree-based indices grow exponentially with respect to the number of variables being indexed. Many modern simulations routinely have on the order of 100 variables that are computed and saved at each time step. It should be obvious that tree-based indexing is simply not practical for large and complex scientific data.

This well-known problem has received a great deal of attention from our colleagues in the field of scientific data management. They have developed a unique technology called "compressed bitmap indices" that have very favorable storage and search complexity [10]. This technology has been applied with great success to index/query problems of some of the world's largest datasets [11]. In a series of collaborative research projects, members of VACET and DOE's Scientific Data Management Center have demonstrated the practicality of combining fast bitmap indexing with high performance visual data analysis to implement a novel approach to query-driven visualization applied to visual data analysis of problems in combustion modeling [7] and large-scale network traffic analysis [12].

## Adaptive Mesh Refinement Visualization

Adaptive Mesh Refinement (AMR) techniques combine the compact, implicitly specified structure of regular, rectilinear with the adaptivity to changes in scale of unstructured grids. AMR has proven particularly useful for modeling multiscale computational domains that span many orders of magnitude of spatial or temporal scales by focusing solvers on regions where "interesting" physics or chemistry occur. Such domains include applications like astrophysics supernova modeling, where the simulation endeavors to model phenomena that occur at scales ranges from sub-kilometer to interplanetary. AMR avoids the inefficiencies inherent in attempting to model this vast computational domain at a single, fine, homogeneous resolution.

Handling AMR data for visualization is challenging, since coarser information in regions covered by finer patches is superseded and replaced with information from these finer patches. During visualization it becomes necessary to manage selection of which resolutions are being used. Furthermore, it is difficult to avoid discontinuities at level boundaries, which, if not properly handled, lead to visible artifacts in visualizations. Due to these difficulties, AMR support as first class data type in production visualization tools has been lacking despite the growing popularity of AMR-based simulations [6].
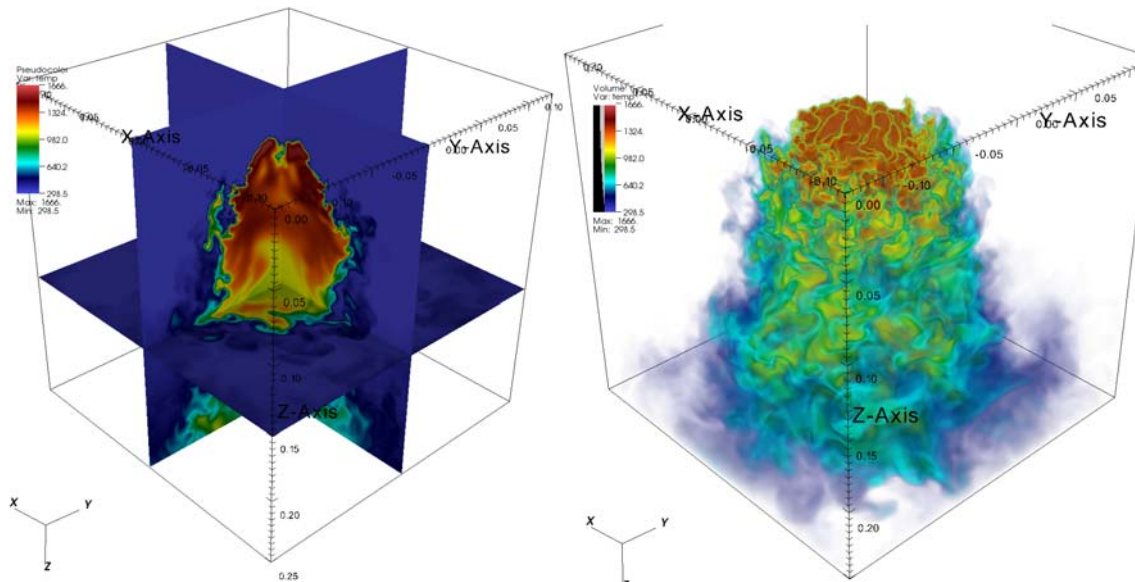
Figure 6. Production-quality visualization of data from an AMR-based simulation of a hydrogen flame. The left panel shows three orthogonal slices colored by temperature (blue is colder, red is hotter). The right panel shows an image produced with volume rendering of the same variable from this dataset. (Simulation data courtesy M. Day and J. Bell, Lawrence Berkeley National Laboratory; images courtesy G. Weber, Lawrence Berkeley National Laboratory).

Through interactions with our computational science stakeholders, VACET is providing production-quality, parallel capable software providing capabilities that fulfills needs in exploratory, analytical and presentation AMR visualization. Our deployment software – VisIt [5] – is an open source visualization tool that accommodates AMR as first class data type. VisIt handles AMR data as a special case of "ghost data," i.e., data that is used to make computations more efficient, but which is not considered to be part of the simulation result. VisIt tags cells in coarse patches that are available at finer resolution as "ghost" cells, allowing AMR patches to retain their highly efficient native format as rectilinear grids. VisIt offers a rich set of production-quality functions, such as pseudocolor and volume rendering plots (Figure 6), for visualization and analysis of massive scale data sets, making it an ideal candidate to replace specialized AMR visualization tools.
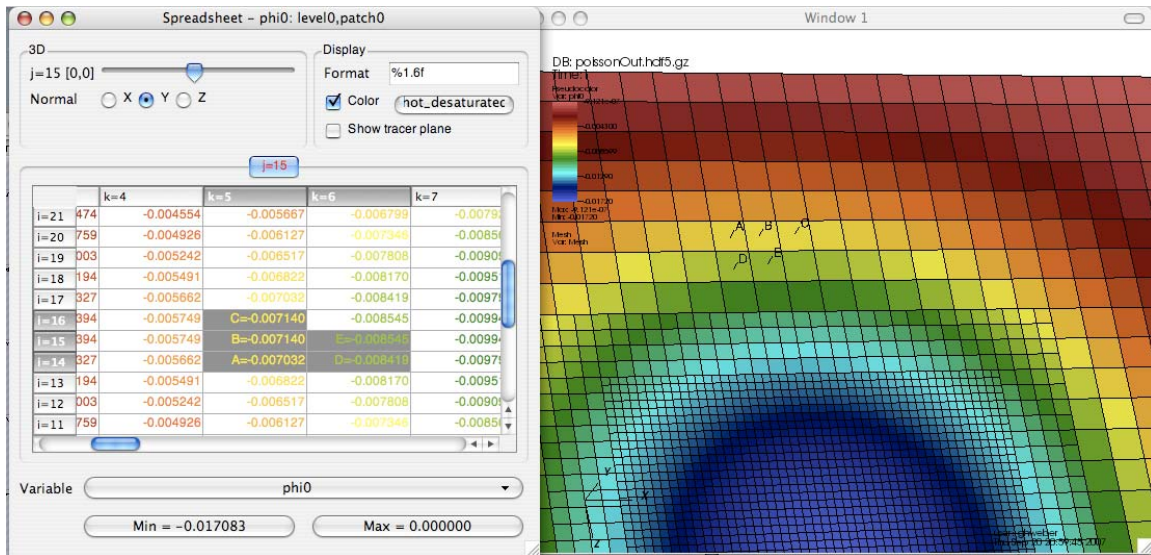
Figure 7. Spreadsheet plots are an important tool for debugging AMR codes. They support direct viewing of numerical data in patch cells. VisIt labels selected cells both in Spreadsheet and 3D visualizations allowing users to recognize correspondences quickly and effectively.

Recently, VACET has focused attention on implementing a set of essential debugging features in VisIt so that one of our stakeholders, the DOE SciDAC Applied Partial Differential Equations Center (APDEC), can fully migrate from their project-written and maintained visual data analysis software (ChomboVis) to VisIt. This migration will result in two benefits crucial to APDEC. The first is a cost savings, as APDEC will no longer need to expend in-project resources on maintaining visualization software. The second is new AMR visualization capabilities that include the ability to run on parallel machines and support for remote and distributed visualization.

We added a new capability in VisIt – AMR spreadsheet plots – that support direct viewing of numerical values on a particular slice of a patch (see Figure 7). This function is essential for debugging and used by AMR code development teams on a daily basis. The new spreadsheet capability is integrated with VisIt's "pick cell" feature, allowing users to "link" them to other plots. Additional new features include the ability to customize the VisIt interface, thereby improving usability so that new users to can quickly navigate and employ features familiar to them in their older, retiring software.

While not as visible as the features above, other recent accomplishments include software architecture and engineering work to produce all-important performance improvements. Optimizations in AMR grid processing have produced a ten-fold savings in memory and support more efficient rendering. Additional performance and memory optimizations improve efficiency for the important use case of rendering patch boundaries. Our new, specialized algorithm is an order of magnitude faster and more memory efficient than the previous implementation.

All of these software enhancements that produce important performance improvements and visualization capabilities crucial to AMR-based computational science projects have been made available to the public through production-quality, parallel-capable, open source visualization software.

Conclusion

This article has but scratched the surface of a number of serious challenges facing modern scientific researchers. At the root of most of these challenges is the fact that we are awash with information, and that gaining understanding from an increasing amount of data is an incredibly challenging task with few, if any, "off-the-shelf" solutions. This article has provided an overview of the value of visualization in scientific knowledge discovery, as well as a couple of examples of current state-of-the-art.

The mission of DOE's SciDAC Visualization and Analytics Center for Enabling Technologies is to gain traction on solutions to this large family of difficult challenges. We use a multi-faceted approach where state-of-the-art technologies from visualization, data analysis, data management, visual interfaces, software architecture and engineering are brought to bear on some of the world's most challenging scientific data understanding problems.

For more information about VACET, please visit our website at www.vacet.org.

References

1. B. McCormick, T. De Fanti et al., "Special Issue on Visualization in Scientific Computing," Computer Graphics, 21:6, November 1987.
2. C. Garth, F. Gerhardt, X. Tricoche, H. Hagen. "Efficient Computation and Visualization of Coherent Structures in Fluid Flow Applications." *Transactions on Visualization and Computer Graphics/IEEE Visualization 2007* (accepted for publication).
3. U.S. Department of Energy, "Scientific Discovery Through Advanced Computing", http://www.scidac.gov/SciDAC.pdf, March 2000.
4. Richard Mount (ed), The Office of Science Data-Management Challenge – Report from the DOE Office of Science Data-Management Workshops, http://www.slac.stanford.edu/cgi-wrap/getdoc/slac-r-782.pdf, May 2004.
5. VisIt Visualization Software. http://www.llnl.gov/visit, September 2007.
6. G.H. Weber, V. Beckner, H. Childs, T. Ligocki, M. Miller, B. van Straalen, E.W. Bethel. "Visualization Tools for Adaptive Mesh Refinement Data." In: *Proceedings of the 4th High End Visualization Workshop* (Tyrol Austria, June 18-22, 2007), pp. 12-25, 2007.
7. K. Stockinger, J. Shalf, K. Wu and E. W. Bethel. "Query-Driven Visualization of Large Data Sets." In *Proceedings of IEEE Visualization 2005*, pp 167-174, Minneapolis NM, October 2005.
8. O. Rübel, G. H. Weber, S. V. E. Keränen, C. C. Fowlkes, C. L. Hendriks, L. Simirenko, N. Shah, M. Eisen, M. Biggin, H. Hagen, J. Sudar, J. Malik, D. Knowles, and B. Hamann. "PointCloudXplore: Visual Analysis of 3D Gene Expression Data Using Physical Views and Parallel Coordinates. In *Data Visualization 2006, Proceedings of EuroVis 2006*, pp203-210, Eurographics Association, Aire-la-Ville Switzerland, July 2006.
9. R. Bellman. "Adaptive Control Processes: A Guided Tour." Princeton University Press, 1961.
10. K. Wu, E. Otoo, and A. Shoshani. "On the Performance of Bitmap Indices for High Cardinality Attributes." In *International Conference on Very Large Data Bases,"* Toronto, Canada. September 2004.
11. K. Wu, W.-M. Zhang, V. Perevoztchikov, J. Laurent, and A. Shoshani. "The Grid Collector: Using an Event Catalog To Speed Up User Analysis in a Distributed Environment." Computing in High Energy and Nuclear Physics (CHEP), Interlaken, Switzerland, September 2004.
12. K. Stockinger, E. W. Bethel, S. Campbell, E. Dart and K. Wu. "Detecting Distributed Scans Using High Performance Query-Driven Visualization." In *Proceedings of SC06 (Supercomputing)* (need better bibliographic citation data here)
13. C. Bajaj, V. Pascucci, and D. Schikore. "Fast Isocontouring for Improved Interactivity." In *Proceedings of the 1996 Symposium on Volume Visualization*, pp 39-46, October 1996.
14. I. Bowman, J. Shalf, K.-L. Ma and E. W. Bethel. "Performance Modeling for 3D Visualization in a Heterogeneous Computing Environment." Technical Report LBNL-56977. Lawrence Berkeley National Laboratory, Berkeley CA, 2004.
15. A. Gyulassy, V. Natarajan, B. Hamann, and V. Pascucci. "Efficient Computation of Morse-Smale Complexes for Three-Dimensional Scalar Functions." *Transactions on Visualization and Computer Graphics/IEEE Visualization 2007* (accepted for publication).
16. J. Stasko and E. Zhang. "Focus+Context Display and Navigation Techniques for Enhancing Radial, Space-Filling Hierarchy." In *Proceedings of IEEE Information Visualization 2000,* pp57-65, Salt Lake City UT, October 2000.
17. W. Lee, S. Ethier, W. Wang, and S. Klasky. "Gyrokinetic Particle Simulation of Fusion Plasmas: Path to Petascale Computing." Journal of Physics: Conference Series 46(2006), pp 73-81, *Proceedings of SciDAC 2006*. Institute of Physics Publishing, July 2006.
18. H. Varian. "The Information Economy." Scientific American, pp 200-201, September 1995.