FULL TITLE
*ASP Conference Series, Vol. **VOLUME**, **YEAR OF PUBLICATION***
NAMES OF EDITORS

Modern Scientific Visualization is More than Just Pretty Pictures

E. Wes Bethel, Oliver Rübel, Prabhat, Kesheng Wu, and Gunther H. Weber

Lawrence Berkeley National Laboratory, Berkeley CA 94720, USA

Valerio Pascucci

Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, UT

Hank Childs and Ajith Mascarenhas

Lawrence Livermore National Laboratory, Livermore CA, USA

Jeremy Meredith and Sean Ahern

Oak Ridge National Laboratory, Oak Ridge TN, USA

Abstract. While the primary product of scientific visualization is images and movies, its primary objective is really scientific insight. Too often, the focus of visualization research is on the product, not the mission. This paper presents two case studies, both that appear in previous publications, that focus on using visualization technology to produce insight. The first applies "Query-Driven Visualization" concepts to laser wakefield simulation data to help identify and analyze the process of beam formation. The second uses topological analysis to provide a quantitative basis for (i) understanding the mixing process in hydrodynamic simulations, and (ii) performing comparative analysis of data from two different types of simulations that model hydrodynamic instability.

1. Introduction

Galileo Galilei's improvements to early telescope design first opened up the heavens, the satellites of Jupiter, sunspots, and even the rotation of the Sun. He proved the Copernican heliocentric model of the solar system: it is the Sun, not the Earth, that is the center of the solar system. Thus, the telescope became the first device to make the "unseeable" "seeable."

Today, scientific visualization plays an equally significant role in contemporary science. Visualization transforms abstract data into readily comprehensible images (e.g., Figure 1) and has become an indispensable part of the scientific discovery process. In 1987, a landmark report from the computer graphics and visualization community (McCormick 1987) coins the term "visualization in scientific computing" and amplifies on the subject of its central role in the scientific discovery process.

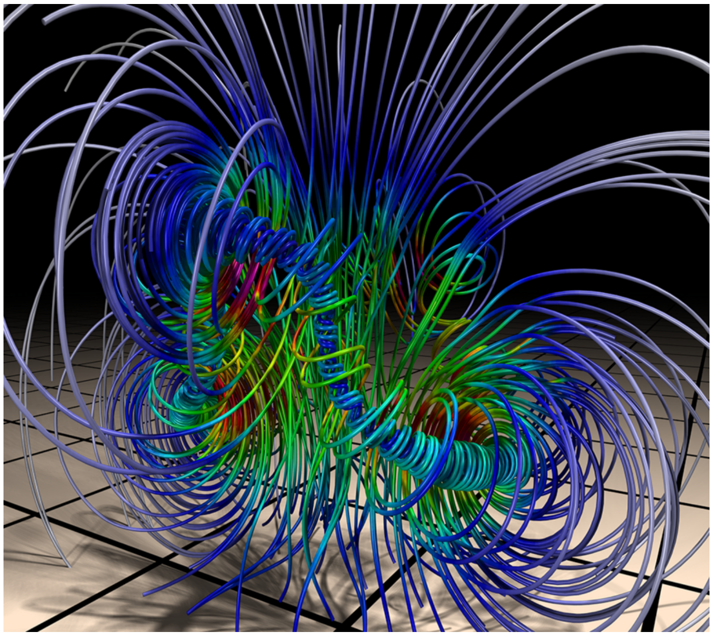

Figure 1. This image shows the merging of two vortex cores, something not possible to see by examining tables of numbers. Image courtesy Dave Pugmire (ORNL), sample AMR data courtesy Phil Colella (LBNL) and the SciDAC Applied Partial Differential Equations Center.

In (Butler 1993), Bergeron describes three broad user-centric visualization use modalities. “Presentation visualization” is where you know what is there and want to show it to someone else. “Analytical visualization” is where you know what you are looking for. “Discovery visualization” occurs when you have no idea what you’re looking for. Discovery visualization is characterized as an “undirected search,” or “unconstrained navigation” through visualization or rendering parameter space.

Each of these use modalities plays a key role at different stages in the scientific process, and each has a unique set of architectural and resource requirements. Discovery visualization is typically employed early in the investigatory process to uncover trends and anomalies in scientific data. It requires unconstrained, interactive navigation through an $n-$dimensional space (data variables, time, and logical axes of structured grids), which in turn requires a potentially great deal of I/O bandwidth and computational power. Exploration visualization is most often conducted in an interactive fashion. Analytical visualization serves to confirm, or not, the presence of features or characteristics in data. It is often implemented as a non-interactive “standalone, post-process,” and as such most often does not have the same I/O nor computational demands as exploration visualization due to the relaxation of interactivity and random-access use patterns. Presentation visualization, like analytical visualization, serves to communicate known characteristics about data in a particular way. It, too, is most typically implemented as a non-interactive post-processing function.

One of the major challenges faced by modern science is the explosion of information. It is widely accepted the most significant bottlenecks in modern science are the management of an ever-increasing amount of information and deriving knowledge or insight from the tsunami of data (Mount 2004). The visualization community has responded to this challenge in a number of different ways. One is to use parallelism to scale existing algorithms to accommodate larger datasets (Childs 2005; Law 2001) or to develop novel implementations of existing algorithms suitable for deployment on parallel, distributed platforms (Bethel 2000). These approaches have the side-effect of increasing the amount of information a human must interpret; they make the "information overload" problem worse, rather than better.

Another set of approaches, which we examine here, focus on reducing the amount of information a human must interpret even though the datasets being examined are larger and more complex. Section 2. presents the topic of "Query-Driven Visualization," where analysis, visualization, and human interpretation is restricted to data deemed to be "scientifically interesting." The idea is to extract meaningful subsets of data, e.g., "interesting data," and then subject those subsets to analysis, visualization, and interpretation. "Interesting" is defined as data that satisfy a multi-dimensional Boolean range query. This approach can be used successfully for all three major visualization use modalities. Another variation on this theme is the subject of Section 3., which focuses on extracting features in data using topological analysis. This approach has the benefit of providing a quantitative basis for comparative visualization and analysis.

2. Query-Driven Visualization (QDV)

2.1. What is QDV?

The term "Query-Driven Visualization" refers to the process of limiting visualization processing and subsequent visual interpretation to data that is deemed to be "interesting." The basic premise is to restrict computational and cognitive load by limiting processing and interpretation to features of interest. This approach is consistent with the needs of many scientific users who need capabilities to help them find and focus on features hidden in large, multidimensional data (Hamann 2002).

One of the significant challenges from the field of data management is data searching, which, incidentally, is the key architectural linchpin for an efficient and high-performance QDV implementation. Many approaches have been used over the years, ranging from well-known constructs like B-trees Knuth (1998) to complex indexing schemes. It is well known that tree-based indexing methods suffer from the so-called "Curse of Dimensionality," which dictates that storage requirements grow exponentially with increasing dimensionality (Bellman 1961). In contrast, storage requirements for bitmap indices grow linearly with increasing dimensionality; bitmap indices are really the only viable approach for index/query for large, multidimensional scientific data. To efficiently answer complex multidimensional, multivariate data queries, we turn to the scientific data management community to leverage a form of compressed bitmap indexing that has proven generally applicable to a broad range of applications, including data analysis from High Energy Physics Experiments (Wu 2004).

2.2. Case Study: Laser-Wakefield Simulation Data Analysis

We now present a specific example where we apply our system to perform visual data analysis of 2D and 3D data produced by a laser wakefield particle accelerator simulation. In these examples, both 2D and 3D datasets contain information about each particle's position, momentum, and a unique particle identifier. The 2D dataset contains about 400K particles in each of the 18 timesteps, and is about 1.3GB in size, including the index structure. The 3D dataset contains about 90M particles in each of the 30 timesteps, and is about 210GB in size, including the index. These simulations model the effects of a laser pulse propagating through a hydrogen plasma. Similar to the wake of a boat, the radiation pressure of the laser pulse displaces the electrons in the plasma, and with the space-charge restoring force of the ions, this displacement drives a wave (wake) in the plasma. Electrons can be trapped and accelerated by the longitudinal field of the wake, forming electron bunches of high energy.

In order to gain a deeper understanding of the acceleration process, we need to address complex questions such as: (i) which particles become accelerated; (ii) how are particles accelerated, and (iii) how was the beam of highly accelerated particles formed and how did it evolve (Geddes 2005). To identify those particles that were accelerated, we first perform selection of particles at a late timestep ($t = 37$) of the simulation by using a threshold for the value for x-momentum, px. By tracing the selected particles over time we will then analyze the behavior of the beam during late timesteps and identify characteristic beam-substructures during beam formation. In this context, we use selection of particles, ID-based tracing of particles over time, and refinement of particle selections based on information from different timesteps as main analysis techniques.

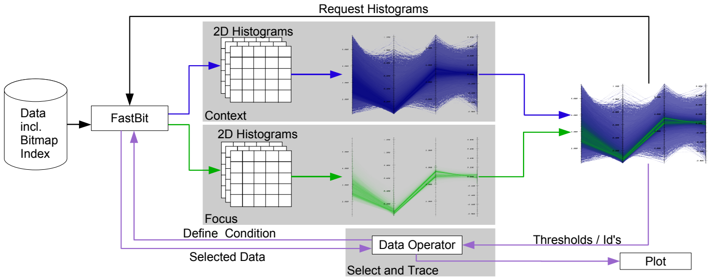

Figure 2. An overview of the major components and data flow paths in our implementation. Large-scale scientific data and indexing metadata are input via a parallel I/O layer, which allows us to achieve high levels of performance and parallel efficiency. We use an API at the I/O layer to perform parallel computation of multidimensional histograms as well as data subsetting. Results of those computations are used downstream in the visualization application for presenting information to the user and in support of interactive data mining actions.

System Overview The fundamental concepts in our implementation center around effective I/O, effective presentation of information, and an effective mechanism that allows users to easily and quickly specify queries. From a high level, the system allows a user to first see an overview of "all the data" using a multidimensional visual information display technique known as "parallel coordinates." The user then specifies the subset of data that is "interesting" by using the parallel coordinates display to define a multivariate range query. This query in turn is given to an I/O layer, which either extracts the data subset or computes new, more finely resolved histograms. Data subsets are sent to downstream processing and analysis modules, while new histograms are in turn used as the basis for constructing and displaying new, refined parallel coordinates plots. In our implementation, the computational complexity of rendering parallel coordinates plots – both context and focus views – is a function of histogram resolution, not the size of the underlying data. Therefore, our approach is particularly well-suited for application to extremely large data.

An overview of the system architecture is shown in Figure 2. Raw scientific data, which is produced by simulation or experiment, is augmented by the computation of indexing data. In our case, this step is performed outside the visual data analysis application as a one-time preprocessing, and our implementation uses FastBit (LBNL-SDM 2008) for creating index structures.

Beam Selection In order to identify the beam, i.e., find those particles that become accelerated, we first concentrate on the last timestep of the simulation (at $t = 37$). Using the parallel coordinates display, we select the particles of interest by applying a threshold of $px > 8.872 * 10^{10}$. As is visible in the parallel coordinates plot, those particles constitute two separate clusters (beams) in x direction (Figure 3 c). Using a pseudocolor plot of the data, we can then see the physical structure of the beam.

Beam Assessment Having identified the particles that are part of the beam, we wish to (i) assess the quality of the beam, and (ii) analyze its formation. We address these questions by enabling the user to efficiently trace selected particles forwards and backwards in time. By tracing particles back in time, we observe that the first bunch following the laser pulse (rightmost in these plots) has lower momentum spread at its peak energy (at $t = 27$) than the second bunch (see Figure 3a and 3b). In practice, the first beam following the laser pulse is therefore typically the one of most interest to the accelerator scientists. The fact that the second beam shows higher or equal acceleration at the last timestep of the simulation is due to the fact that the first beam will outrun the wave later in time and therefore switches into a phase of deceleration while the second beam is still in an acceleration phase. In practice, when researchers want to select only the first beam, they usually perform selection of the particles using thresholding in px at an earlier time (e.g. $t = 27$) when the beam-particles of interest have maximum momentum, rather than the last timestep, here $t = 37$. By performing selection at an earlier time, one avoids selecting particles in the second beam while being sure to select all particles in the first beam. In this specific use case we are interested in analyzing and comparing the evolution of these two beams, which is why we performed selection at the last timestep.

Beam Refinement To analyze formation and evolution of the beam, we trace the selected particles further back in time to the point when the particles entered the simulation and were injected into the beams. Based on the information at timestep $t = 14$, we then refine our initial selection of the beam. By applying

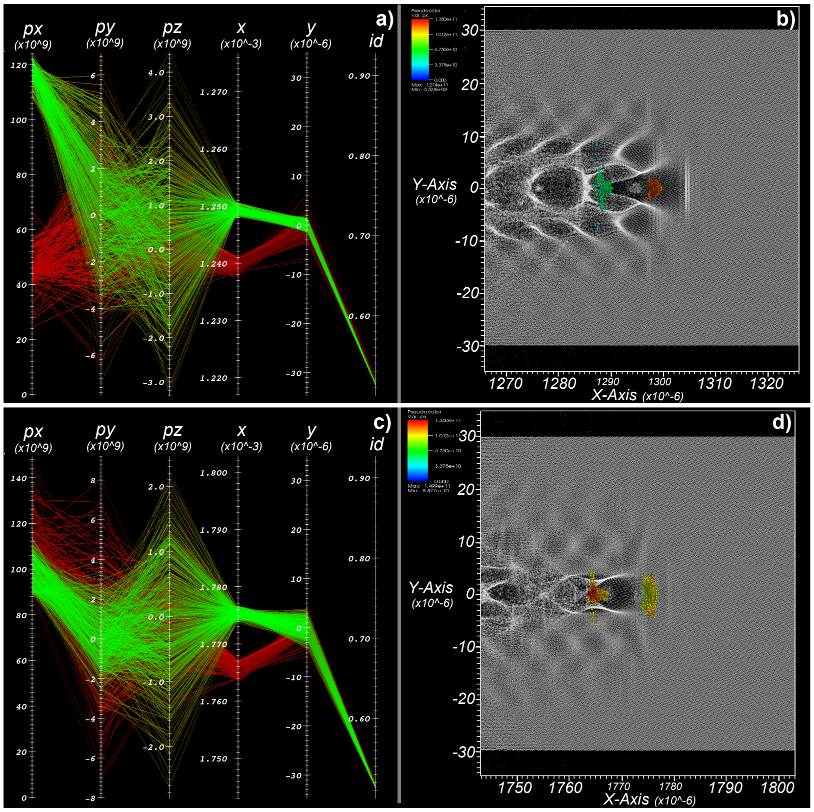

Figure 3. a) Parallel coordinates and b) pseudocolor plot of the beam at $t = 27$. Corresponding plots c,d) at $t = 37$. The context plot, shown in red, shows both beams selected by the user after applying a threshold of $px > 8.872 * 10^{10}$ at $t = 37$. The focus plot, shown in green, indicates the first beam that is following the laser pulse. In the pseudocolor plots b) and d), we show all particles in gray and the selected beams using spheres colored according to the particle's x-momentum, px. The focus beam is the rightmost bunch in these images. At timestep $t = 27$, the particles of the first beam (green in figure a) show much higher acceleration and a much lower energy spread (indicated via px) than the particles of the second beam. At later times, the lower momentum of the first beam indicates it has outrun the wave and moved into decelerating phase, e.g at timestep $t = 37$.

an additional threshold in x, we can select those particles of the beam that are injected into the first wake period behind the laser pulse (see Figure 4a and 4b). By comparing the temporal traces of the selected particle subset (green) with the traces of the whole beam (red), we can readily identify important characteristics of the beam (see Figure 4c). After being injected, the selected particle subset (green) first defines the outer part of the first beam at timestep $t = 15$, while additional particles are injected into the center of the beam. Later on at timesteps $t = 16$ and $t = 17$, the selected particles become strongly focused and define the center of the first beam. By refining selections based on information at an earlier time, we are able to identify characteristic substructures of the beam.

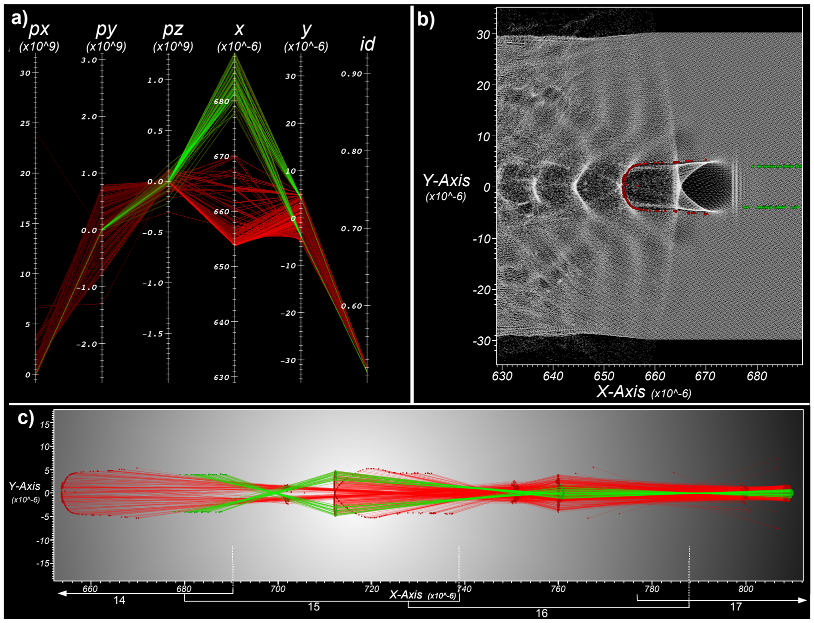

Figure 4. a) By applying an additional threshold in x at timestep $t = 14$, we separate the two different set of particles entering the simulation. b) The refinement result, shown in physical space, includes all non-selected particles (gray) to provide context. c) Particle traces of the complete beam and the refined selection. In all plots we show the complete beam in red and the refined selection in green. After entering the simulation, the selected particles (green) define first the outer part of the first beam at timestep $t = 15$. Later on at timesteps $t = 16$ and $t = 17$, these particles become highly focused and define the center of the first beam.

3D Analysis Example We now describe an example analysis of the 3D particle dataset. Figure 5(a and b) shows the beam selection step for this dataset. At a much earlier timestep $t = 12$ ($x \approx 5.7 * 10^{-4}$ compared to $x \approx 1.3 * 10^{-3}$ in the 2D case) particles are trapped and accelerated. In order to get an overview of the main relevant data, the user removed the background from the data first by applying a threshold of $px > 2.0 * 10^9$. The user then selected particles in the first bunch via thresholding based on the momentum in x direction ($px > 4.586 * 10^{10}$) and x position ($x > 5.649 * 10^{-4}$) to exclude particles in the secondary periods from the selection. Figure 5b shows a volume rendering of the plasma density along with the selected particles revealing the physical location of the selected beam within the wake. Figure 5c shows the traces of the particles selected earlier in Figure 5a. We selected the particles at timestep $t = 12$ then traced them back to timestep $t = 9$ where most of the selected particles enter the simulation window and forward in time to timestep $t = 14$. As one can see in the plot, the selected particles are constantly accelerated over time.

Performance Prior to our collaboration with the accelerator scientist, his usual method for performing the type of analysis we describe in earlier sections was entirely manual, and performed in a serial fashion. He would examine phase-space plots at different timesteps to visually identify particles having both high

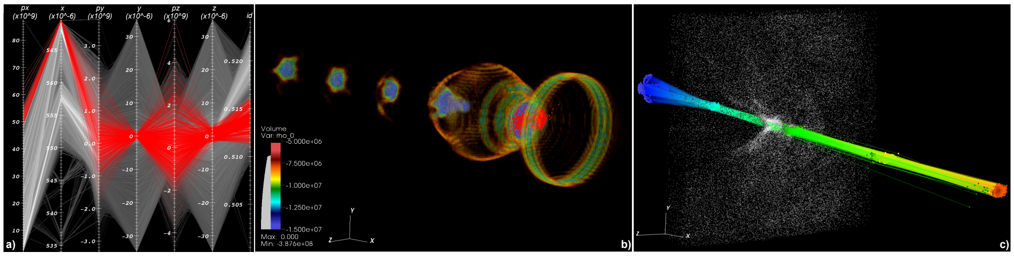

Figure 5. a) Parallel coordinates of timestep $t = 12$ of the 3D dataset. Context view (gray) shows particles selected with $px > 2 * 10^9$. The focus view (red) shows particles satisfying the condition $(px > 4.856 * 10^{10})$ AND $(x > 5.649 * 10^{-4})$, which form a compact beam in the first wake period following the laser pulse. b) Volume rendering of the plasma density and the selected focus particles (red). c) Traces of the beam. We selected particles at timestep $t = 12$, then traced the particles back in time to timestep $t = 9$ when most of the selected particles entered the simulation window. We also traced the particles forward in time to timestep $t = 14$. Color indicates px. In addition to the traces and the position of the particles, we also show the context particles at timestep $t = 12$ in gray to illustrate where the original selection was performed. We can see that the selected particles are constantly accelerated over time (increase in px).

momentum and a high degree of spatial coherence. Having identified a set of candidate particles, he would generate a list of particle IDs, then run an IDL script (in serial) to extract target particles from all simulation timesteps, then perform further analysis. The process of extracting particles from all timesteps would require hours of runtime due to a combination of the serial nature of processing and the somewhat inefficient (but easy to implement) search algorithm.

We conducted an extensive performance analysis of our system that focuses on the following factors: (1) scalability and parallel performance of the algorithm for computing 2D conditional histograms, which form the basis for visual information display in parallel, from multidimensional time-varying datasets, (2) the scalability and parallel performance of an algorithm that performs particle extraction from a multi-terabyte sized dataset. The results of that study, which are presented in detail in (Rübel 2008), show that our techniques exhibit excellent scalability, parallel performance, and absolute wallclock runtime. We were able to sucessfully replace a manual process that required hours of runtime with one that executes in parallel on a Cray XT4 system in only a few seconds. This type of performance helps to improve scientific productivity during the process of exploratory visual data analysis.

3. Robust Topological Techniques for Large Scale Data Analysis

A deeper understanding of simulated phenomena requires powerful data analysis tools that go beyond generating images for display. Often, this analysis takes the form of defining, detecting, and quantifying features in data, features that can and do exist at multiple scales in data. The definition of a feature is application-dependent and often difficult to capture with mathematical rigor. In the analysis of data produced by hydrodynamic instability simulations, a "bubble" and "bubble counts" are examples of features that are easy to spot visually but difficult to define with rigor. They are easy to spot visually in early stages of the simulation and appear as uniform, similarly-sized bumps in the interface surface (see Figure 7). In later stages, when turbulence dominates the mixing process, a bubble can show multi-scale properties, where a large bubble consists of many smaller bubbles.

Recent developments in the field of computational topology provide the mathematical foundations and algorithms to tackle problems in this area. Because the mathematics is developed in an abstract setting, it can be adapted to a specific problem, such as the analysis of bubble structures in the hydrodynamic instability simulation. Because the algorithms are strictly combinatorial, they are robust and can provide rigorous guarantees about the computation results. Such guarantees are crucial features that are missing in many numerical algorithms that compute topological features. Our analysis algorithms are scalable – we have applied them to datasets having up to 3072^3 grid resolution. As described in Section 3.2., we have for the first time provided a quantitative comparison and validation of two different classes of simulations of the same physical phenomenon.

The contents of this section are a summary of several publications and prior research. For a more detailed description on the data layout scheme see (Pascucci 2001), for topology-based analysis (Cole 2003; Edelsbrunner 2003; Bremer 2003, 2004; Natarajan 2006), and for application of these techniques to the analysis of simulation data, including hydrodynamic instability, see (Laney 2006; Gyulassy 2007; Miller 2006).

3.1. Morse Theory for Robust, Multi-scale Feature Analysis

To begin, we present some background from Morse theory (Milnor 1963) and from combinatorial and algebraic topology (Alexandrov 1998; Munkres 1984).

Smooth maps on manifolds. Let $f\colon \mathbb{M} \to \mathbb{R}$ be a smooth map. A point $x \in \mathbb{M}$ is a *critical point* of f if the gradient of f vanishes at x, and the value $f(x)$ is a *critical value*. Non-critical points and non-critical values are called *regular points* and *regular values*, respectively. A critical point x is *non-degenerate* if the Hessian (matrix of second-order partial derivatives) at x is non-singular. The *index* of a critical point x is the number of negative eigenvalues of the Hessian. For $d = 3$, there are four types of non-degenerate critical points: the *minima* (index 0), the 1-*saddles* (index 1), the 2-*saddles* (index 2), and the *maxima* (index 3). A function f is *Morse* if all critical points are non-degenerate with distinct values.

Morse-Smale complex. An *integral line* is a maximal path on $\mathbb{M}$ whose tangent vectors are parallel to the gradient of f. The *stable manifold* of a critical point x is the union of x and all integral lines that *end* at x. The *unstable manifold* of x is defined symmetrically as the union of a critical point x and all integral lines that *start* at x. One can superimpose the stable and unstable manifolds of all critical points to create the *Morse-Smale complex* (or MS complex) of f (Edelsbrunner 2003; Bremer 2004), see Figure 6(a)–(d). The *nodes* of this complex are the critical points of f, its *arcs* are integral lines

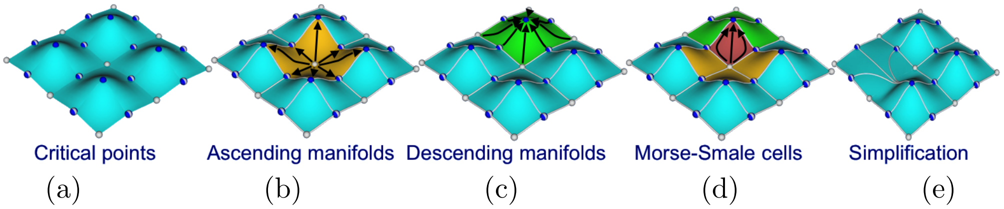

Figure 6. MS-complex construction, simplification and topologically valid approximation: (a) Morse function with critical points shown; (b) Stable manifolds; (c) Unstable manifolds; (d) MS-complex; (e) MS-complex and manifold after simplification. Maxima are solid blue, minima are solid white, and saddles are mixed.

starting or ending at saddles and its *regions* are the non-empty intersections of stable and unstable 2-manifolds. More details on the definition of the MS-complex on 2-manifold triangle meshes and algorithms to compute it are given by Bremer et al. (Bremer 2004).

Simplification. It is often useful to simplify a MS-complex to remove noise and to perform multi-scale function analysis. Following (Bremer 2004), we perform *cancellations* of arc-connected maximum-saddle and minimum-saddle critical point pairs to simplify an MS-complex. Cancellations are ranked by their *persistence* – the absolute difference in function value between the canceled critical point pair. Figure 6(e) shows an example of a topological simplification and a corresponding approximation of f.

Computation. In practice, one usually deals with piecewise linear (PL) functions given at the vertices of a triangulation. See (Bremer 2004) for a detailed discussion on how to translate concepts from the generic smooth functions discussed above to PL-functions. Starting from saddles, we compute the arcs of the MS-complex as the steepest, non-crossing monotone lines. Because we avoid mesh refinement to handle degeneracies, and instead directly handle merged lines and multi-saddles, we can use efficient static data structures to store the triangulation and can compute MS-complexes of large data sets. We are now ready to describe how we use the MS-complex to analyze the Rayleigh-Taylor simulations developed at Lawrence Livermore National Laboratory (Cook 2001). This analysis is described in detail in (Laney 2006).

3.2. Analysis of Turbulent Mixing in Hydrodynamic Instabilities

Understanding the turbulent mixing of fluids is one of the fundamental research problems in the area of fluid dynamics. Turbulent mixing occurs in a broad spectrum of phenomena ranging from boiling water to astrophysics and nuclear fusion. Rayleigh-Taylor instability (RTI) occurs when two fluids of different density are accelerated opposite the mean density gradient. That is, a heavier fluid is accelerated against a lighter fluid by a force like gravity. Figure 7 shows the mixing layers and the progression of the mixing process. The heavy fluid accelerates downward, forming "spikes," while the light fluid moves upward forming "bubbles." The bubbles and spikes are thought to be one way to

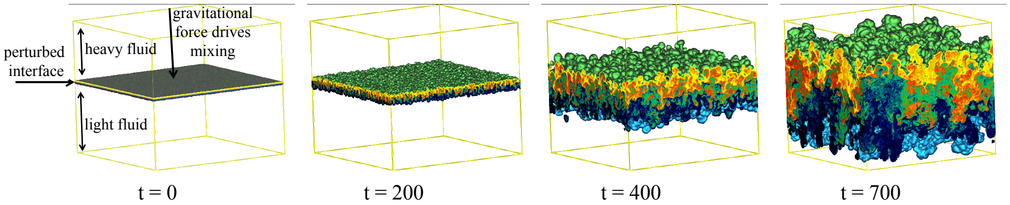

Figure 7. An overview of the 1152^3 simulation (periodic in x and y) of the Rayleigh-Taylor instability at start ($t = 0$), early ($t = 200$), middle ($t = 400$), and late ($t = 700$) time. The light fluid has a density of 1.0, the heavy fluid has a density of 3.0. Two envelope surfaces (at densities 1.02 and 2.98) capture the mixing region. The boundaries of the box show the density field in pseudocolor. We analyze the upper envelope to study bubble structures and the midplanes to study mixing trends.

characterize the large-scale behavior of the mixing process. Scientists analyzing these simulations are particularly interested in the number of bubbles (and spikes) and their respective evolution. Large-scale models have been proposed based on bubble dynamics in which bubble growth, movement, and interaction are modeled (Alon 1996). Our analysis is performed on the envelope surfaces describing the boundary between undisturbed and “mixed” fluids.

We use multiresolution data layout and access techniques described in (Pascucci 2001) for efficient data management and the techniques described in Section 3.1. for the analysis.

Segmentation of bubbles. One of the challenges in analyzing mixing behavior is that there exists no prevalent mathematical definition for what constitutes a bubble/spike. In general, a bubble can be understood as a three dimensional feature composed of lighter density fluid moving upwards (in the Z-direction) into a heavier density fluid. We use the topological concepts introduced in Section 3.1. to define bubbles, spikes and other features of interest. Consider the images of the segmented mixing envelope surface at different times shown to the left, bottom, and right of the plot in Figure 8. During early time steps (Figure 8 upper left and middle left), it is natural to consider the mixing envelope as a time-varying *functional* surface defined over the XY-plane (i.e, a function with a single value per XY coordinate such as in a terrain) and to associate local maxima with bubbles. This analogy fails at later time steps because the surface becomes non-functional.

However, we can generalize this approach by treating the envelope surface as the domain of a function whose value at a point x is the Z-coordinate of x. It is natural to connect the maxima of this function to bubbles and compute the stable manifold of each maximum as a segmentation of the surface into bubbles. As can be seen in Figure 8, this segmentation corresponds very well to the human notion of a bubble. Symmetrically, we use the unstable manifolds of minima to define spikes. Potentially, other functions could be defined on the envelope surface that would result in a robust segmentation. For example, the Z-velocity at all of the points on the envelope surface could be incorporated to capture the fact that bubbles should be moving upwards into the heavy fluid.

In general, topologically based segmentations are often linked to important features: maximal and minimal Z-velocities on the midplanes correspond to cores of rising and falling sections of fluids; density extrema correspond to pockets of unmixed fluids. Topological methods are flexible and enable analysis of a variety of phenomena using the same methodology. Furthermore, the MS-complex can be computed combinatorially (Edelsbrunner 2003; Bremer 2004), which translates into provably correct and stable algorithms that are crucial when dealing with large and complex data.

Multi-Scale analysis and persistence selection. The MS-complex, just as any other segmentation, captures noise as well as features. We use a simplification scheme to remove noise and to construct a series of approximations at decreasing resolution. Unlike many other techniques, topological segmentations allow a simplification scheme that is optimal in the L_∞-norm. One can formulate the problem of coarsening a segmentation in the following manner: given a function f and a segmentation S of the domain of f, what is the minimal change on f such that S is coarsened? If the segmentation one considers is the MS-complex of f, then it can be shown (Bremer 2004) that canceling a critical point pair with persistence p in f requires an approximation $\hat{f}$ with $||f - \hat{f}||_\infty \geq p/2$. Therefore, canceling critical points in order of increasing persistence corresponds to an L_∞-optimal simplification.

Cancellations of critical points can be used both to remove noise and, more in general, to analyze the trends in the data at multiple scales. This type of multi-scale analysis is not readily available in classical image processing techniques and constitutes a fundamental advantage of our approach.

For each MS-complex, we compute a sequence of cancellations that optimally simplify the complex down to its minimal configuration. We can thus define a *family of segmentations* of the envelope surface ranging from persistence $p = 0.0$, where both signal and noise features are segmented, to persistence $p = 1.0$ (full function range), where the entire surface is collapsed into a single topological feature. We can then create statistics showing the number of bubbles over time using a range of persistence thresholds. These results show that the mixing behavior can differ significantly across scales. Using the simplification sequences, we can capture the behavior on all scales without recomputing the MS-complex. Domain scientists interact with a visualization of the segmented surface and select an appropriate persistence value based on their physical intuition of a correct segmentation of bubbles.

Results: Bubble Counts and Quantification of Mixing Phases The input data consists of 758 time-steps, each time-step containing about 5.8GB of density data defined on a 1152^3 grid. Our analysis was performed on 68 dual-processor nodes of a cluster running Linux. For each time-step, we extract an isosurface at density value 2.98 and compute the Morse-Smale complex with function f set to the z-coordinate of each point. Isosurface extraction takes about 15 seconds at early time-steps and about 30 seconds at the late time-steps where the surface is more complex. Morse-Smale complex computation takes about 2 seconds at early time-steps and about 25 seconds at late time-steps. We compute a digest

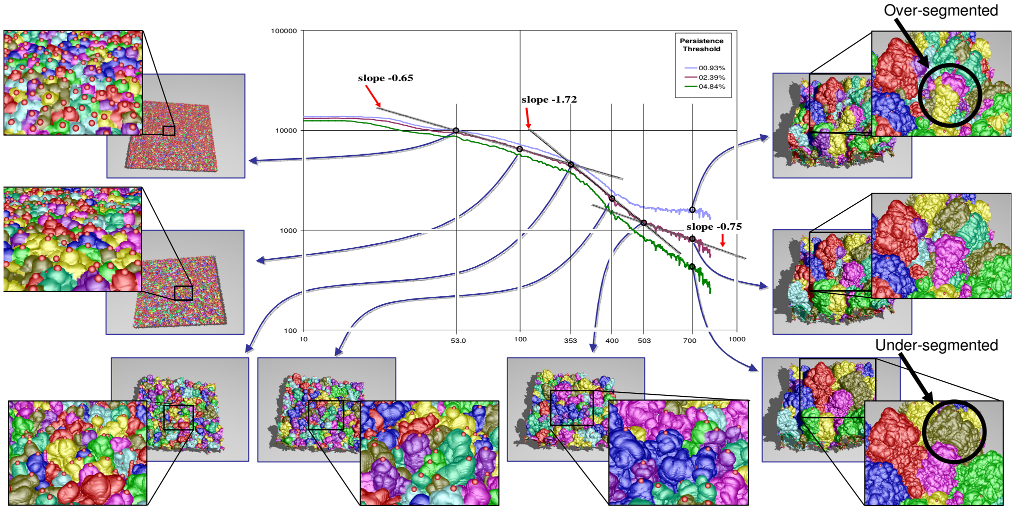

Figure 8. The plot depicts bubble counts of the envelope surface for three persistence values. To the right of the plot, the MS-segmentation at three persistence values for time 700. To the left and below the plot, the bubble segmentation along the curve of medium persistence at various times. Each maximum along with the Morse cells of its child-maxima are colored the same.

of all required data from the Morse-Smale complex that allows us to efficiently calculate bubble counts for any choice of persistence value.

Figure 8 shows bubble counts for three choices of persistence values; the inset figures show the segmentation of the isosurface into bubbles. Since it was not clear what persistence value is the "best" choice, our flexible analysis pipeline supports computation of the MS-complex and resultant bubble count using a number of different persistence values. We provided scientists with a variety of visualizations and statistics that they can study to choose an appropriate persistence that produces results that make sense to them. It is important to note that the points of inflection, where the slope of each curve changes, occur at the same time step for all three scales. This fact indicates that the key results of our analysis are fundamentally insensitive to the choice of persistence value: the measurements of interest for these data are primarily focused on the trends of the derivative rather than the absolute value of the bubble count.

After an extensive expert scientific evaluation of results computed from different persistence values, we settled on a bubble count that corresponds to a persistence value of 2.39% of the function range at the end of the simulation. Figure 9, shows the result of this analysis summarized by the final bubble count curve and its derivative. The plot reveals four phases of the mixing process, providing for the first time quantification of the mixing rates and transition times between the different stages.

Feature-based Comparison of Large Eddy Simulations(LES) and Direct Numerical Simulations(DNS) Our data management and analysis techniques provide a method to compare two different data sets of the same phenomenon simu-

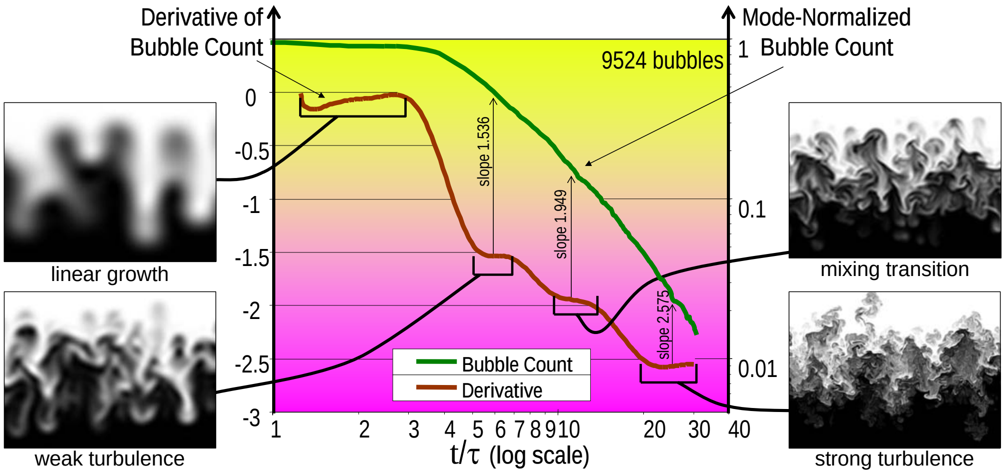

Figure 9. Bubble count and its derivative. The derivative curve shows four regions where the slope is virtually constant, each of which corresponds to distinct stages of the turbulent mixing process.

lated using different classes of simulations. Large Eddy Simulations (LES) use smaller resolution meshes and less computational resources than Direct Numerical Simulations (DNS). Grid resolutions for LES capture large-scale effects, but must model the sub-grid scale effects explicitly running, which can lead to potential modeling errors. On the other hand, DNS simulations are based on first-principles of continuum mechanics and therefore not prone to the same modeling errors. However, DNS has high computational requirements, since it requires grids with resolutions that must be fine enough to capture physical effects at all scales.

Validating the results of LES simulations against a corresponding DNS applied to the same phenomenon is of great importance to scientists. We have conducted a feature-based comparison that is shown in Figure 10. The LES grid dimension was 1152^3 and the DNS grid dimension was 3072^3. We perform the analysis as described previously and compute bubble counts for both types of simulations. After normalizing the counts, we can plot them together and compare them. Figure 10 shows that the simulations are in close agreement and produce similar trends in the mixing layer. This result shows that topologically based analysis provides the basis for performing comparative analysis of data from different simulations run at different resolutions.

4. Conclusion

While the main product of scientific visualization is visual – images and movies – its main objective is scientific insight. Simply increasing visualization capacity to respond to ever-increasing data size and complexity does not necessarily result in a corresponding increase in insight. We have presented two case studies, one implementing a high performance exploratory visualization use model, the other implementing an advanced analytical visualization use model, that emphasize scientific insight rather than pretty pictures.

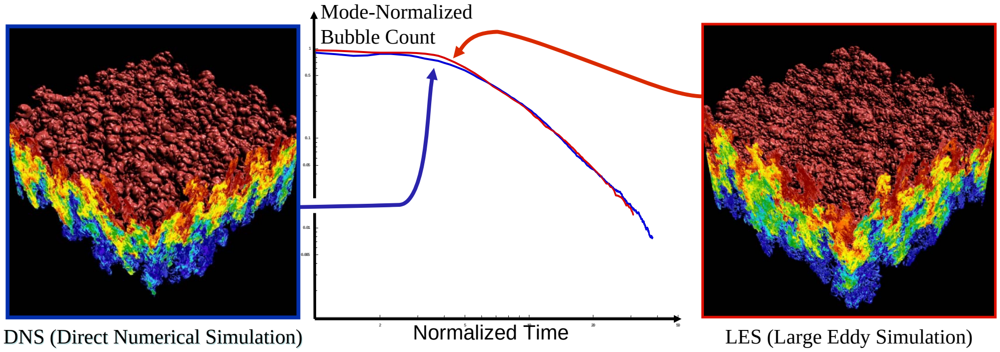

Figure 10. Feature-based comparison and validation of Large Eddy Simulation(LES) and Direct Numerical Simulation(DNS) of Rayleigh-Taylor Instabilities.

Acknowledgments. This work was supported by the Director, Office of Science, Offices of High Energy Physics and Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 through the Scientific Discovery through Advanced Computing (SciDAC) program's Visualization and Analytics Center for Enabling Technologies (VACET) and the COMPASS project. This research used resources of the National Energy Research Scientific Computing Center (NERSC), which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

P. S. Alexandrov. *Combinatorial Topology.* Dover, Mineola, NY, USA, 1998.

U. Alon and D. Shvarts. Two-Phase Flow Model for Rayleigh-Taylor and Richtmeyer-Meshkov Mixing. In *Proc. Fifth International Workshop on Compressible Turbulent Mixing.* World Scientific, 1996.

R. Bellman. *Adaptive Control Processes: A Guided Tour.* Princeton University Press, 1961.

Wes Bethel, Brian Tierney, Jason Lee, Dan Gunter, and Stephen Lau. Using High-Speed WANs and Network Data Caches to Enable Remote and Distributed Visualization. In *Supercomputing '00: Proceedings of the 2000 ACM/IEEE Conference on Supercomputing*, Washington, DC, USA, 2000. IEEE Computer Society.

P.-T. Bremer, H. Edelsbrunner, B. Hamann, and V. Pascucci. A Multi-resolution Data Structure for Two-dimensional Morse-Smale Functions. In G. Turk, J. J. van Wijk, and R. Moorhead, editors, *Proc. IEEE Visualization '03*, pages 139–146, Los Alamitos California, 2003. IEEE, IEEE Computer Society Press.

P.-T. Bremer, H. Edelsbrunner, B. Hamann, and V. Pascucci. A Topological Hierarchy for Functions on Triangulated Surfaces. *IEEE Trans. on Visualization and Computer Graphics*, 10(4):385–396, 2004.

David M. Butler, James C. Almond, R. Daniel Bergeron, Ken W. Brodlie, and Robert B. Haber. Visualization Reference Models. In *VIS '93: Proceedings of the 4th conference on Visualization '93*, pages 337–342, 1993.

Hank Childs, Eric Brugger, Kathleen Bonnell, Jeremy Meredith, Mark Miller, Brad Whitlock, and Nelson Max. A Contract Based System For Large Data Visualization. *Visualization Conference, IEEE*, pages 191–198, 2005.

K. Cole-McLaughlin, H. Edelsbrunner, J. Harer, V. Natarajan, and V. Pascucci. Loops in Reeb Graphs of 2-Manifolds. In *Proceedings of the 19th Annual Symposium on Computational Geometry*, pages 344–350. ACM Press, 2003.

A.W. Cook and P. E. Dimotakis. Transition Stages of Rayleigh-Taylor Instability Between Miscible Fluids. *J. Fluid Mech.*, 443:66–99, 2001.

H. Edelsbrunner, J. Harer, and A. Zomorodian. Hierarchical Morse-Smale Complexes for Piecewise Linear 2-Manifolds. *Discrete Comput. Geom.*, 30:87–107, 2003.

Cameron Guy Robinson Geddes. *Plasma Channel Guided Laser Wakefield Accelerator.* PhD thesis, University of California, Berkeley, 2005.

LBNL Scientific Data Management Research Group. FastBit: An Efficient Compressed Bitmap Index Technology, 2008. `http://sdm.lbl.gov/fastbit/`.

A. Gyulassy, M. Duchaineau, V. Natarajan, V. Pascucci, E. Bringa, A. Higginbotham, and B. Hamann. Topologically Clean Distance Fields. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1432–1439, November/December 2007.

Bernd Hamann, E. Wes Bethel, Horst Simon, and Juan Meza. The NERSC Visualization Greenbook: Future Visualization Needs of the DOE Computational Science Community Hosted at NERSC. *The International Journal of High Performance Computing Applications*, 17(2):97–124, 2002.

Donald. E. Knuth. The Art of Computer Programming Volume 3 (2nd ed.), Sorting and Searching Addison-Wesley Professional, 1998.

D. Laney, P. T. Bremer, A. Mascarenhas, P. Miller, and V. Pascucci. Understanding the Structure of the Turbulent Mixing Layer in Hydrodynamic Instabilities. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):1053–1060, 2006.

Charles Law, Amy Henderson, and James Ahrens. An Application Architecture for Large Data Visualization: A Case Study. In *Proceedings of the 2001 IEEE Symposium on Parallel and Large-Data Visualization and Graphics (PVG '01)*, pages 125–128. IEEE Press, 2001.

B. H. McCormick, T. A. DeFanti, and M. D. Brown (eds.). Visualization in Scientific Computing. *Computer Graphics*, 21(6), 1987

P. Miller, P.-T. Bremer, W. Cabot, A. Cook, D. Laney, A. Mascarenhas, and V. Pascucci. Application of Morse Theory to Analysis of Rayleigh-Taylor Topology. In *Proocedings of the 10th International Workshop on the Physics of Compressible Turbulent Mixing*, 2006.

J. W. Milnor. *Morse Theory.* Princeton Univ. Press, New Jersey, USA, 1963.

Richard Mount (ed.). The Office of Science Data-Management Challenge. Report from the DOE Office of Science Data-Management Workshops. Technical Report SLAC-R-782, Stanford Linear Accelerator Center, March-May 2004.

J. R. Munkres. *Elements of Algebraic Topology.* Addison-Wesley, Redwood City, CA, USA, 1984.

V. Natarajan, Y. Wang, P.-T. Bremer, V. Pascucci, and B. Hamann. Segmenting Molecular Surfaces. *Comput. Aided Geom. Des.*, 23(6):495–509, 2006.

Valerio Pascucci and Randall J. Frank. Global Static Indexing for Real-time Exploration of Very Large Regular Grids. In *Supercomputing '01: Proceedings of the 2001 ACM/IEEE conference on Supercomputing*, New York, NY, USA, 2001.

Oliver Rübel, Prabhat, Kesheng Wu, Hank Childs, Jeremy Meredith, Cameron G. R. Geddes, Estelle Cormier-Michel, Sean Ahern, Gunther H. Weber, Peter Messmer, Hans Hagen, Bernd Hamann, and E. Wes Bethel. High Performance Multivariate Visual Data Exploration for Extemely Large Data. In *SuperComputing 2008 (SC08)*, Austin, Texas, USA, November 2008. LBNL-716E.

K. Wu, W.-M. Zhang, V. Perevoztchikov, J. Lauret, and A. Shoshani. The Grid Collector: Using an Event Catalog to Speedup User Analysis in Distributed Environment. In *Computing in High Energy and Nuclear Physics (CHEP) 2004*, Interlaken, Switzerland, September 2004.