Using Visualization and Data Analysis to Understand Critical Structures in Massive Time Varying Turbulent Flow Simulations

Kelly P. Gaither (1), Hank Childs (2), Karl W. Schulz (1), Cyrus Harrison (3), William Barth (1), Diego Donzis (4), and P.K. Yeung (5)

(1) = University of Texas at Austin, (2) = Lawrence Berkeley National Laboratory,
(3) = Lawrence Livermore National Laboratory,
(4) = Texas A&M University, (5) = Georgia Tech University

DISCLAIMER: This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

Acknowledgements: The authors wish to thank the Texas Advanced Computing Center for access to Ranger and Longhorn and for user support provided on both those resources. This work was supported in part by the Longhorn XD Visualization grant from the Office of Cyberinfrastructure at the National Science Foundation. This work was also supported by the Director, Office of Advanced Scientific Computing Research, and Office of Science, 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 Analysis Center for Enabling Technologies (VACET).

Submitted to special issue of Computer Graphics and Applications (2012), pp. 1–11

Using Visualization and Data Analysis to Understand Critical Structures in Massive Time Varying Turbulent Flow Simulations

Kelly P. Gaither, Hank Childs, Karl W. Schulz, Cyrus Harrison, William Barth, Diego Donzis, and P.K. Yeung

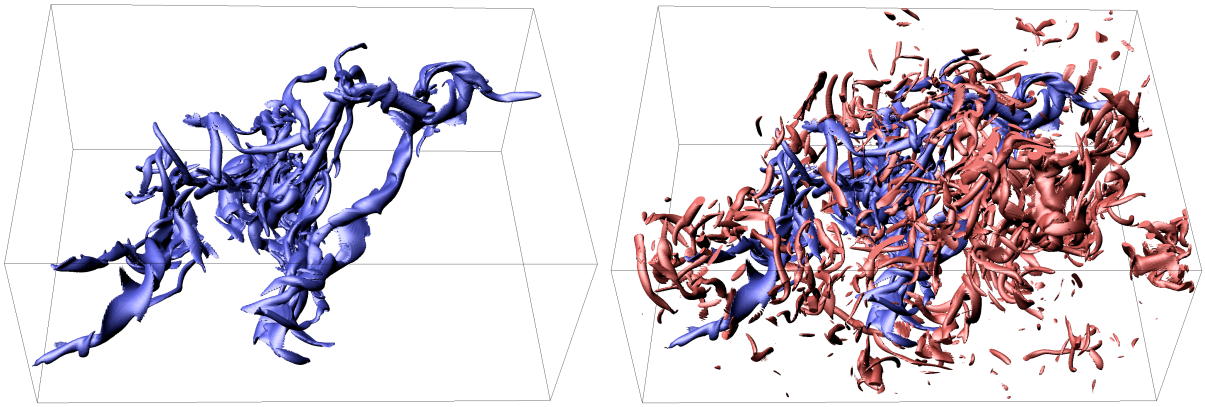

Figure 1: *On the left, a single, connected component of high enstrophy, colored blue. This component is surrounded by hundreds of smaller connected components of high enstrophy. On the right, we render those smaller components as well, in red. The region displayed in these images represents less than one millionth of the simulation's total volume; the* 4096^3 *turbulent flow data set we studied had millions of smaller components (like the red ones) and thousands of larger ones (like the blue one). Our study to better understand the turbulent flow required only the large components, so a critical first step was to identify and exclude the smaller ones.*

Abstract

Turbulence, the most common state of fluid motion in nature and engineering, is a Grand Challenge for the physical and computer sciences. Effectively simulating the wide range of non-linearly interacting three-dimensional fluctuations typical of applications requires the largest supercomputers in the world today. The concept of small-scale universality is critical to our fundamental understanding and also to model and predict turbulent flows. Two important descriptors of small-scale motions are the energy dissipation rate and vorticity. Although it is known that intense vorticity and dissipation tend to concentrate in filamentary and sheet-like structures respectively, their evolution is much less understood or quantified. This paper details the crucial role that visualization and data analysis play in analyzing and understanding a turbulent flow simulation of 4096^3 *cells per time slice (= 68 billion cells) and 17 time slices (= 1 trillion total cells). The visualization techniques presented in this paper allow us to investigate the dynamics of intense events individually or as they form clusters. Understanding the geometrical and dynamical descriptions of these intense events allow scientists to get closer to a more complete understanding of turbulent flows and more accurate models for engineering applications.*

1 Introduction

Massive supercomputers are needed to capture the wide range of non-linearly interacting three-dimensional turbulent fluctuations typical of real world applications. Critical to our fundamental understanding and our ability to model and predict turbulent flow is the concept of small-scale universality, which means that small scale turblence is generally assumed to be relatively insensitive to orientation. Two important descriptors of these small-scale motions are energy dissipation rate and vorticity, representing local straining and rotation respectively. Although it is known that intense vorticity and dissipation tend to concentrate in filamentary and sheet-like structures, their evolution is much less understood or quantified. It is only recently that computational power has allowed us to reach high Reynolds numbers closer to those in applications with resolution to capture the finest details at the smallest scales.

Visualization and analysis of these turbulent data sets provides a powerful mechanism for understanding these complex data sets. However, there are significant technical challenges that must be overcome to extract the underlying information. Visualizing and understanding the interesting bits of information in data with a wide range of length scales requires a varied toolbox of methods that combine analysis tools capable of gleaning information at very small scales with visualization tools capable of illustrating global behaviour. Further, as simulations resolve to finer and finer mesh resolutions, previous visualization and analysis approaches become increasingly unwieldy. In this paper, we present methods for visualizing and analyzing massive turbulent flow simulations, with the goal of being able to describe and understand the characteristics of coherent structures over time. The visualization techniques presented in this paper provide an opportunity to investigate the dynamics of intense events individually or as they form clusters as part of a multi-scale hierarchy. Understanding the geometrical and dynamical descriptions of these intense events constitutes a critical step towards a more complete understanding of turbulent flows and more accurate models for engineering applications.

The contributions of this work to the state of the art are:

- The development of a system for automatic feature detection, extraction and classification that is well suited to large-scale, temporally infrequent data, by augmenting traditional methods (i.e. spatial location) with “fingerprints" provided from shape characterization methods;
- A discussion of the infrastructure necessary to visualize and analyze a very large turbulent flow simulation (4096^3 cells per time slice and over one trillion cells overall); and
- Novel applications of chord length distributions to inform and compare shape.

We begin by providing a brief overview of the direct numerical simulation of turbulent flow and presenting the central science questions that provide the impetus for this work [Sections 2 and 3]. We then present the application of our techniques to the visualization and analysis of turbulent flow data [Section 4], including details presented on informing shape [Section 4.1] and component movement over time [Section 4.3]. We then describe how chord length distributions can assist with characterizing shape in time varying flow [Section 4.4] and present our results [Section 5], a performance overview [Section 6], and conclusions and future work [Section 7]. Comparisons to previous work are made throughout the paper as methods are presented.

2 Turbulent Flow

Turbulence is the most common state of fluid motion and a very difficult problem in the physical sciences. Advances in understanding turbulence and the ability to model its effects are critical in many applications such as aerospace vehicle design, combustion processes, and environmental quality. Turbulent flows are characterized by nonlinear stochastic fluctuations in time and three-dimensional space over a wide range of scales. Direct Numerical Simulation (DNS) [MM98] computes fluctuations at all scales. The incompressible DNS formulation computes instantaneous velocity and pressure fields according to the Navier Stokes equations as:

$$\frac{\partial u}{\partial t} + u \cdot \bigtriangledown \frac{p}{\rho} + v \bigtriangledown^2 u.$$

The main difficulty in DNS is that the range of scales, and hence the computational demands in both speed and memory, increases strongly with the Reynolds number, usually high in applications. Numerical simulation of high Reynolds number turbulence is a grand challenge problem in high-performance computing, with tremendous opportunities for scientific progress [Jim03].

The simulation we study [DYP08] computed energy dissipation and enstrophy on a 4096^3 rectilinear data set. The data set consists of 17 timesteps, each computationally expensive to compute and quite large in size (0.5 TB each).

3 Central Science Questions

Traditional analysis of this turbulent flow data focuses on the shape and structure of enstrophy – the integral of the square of the vorticity. Scientists want to understand what dissipation looks like in the area surrounding areas of high enstrophy. Classically, they have done this by placing a volume rendering of dissipation around isosurfaces of enstrophy. Figure 2 shows that, although this approach can enable scientists to effectively infer trends visually at low resolutions, as the resolutions approaches the 4096^3 size, visual inference of small scale features is difficult to impossible.

Since we are dealing with such a large data size, our approach is to augment the visual approach with statistical analysis; we want to use statistics to find the things “worth

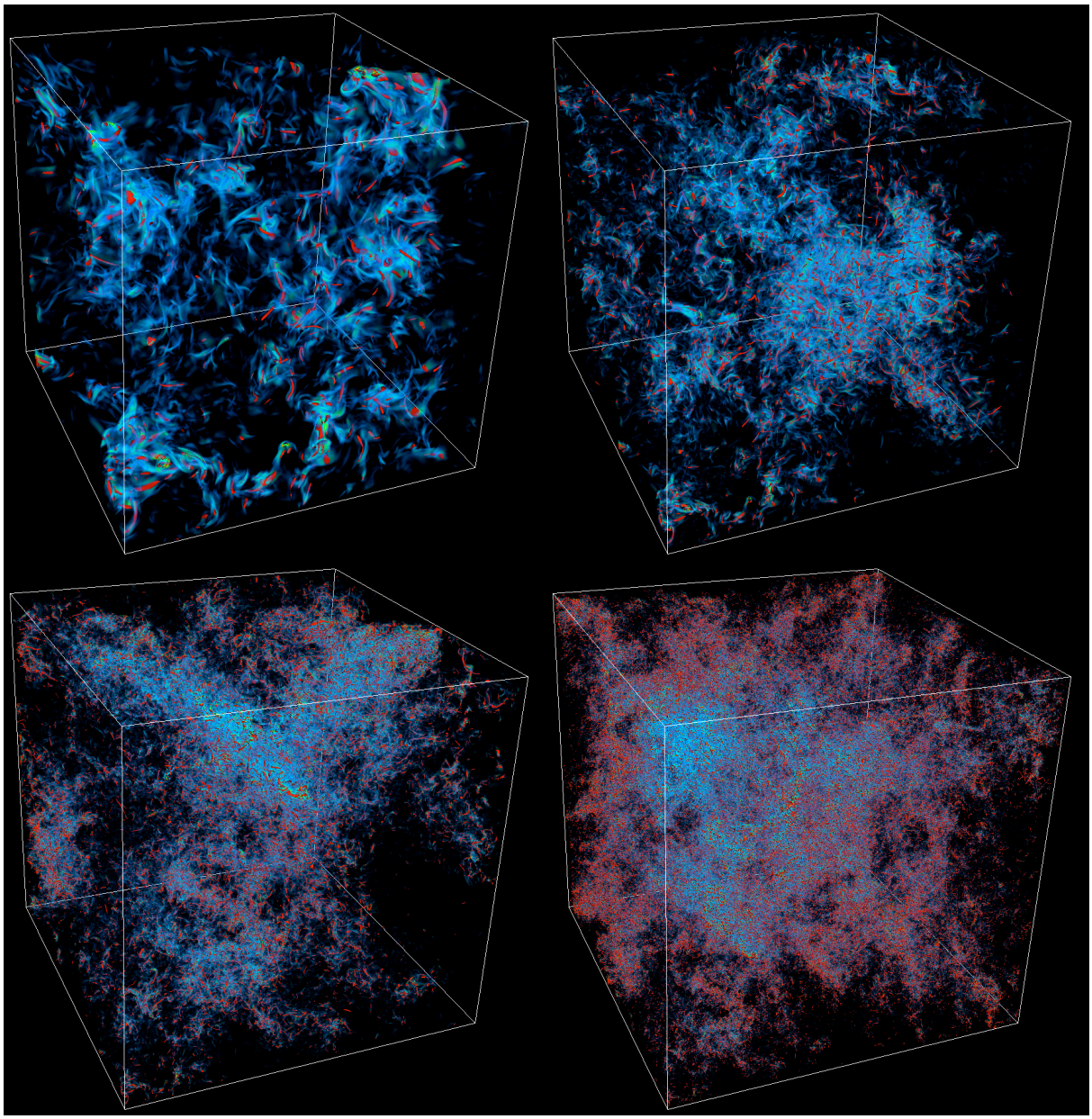

Figure 2: *Contour of enstrophy and volume rendering of dissipation for four different data sizes. The upper left image is a* 512^3 *portion of the data set, the upper right image is* 1024^3*, the lower left image is* 2048^3 *and the lower right image is the full* 4096^3 *data set. These images illustrate how visualization techniques that are effective at smaller data sizes can lose their efficacy at massive scale because the number of features grow so quickly.*

seeing." Both fields (energy dissipation and enstrophy) exhibit what we call “intermittency," the property of very intense but localized events in space. We are interested in seeing how these fields look in space, particularly if high-activity regions for these two fields happen at the same locations or close to each other, and in understanding how these structures of enstrophy behave over time. In particular, we would like to get a reasonable understanding of persistence, structure interactions and path trajectory. In summary, our central science questions are:

- Can we identify and understand enstrophy structures in these turbulent flow data sets?

- Can we characterize shape and size of these enstrophy structures over time?
- Can we track and characterize clusters of enstrophy over time?
- What does dissipation look like in areas surrounding clusters of enstrophy?

4 Visualization and Analysis of Turbulent Flow Data

Visualizing turbulent flow to get a global understanding of the data is not new. There have been a significant number of publications dedicated to visualizing and analysing the complex nature of these typically large data sets. Hin, et. al [HP93] presented a method for the visualization of three-dimensional turbulent flow using particle motion animation. Mynett, et. al [MSH95] detailed a method for tracing particles in turbulent flow simulations. Johnson, et. al [JCG08] presented a method for the interactive visualization of large scale turbulent flow simulations. Each of these papers used visualization to gain a better global understanding of these complex data sets. In this work, we attempt to gain a better understanding of all aspects of the turbulent flow - scale, complexity, production, destruction, and correlation by using a combination of data analytics and visualization techniques.

Rather than developing a new standalone visualization system for investigating turbulent flow, we chose to begin with a popular visualization system that works well at large scales. VisIt is a freely downloadable open-source visualization package that has proven capable of handling very large data sets [CPA*10], making it a good fit for our 4096^3 turbulent flow data. Although we made use of well known visualization and analysis techniques, we were also interested in the less traditional techniques VisIt provided, such as connected component analysis and chord length distribution computation. To investigate the scale, complexity, production, destruction, and correlation of features, we analyzed this data in five distinct phases:

1. **Visual Inspection:** We applied a variety of visualization techniques, such as contouring, volume rendering, and slicing to better understand the global nature of the flow fields and to determine whether there existed readily apparent structures from visual inspection alone. This work gave us a rough knowledge of length scales of enstrophy structures and provided intuition about analysis directions.
2. **Feature Identification:** The application scientists were aware a priori that regions of high enstrophy (“worms") required further study. We discuss our method for identifying these features in 4.1.
3. **Shape Characterization:** Tracking the worms through conventional methods was not possible because the data was stored so infrequently in time. To overcome this problem, we calculated chord length distributions, using them as “fingerprints" to augment traditional tracking methods. We discuss the calculation of chord length distributions in 4.2.
4. **Feature tracking:** We tracked the paths of all worms over time, including component creation, dissipation, amalgamation, bifurcation, and continuation. We discuss our feature tracking using shape characterized by chord length distributions in 4.3.
5. **Shape analysis over time:** We use the tracking discussed in 4.3 and the chord length distributions discussed in 4.2 to study changes in the worm’s shapes over time. The shape analysis is discussed in Section 4.4.

4.1 Feature Identification

We isolated regions where enstrophy was greater than a fixed value, α, provided by the applications scientist. Cells that were entirely above α were retained. Cells that were entirely below α were discarded. Cells that straddled α were subdivided to isolate the portion with values greater than α. This process created tetrahedrons, pyramids, and other cell types to represent the desired portion of the original cell; their faces corresponded to either the original cell boundary or to an isosurface at α. All remaining cells were then put through a connected components identification process. This allowed the total number of components to be counted and for each cell to be classified with the component it belonged to (i.e. which worm is a given cell part of). After the connected sub-meshes were identified and labelled, individual component properties (centroid, bounding boxes, volume, etc.) were calculated. Components that had a total volume below some fixed value, β, were eliminated. This reduced the number of components for each time slice from millions to hundreds. Removing these small worms allowed us to track the larger, more salient structures.

As mentioned in the motivation, it was critical for our effort to employ techniques that work well for the very large data sizes inherent to turbulent flow. All of the techniques discussed in this section meet this criteria. (See Section 6 for more discussion of performance and scalability.)

4.2 Using Chord Length Distribution to Characterize Size and Shape

A chord length distribution is a probability distribution over lengths. It characterizes shape and is well studied in the nuclear reactor community [MRD03]. If a line intersects a shape Q to form a chord, the chord length distribution for that shape, $\Theta_Q(l)$, describes the probability that a chord will be a certain length, l. Chord length distributions are well suited for the scale and complexity of turbulent flow data: the method is indifferent to complexity of shape and they can be calculated efficiently in a distributed memory parallel setting by using a stochastic, line scan-based approach. Further, the distributions can be calculated over all components simultaneously, meaning the presence of a large number of components (as is the case with turbulence) does not adversely affect runtime.

Chord length distributions can be efficiently calculated in a distributed memory parallel setting using a stochastic, line scan-based approach. Every processor utilizes the same set of uniform density, random lines. Each processor intersects its own components. The resulting intersections are exchanged so that the intersections from a single line are wholly contained on a single processor (necessary since components are distributed over processors). The intersections are analyzed, and the results over all processors are summarized. This process is further described in [Chi06]. Fortunately, the distributions can be calculated over all components simultaneously, making this efficient for a large number of components. We calculated the chord length distribution for each of the large worms (components with volume $> \beta$). When evaluating a chord, we query which component the chord came from and only update the distribution specific to that chord. Note that no chord can span multiple components because that would infer that those individual components are connected.

We used one hundred million lines to calculate the chord length distributions for each time slice. This number of lines was required to obtain sufficient statistics for smaller components. We conducted a convergence study examining the effect of the total number of lines on the chord length distribution. We looked at line counts of one hundred thousand, one million, ten million, and one hundred million, running two tests for each line count with different sets of random lines. For each line count, we looked at the L2-norm between the resulting chord length distributions. The average error only dropped below 1% when using one hundred million lines.

4.3 Using Chord Length Distribution to Characterize Movement and Particle History

Tracking the behaviour of features over time is not a new concept. Silver's work detailed seminal efforts for classifying time-varying features in turbulent flow fields [SW98]. The authors characterized all events as being one of continuation, creation, dissipation, bifurcation, and amalgamation and presented a general solution for tracking features over time. They detailed the process of matching an event in one time step with a corresponding event in the subsequent time step. This method of classification was derived from viewing animations of three-dimensional data sets. The tracking method presented in this paper is an extension of that fundamental work. Rather than viewing animations to help inform our decision making, we start with the components identified in section 4.1 and analyze their properties: spatial location, bounding box information, volume, and the chord length distributions discussed in section 4.2.

Of particular significance to us is the movement and behaviour of the connected components over time. Recall that a connected component is defined as a collection of cells that are connected spatially. These connected components computed in VisIt allow us to spatially segment clusters of activity, thus giving us isolated components to track over time. To classify the behaviour of a connected component over time, we can track the following:

Component Continuation. The path a connected component takes over time can be found by segmenting, recording and tracking the continuation of components over time.

Component Creation. The creation or birth of a new connected component can be tracked by finding the introduction of a new particle with no predecessor.

Component Dissipation. The dissipation or death of a connected component can be found by tracking the disappearance of an existing connected component.

Component Amalgamation. The amalgamation of two or more connected components is closely related to the vortex stretching mechanism or vortex reconnection. We can isolate vortex reconnections by tracking the merging of two or more connected components.

Component Bifurcation The bifurcation of a connected component into two or more subsequent connected components can be useful in understanding when there is a direct energy transfer from large to small scale. This bifurcation is closely related to a "vortex stretching" mechanism.

Isolating this behaviour allows us to better understand what is happening in localized regions of the flow field and regions where intense events occur. To do this, we must first understand where change is maximized in the flow field.

VisIt allows us to output a file for every time step containing statistical information about each of the connected components including the centroid, a unique identifier for the connected component, the number of cells in the connected component, the volume, and an axis aligned bounding box. Additionally, we created and used a database of information containing chord length distributions for each of these connected components over time.

Initially, we construct a matrix containing the spatial distances from one component in a given timestep to all other components. We also include the shape information informed from the chord length distributions. This shape information quantifies the shape change between components in subsequent timesteps. We compute the nearest neighbor for all components throughout all timesteps. Additionally, we heuristically set a threshold value to determine whether components are spatially near each other. This threshold value is based on component distances across all time steps and was set by looking at the distribution of these distances.

4.3.1 Component Dissipation – Destruction

Of particular interest in turbulent flow is destruction. Tracking component dissipation is a good way to study and analyze destruction in turbulent flow. Examining the ratio of component creation to dissipation gives us a reasonable way to study the ratio of turbulent production versus turbulent destruction. We can track component dissipation by using our matrix containg spatial distance and shape:

- For a given connected component, find the nearest neighbor in the subsequent timestep.
- If the spatial distance is greater than our preset threshold, flag that component as having dissipated.

4.3.2 Component Creation – Production

In contrast to component dissipation is the concept of component production. We can study production by studying the creation of connected components over time, allowing us to understand how frequently particles are produced. We track component creation using our matrix containg spatial distance and shape:

- For a given connected component, find the nearest neighbor in the previous timestep.
- If the spatial distance is greater than our preset threshold, flag that component as a creation.

4.3.3 Component Amalgamation – Vortex Reconnection

When two or more connected components merge or amalgamate at a given location in time to form a single connected component in the subsequent timestep, we flag those components as amalgamations. This means that when we see an amalgamation component, we know that it will merge into or become part of a component in the subsequent timestep. The physical equivalent of component amalgamation is vortex reconnection. Tracking these vortex reconnection events allows us to better understand how and why these components merge and what is happening in the flow in the surrounding area. We track component amalgamation after both component dissipation and component creation is detected:

- Cycle over all nearest neighbors for the connected components in the matrix. Because we have flagged component dissipations and creations already, all nearest neighbors fall within the threshold value for what we consider close. We increment a counter to count the number of components that are clustered close to components in the next timestep.
- For all components in all timesteps, find those components whose near neighbor count is greater than 1 and whose shape changes appreciably given a predefined threshold heuristically chosen. Flag all those neighboring components from the previous timesteps as amalgamations.

4.3.4 Component Bifurcation – Direct Energy Transfer

To provide information about destruction, we track direct energy transfers from large to small scales. A bifurcation or direct energy transfer happens when a connected component in one timestep splits into two or more connected components in the subsequent timestep. These critical events offer additional understanding of how components split and progress over time, allowing us to examine locally which components split and what the behavior of the flow is in the surrounding area. By tracking these amalgamations or energy transfers, we can examine what is occurring prior to the bifurcation. We track these component bifurcations after flagging the component amalgamations:

- Cycle backward in time over all nearest neighbors for the connected components in the matrix. Because we have flagged component dissipations, creations and amalgamations, we have the counter for the neighbors near a given component.
- For all components in all timesteps, find those components whose near neighbor count is greater than 1. Flag all those neighboring components from the previous timesteps whose shape changes appreciably as bifurcations. Bifurcations represent those components in a given time step that split into 2 or more components in the subsequent timestep.

4.3.5 Component Continuation

Tracking components that continue from one timestep to another allows us to track the path a connected component takes over time. Using this information, we are able to isolate components that continue in subsequent timesteps and whose shape and volume does not change measurably over time. Because we have flagged all dissipations, creations, amalgamations and bifurcations, we are left with components that have single nearest neighbors that fall within the predefined threshold.

4.4 Using Chord Length Distribution to Analyze Change in Shape Over Time

One goal in this study was to charactere the “thickness" of the high enstrophy “worms." Although thickness is well defined for simple shapes, it is tougher to define for complex shapes like our worms (see Figure 3). In [HR97], the authors present a definition where the thickness at each point is the radius of the largest sphere that contains the point and is fully contained within the shape. However, this process is not well suited to a distributed memory environment or to the unstructured meshes we produce when isolating high enstrophy regions (see 4.1). We must focus on techniques that work well with the big data inherent to turbulent flow. Therefore, we favored a definition that describes the proportion of the volume at a given thickness. Let $C(P,\omega)$ be the chord of some shape Q that goes through point P at direction ω and define $L(P,\omega,l)$ as 1 if $C(P,\omega)$ is of length l and 0 otherwise. Then, $V(l)$, the proportion of the volume at thickness l, is:

$$V(l) = \int_Q \frac{\int L(P,\omega,l)d\omega}{\int 1 d\omega} dV$$

Further, this quantity can be derived, in a straightforward fashion, from the line scans used to calculate a chord length distribution [Chi06]. Once thicknesses are calculated, we can use the tracking from 4.3 to study how thickness changes over time.

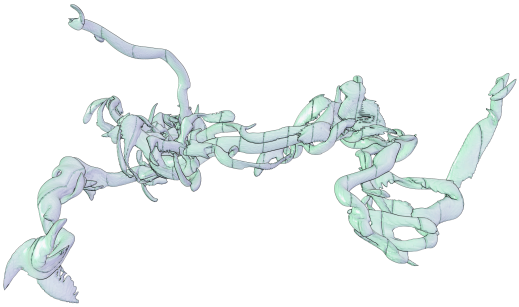

Figure 3: *This component of high enstrophy presents difficulties in measuring thickness. Some regions appear to have a "major axis" where the orthogonal axis can be considered thickness. Other regions are more rounded and the thickness there is ambiguous.*

5 Results

As we have discussed, we are interested in gaining a better understanding of the interaction and correlation of enstrophy and energy dissipation. Using connected component analysis, we successfully isolated high enstrophy "worms," allowing us to focus our analysis in and around those components. As shown in Figures 2 and 3, isosurfaces of enstrophy look like filaments that appear to be concentrated in areas of localized high dissipation. The success of our methods, as applied to turbulent flow, is measured by whether or not we can respond to our central science questions stated previously.

5.1 Can we identify and understand enstrophy structures in these turbulent flow data sets?

We started with the raw statistical files on connected components. These raw files, alone, however, provided entirely too much information to work with and are no more informative than looking at a global volume visualization of the entire data set. For this reason, we chose to work with a subset of the data by iteratively refining the threshold against two criteria: 1) isovalues of enstrophy defined by the applications scientists, and 2) keeping connected volumes of enstrophy above a specified threshold. This threshold value was chosen heuristically such that we could see and track these components over time and to reduce the large number of components to initially analyze.

5.2 Can we inform shape and size of these structures over time?

We were interested in better understanding the basic shape that these high enstrophy "worms" take on over time. Initially, we sought to understand the correlation between volume and thickness using our pre-computed chord length distributions to calculate $V(l)$. We can calculate the average thickness for each component:

$$\int l \cdot V(l) dl$$

We speculated that components with large volumes would be thicker, but found that this wasn't true: large volumes had consistently low thickness and components with high thickness had small volumes as shown in Figure 4.

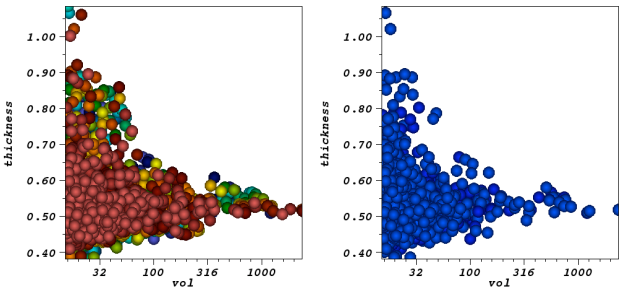

Figure 4: *A scatter plot of volume versus thickness for each worm. The image on the left included all time slices and is colored by time slice. The image on the right included only the first time slice. The shapes of the scatter plots don't substantially change from left to right, indicating that the association between thickness and volume do not change significantly over time.*

After studying this visually, we saw that large components are formed out of many thin shapes, rather than a large mass. Components with small volumes span thicknesses from small to large. The components with high thickness are likely ones that contain some sub-feature that increase its size (for example a larger sphere or a long tube). For large components, however, the presence of a large subfeature is less meaningful for the overall distribution and is unable to skew the thickness up appreciably.

We then plotted average thickness against time for each of the worms (with four of these plots shown in Figure 5). The worm represented in red maintains a fairly constant and small thickness over time. The spatial extents of the volume are 256^3. The cyan curve shows a worm that varies more significantly and shows a trend where the worm becomes thinner in the middle of the simulation before ultimately thickening out. The green and dotted blue curves show two worms that merge at the tenth time slice.

This analysis served two purposes. First, it served to validate the tracking process. Because the change in thickness from time slice to time slice appeared to follow consistent trends over a larger time window, we were able to conclude that the tracking itself was successful. Second, we were able to look for global trends in the changes in worm thickness. We observed no unifying trend for the worms: some got thicker, some got thinner, some got thicker and then thinner,

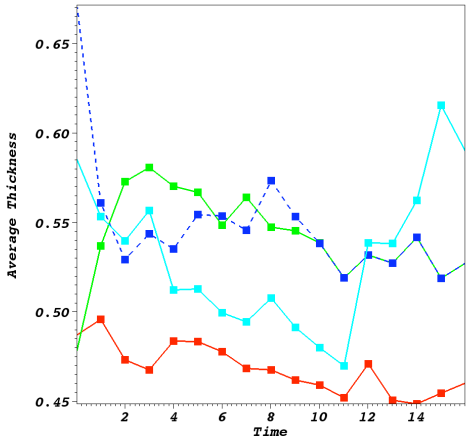

Figure 5: *The average thickness of four worms over time.*

some stayed about the same thickness, etc. Figure 6 shows a single plot with all of the worms.

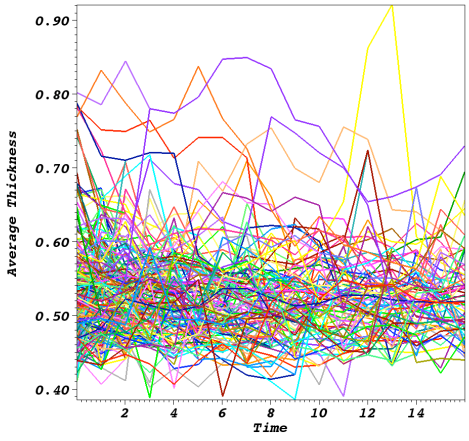

Figure 6: *Plotting the change in average thickness over time for all worms. The majority of worms had average thickness ranging from 0.45 to 0.6.*

5.3 Can we track and characterize clusters of enstrophy over time?

For this work, we isolated high enstrophy components and tracked their behavior over time. We successfully characterized and tagged component dissipations, creations, amalgamations, bifurcations and continuations, meaning that we were able to detect turbulent production and destruction, vortex reconnections, and direct energy transfers. We obtained some expected results as well as some unexpected results

Total Components Dissipated	Total Components Created	Total Components Amalgamated	Total Components Bifurcated
808	463	93	363

Table 1: *Number of component dissipations, creations, amalgamations, bifurcations, and continuations over the 17 time steps.*

that bear further investigation. The general statistics of the tracking are shown in Table 1.

It is interesting to note that the number of structures or components has to be roughly constant in a stationary state. If we look at the number of components dissipated, created, amalgamated, and bifurcated over time, we should in theory get a zero sum. We can see that if we add 463 creations - 808 dissipations - 93/2 amalgamations + 363 bifurcations, we get -27 which is a small number consistent with a statistically stationary state. We divide the amalgamations by 2 since we are only losing half of the components from one timestep to another. We can see that there are about half as many components created than dissipated. This is interesting from the physical perspective and bears further consideration. We can also analyze the component behaviour at each of the time steps and begin to get an idea of how the flow progresses. The number of component creations (or productions) detected over the time series is shown in blue in Figure 7. It shows a relatively constant number of components being produced.

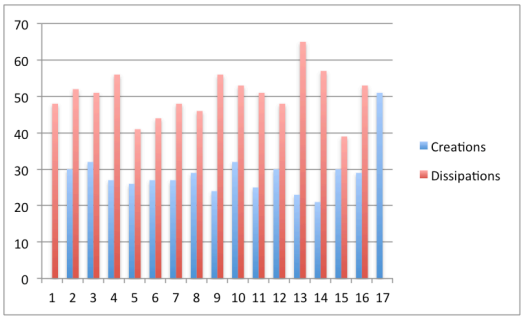

Figure 7: *Component dissipations and creations over time.*

Also shown in Figure 7 is the number of component dissipations over time. Of particular interest is the relative number of component dissipations (or destruction) versus component creations (or production). On average, two to three times as many component are destroyed than produced.

Tracking amalgamations (vortex reconnections) allows us to flag the merge of multiple components (more than two), giving us the ability to more closely examine the properties in the surrounding area. Figure 8 shows the number of component amalgamations. The graph shows the general trend

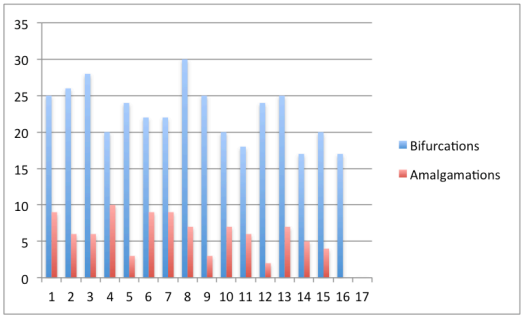

Figure 8: *Component amalgamations and bifurcations over time series.*

over time suggesting that the data has a steady rate of merging and mixing over time.

We can track component bifurcations (direct energy transfers) in a similar manner. Figure 8 also shows the number of component bifurcations over time.

While there is still much investigation to be done, we are encouraged by the successes that we have had in our ability to track and characterize the connected components as these components directly equate to clusters of enstrophy present in the turbulent flow.

5.4 What does dissipation look like in areas surrounding clusters of enstrophy?

Figure 9 shows two images: one with a volume rendering of dissipation and one with that volume rendering in the context of large components of high enstrophy. Suprisingly, the highest dissipation regions visually correlated with the large components; they were visible when only volume rendered, but obscured when the components' surfaces were added. This led to a major finding for our science stakeholders: values of high dissipation occur in large components of high enstrophy, but it does not occur in small components. Figure 10 shows the histograms that supported the initial observation.

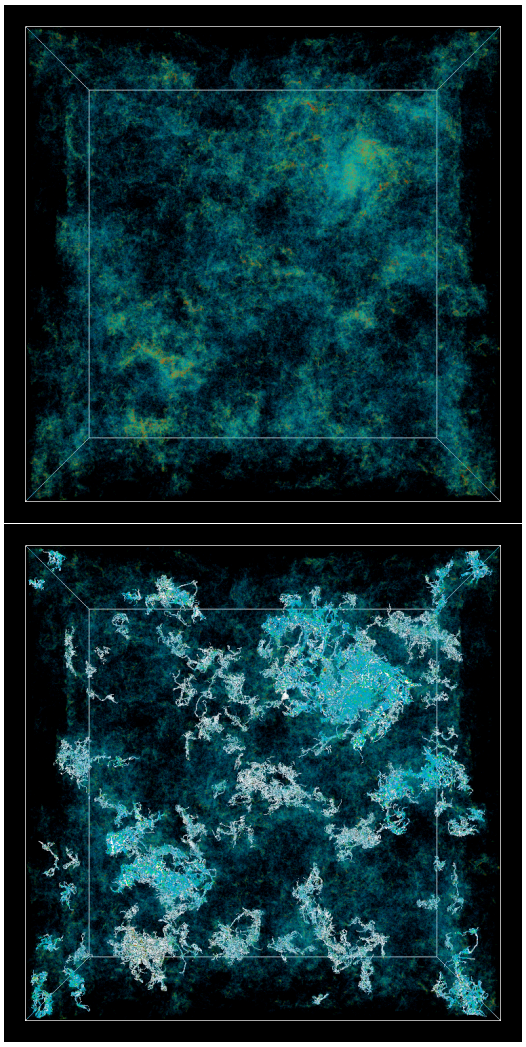

Figure 9: *Top: Volume rendering of energy dissipation. Bottom: Volume rendering of dissipation with high enstrophy components shown in white.*

6 Performance, Scalability, and Parallelism

Our analysis was run on Longhorn, a 2048 core, 256 node, 512 GPU cluster hosted at the Texas Advanced Computing Center. Each node contains two quad-core Intel Xeon E5540 "Gainestown" processors and 48 GB of local RAM. All nodes are connected via a Mellanox QDR InfiniBand switch, and we use MVAPICH2 v1.4 for our MPI implementation. Our parallel analysis routines used one quarter of the machine (64 nodes).

The analysis itself was done in phases. Steps 1-3, visual inspection, feature identification, and shape characterization were all carried out using VisIt and its distributed-memory parallelization techniques. Most of the algorithms used were embarrassingly parallel and thus scalable. Connected component identification is not embarrassingly parallel, but the algorithm we used performs well at large scale [HCG11]. Similarly, chord length distribution calculation requires an all-to-all coordination phase, but again relies on a scalable algorithm [Chi06]. Both algorithms allowed the components to remain distributed over their originating processors, which was crucial for ensuring that our approach scaled to massive data.

The execution time for the parallel processing of the 4096^3 data varied: I/O performance is variable based on contention for the file system and operating system caching effects; communication times vary based on contention and the

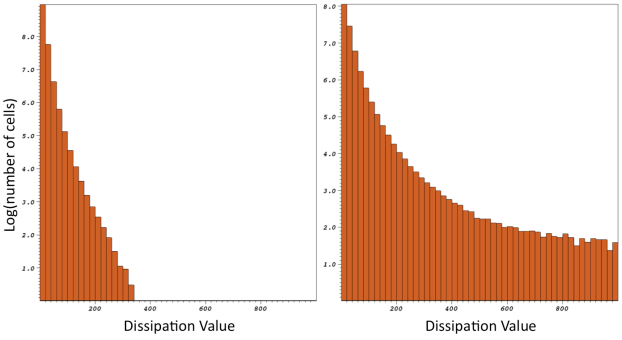

Figure 10: *Both histograms show dissipation values ranging from 0 to 1000 for cells with enstrophy* $> \alpha$*. The left histogram is over cells that are part of a* ***small*** *connected component (volume* $< \beta$*). The right histogram is over cells that are part of a* ***large*** *connected component (volume* $> \beta$*). These histograms show that high dissipation values occur in large components, but not in small ones.*

Phase	Time	Comment
Read	~1min	High variability based on caching and contention
Isovolume	~10sec	Produces unstructured mesh from rectilinear data, increasing memory footprint
Connected Component Identification	~1min	Large communication phase dominates performance
Chord Length Distribution	~100min	100M lines with billions of cells Room for further optimization

Table 2: *Approximate time spent for each processing phase.*

amount of data being communicated, which varied greatly based on enstrophy value (α) and volume thresholds for worms large enough to consider (β), among others. Hence we provide approximate estimates of time spent in the various phases in Table 2. The calculation of chord length distribution took the longest, but this was because it required calculating trillions of intersections.

Steps 4 and 5, feature tracking and shape analysis over time, were done with a serial, standalone, custom program. These programs operated on statistics that were extracted from the 4096^3 data, and thus were not proportional to the total data size. In fact, our total statistical data was approximately ten megabytes. Performance and scalability were not concerns for these programs; the feature tracking and thickness analysis took on the order of seconds, which paled in comparison to their development time, verification time, etc.

7 Conclusions and Future Work

We have presented methods for visualizing and analyzing turbulent flow data that employ new methods for shape characterization and feature detection, extraction and classification. These methods, based on chord length distribution, provide greater insight into these massive complex flow fields and allow us to better understand the physical processes behind turbulence. Using the chord length distribution has allowed us to find a signature or fingerprint for extracted components in what is otherwise very complex flow. This work gave us insight that visual inspection alone would not have otherwise shown. Additionally, we were able to use all these methods at massive scales. It should be noted that this is the trend for turbulent flow simulations and is where the kinds of tools presented in this paper are most critical.

We are encouraged by the results that we have had with this work and will pursue the development of additional techniques to glean further insight into these large-scale data sets. In addition to characterizing the radius of the connected components, we are interested in getting a classification of the component shape (spherical, cylindrical, ellipsoidal, etc.). We suspect and have learned from this study that the components are a combination of these shapes.

Next steps for the feature tracking methods include the automatic correlation of multiple variables surrounding a region of interest. Of significant interest is the ability to determine whether components die and reappear in subsequent time steps. We would also like to better understand how our analysis would change with more temporal data. This is not possible with the current data set, but may be possible with future data sets. Finally, we would like to better understand how our analysis can be augmented by the subset of traditional topological analyses that do scale to large data.

8 Acknowledgments

The authors wish to thank the Texas Advanced Computing Center for access to Ranger and Longhorn and for user support provided on both those resources. This work was supported in part by the Longhorn XD Visualization grant from the Office of Cyberinfrastructure at the National Science Foundation. This work was also supported by the Director, Office of Advanced Scientific Computing Research, and Office of Science, 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 Analysis Center for Enabling Technologies (VACET).

References

[Chi06] CHILDS H.: *An Analysis Framework Addressing the Scale and Legibility of Large Scientific Data Sets.* PhD thesis, Computer Science Department, University of California, Davis, One Shields Avenue, Davis, Ca, 95616, 2006.

[CPA*10] CHILDS H., PUGMIRE D., AHERN S., WHITLOCK B., HOWISON M., PRABHAT, WEBER G., BETHEL E. W.: Extreme Scaling of Production Visualization Software on Diverse Architectures. *IEEE Computer Graphics and Applications 30*, 3 (May/June 2010), 22–31.

[DYP08] DONZIS D., YEUNG P., PEKUROVSKY D.: Turbulence simulations on o(10000) processors. In *Proceedings of TeraGrid 2008 Conference* (2008).

[HCG11] HARRISON C., CHILDS H., GAITHER K. P.: Data-Parallel Mesh Connected Components Labeling and Analysis. In *Proceedings of EuroGraphics Symposium on Parallel Graphics and Visualization* (April 2011), pp. 131–140.

[HP93] HIN A. J. S., POST F. H.: Visualization of turbulent flow with particles. In *Proceedings of the 4th conference on Visualization '93* (Washington, DC, USA, 1993), VIS '93, IEEE Computer Society, pp. 46–53.

[HR97] HILDEBRAND T., RÃIJEGSEGGER P.: A new method for the model-independent assessment of thickness in three-dimensional images. *Journal of Microscopy 185*, 1 (1997), 67–75.

[JCG08] JOHNSON G. P., CALO V., GAITHER K. P.: Interactive visualization and analysis of transitional flow. *IEEE Trans. Vis. Comput. Graph. 14*, 6 (2008), 1420–1427.

[Jim03] JIMÉNEZ J.: Computing high-Reynolds-number turbulence: will simulations ever replace experiments? *Journal of Turbulence 4*, 1 (2003), 1–13.

[MM98] MOIN P., MAHESH K.: Direct numerical simulation: A tool in turbulence research. *Annual Review of Fluid Mechanics 30*, 1 (1998), 539–578.

[MRD03] MAZZOLO A., ROESSLINGER B., DIOP C.: On the properties of the chord length distribution, from integral geometry to reactor physics. *Annals of Nuclear Energy 30*, 14 (2003), 1391–1400.

[MSH95] MYNETT A. E., SADARJOEN I. A., HIN A. J. S.: Turbulent flow visualization in computational and experimental hydraulics. In *IEEE Visualization* (1995), pp. 388–391.

[SW98] SILVER D., WANG X.: Tracking scalar features in unstructured datasets. In *IEEE Visualization* (1998), pp. 79–86.