J. Parallel Distrib. Comput. 120 (2018) 115–126

Contents lists available at ScienceDirect

J. Parallel Distrib. Comput.

journal homepage: www.elsevier.com/locate/jpdc

Remote visual analysis of large turbulence databases at multiple scales

Jesus Pulido [a,b,*], Daniel Livescu [b], Kalin Kanov [c], Randal Burns [c], Curtis Canada [b], James Ahrens [b], Bernd Hamann [a]

[a] Department of Computer Science, University of California, Davis, CA 95616, USA
[b] Los Alamos National Laboratory, Los Alamos, NM 87544, USA
[c] Johns Hopkins University, Baltimore, MD 21218, USA

HIGHLIGHTS

- Data-level wavelet compression reduces bandwidth, memory and compute footprint.
- Latency is improved between database components and multi-user support is introduced.
- Remote visualization is enabled on a large database cluster using commodity hardware.
- New wavelet analysis tools are demonstrated for existing turbulence data cluster.

ARTICLE INFO

Article history:
Received 25 January 2017
Received in revised form 7 May 2018
Accepted 31 May 2018
Available online 15 June 2018

Keywords:
Databases
Wavelets
Data reduction
Remote visualization
Distributed systems
Turbulence

ABSTRACT

The remote analysis and visualization of raw large turbulence datasets is challenging. Current accurate direct numerical simulations (DNS) of turbulent flows generate datasets with billions of points per time-step and several thousand time-steps per simulation. Until recently, the analysis and visualization of such datasets was restricted to scientists with access to large supercomputers. The public Johns Hopkins Turbulence database simplifies access to multi-terabyte turbulence datasets and facilitates the computation of statistics and extraction of features through the use of commodity hardware. We present a framework designed around wavelet-based compression for high-speed visualization of large datasets and methods supporting multi-resolution analysis of turbulence. By integrating common technologies, this framework enables remote access to tools available on supercomputers and over 230 terabytes of DNS data over the Web. The database toolset is expanded by providing access to exploratory data analysis tools, such as wavelet decomposition capabilities and coherent feature extraction.

1. Introduction

Extremely large datasets commonly arise in science and engineering today, and it is often prohibitive to store an original massive dataset at multiple sites or transmit it over computer networks in its entirety. Regardless, such datasets represented tremendous scientific value for the broader scientific community. It is imperative to deploy effective technologies enabling the remote access to vast data archives for the purpose of having a large pool of scientists harness their value and make new discoveries. Our analysis framework presented here was driven specifically by the needs articulated by scientists from Johns Hopkins University (JHU) and Los Alamos National Laboratory. JHU hosts a large digital repository of data from several disciplines, including massive simulation data from numerical simulations of turbulent physics. The framework we present makes remote visual analysis possible via an effective protocol controlling the distribution of analysis and visualization steps to be performed on the server side (JHU site) and a remotely connected visualization client.

Furthermore, our framework incorporates the power of approximating a dataset by using cubic (bi-cubic, tri-cubic) B-spline wavelets. The utilization of wavelet approximation allows a user to generate initial previews or simply coarse approximations of a dataset quickly, making possible the efficient identification of specific regions of interest that might warrant an analysis at a more detailed level. We demonstrate our framework for datasets available in the JHU repository and for typical scientific analysis scenarios.

* Corresponding author at: Department of Computer Science, University of California, Davis, CA 95616, USA.
E-mail addresses: jpulido@ucdavis.edu (J. Pulido), livescu@lanl.gov (D. Livescu), kalin@cs.jhu.edu (K. Kanov), randal@cs.jhu.edu (R. Burns), cvc@lanl.gov (C. Canada), ahrens@lanl.gov (J. Ahrens), hamann@cs.ucdavis.edu (B. Hamann).

https://doi.org/10.1016/j.jpdc.2018.05.011
0743-7315/

Most datasets encountered in applications in the physical sciences, similar to most natural images, present lower-dimensional structures whose detection, extraction, and characterization are active areas of research. The search for more efficient algorithms to detect and manipulate such structures has led to the development of a multitude of multi-resolution geometric methods supporting data analysis at multiple scales. One area that significant benefits from the application of such methods is fluid turbulence. In general, turbulent flows contain localized, highly intermittent structures as well as more stable, coherent structures. The characterization of these structures, which interact nonlinearly as they are advected by the background flow and significantly alter the local topology, is an open, fundamental question in the study of turbulence. Multi-resolution representation methods seem ideally suited for such an effort due to their localization in both the real and frequency spaces.

Recently, Pulido et al. [23], studied several multi-resolution representation methods in depth, including B-spline, Daubechies and Coiflet wavelets, curvelets and surfacelets. These methods were compared to determine their ability to capture the structure of fully developed turbulence using a truncated set of coefficients. Methods were evaluated based on their ability to approximate scalar and vector fields, including density and velocity, as well as derived data such as derivatives, spectra, and properties of constant density surfaces. The main criteria used for comparing methods were computational efficiency, numerical accuracy, and degree of data compression. While different methods performed better with respect to various metrics, B-spline wavelets consistently ranked at or near the top of the metrics considered. Except for some of the orthogonal wavelet methods, the use of multi-resolution representation techniques to study turbulence is relatively new and an emerging area of research. In particular, the B-spline wavelets have only been sparsely used for such purpose. Here, we present the addition of the B-spline wavelet representation to the JHU turbulence database. By overcoming bottlenecks in the system, we demonstrate this tool for both remote visualization and novel multi-scale analysis of turbulence data.

1.1. Contributions

This paper contains the following contributions:

- Wavelet compression is introduced at the data-level to reduce access costs, bandwidth, memory and compute footprint, therefore improving latency between database components to support many remote users.
- Remote visualization is made possible for a multi-terabyte database cluster supporting commodity hardware.
- New analysis tools are demonstrated for two datasets for these types of turbulence: a) Homogeneous Buoyancy Driven Turbulence (HBDT) [19] and b) Forced Magneto-Hydrodynamic turbulence (FMHDT) [2].

In Sections 2 and 3, we discuss related works in remote visualization and wavelet compression, describe the public JHU database cluster, and provide a brief introduction to wavelet methods. Section 4 presents a pipeline that introduces wavelet methods and visualization support to the JHU database cluster. Section 5 discusses results focused on measures assessing the performance, quality, and efficiency of using wavelet methods for analysis. Additionally, multi-scale analysis is performed, driven by domain scientists, and results are presented using the newly implemented analysis tools. Our conclusions in Section 6 summarize our results and contributions.

2. Related work

With the continued and rapid growth of the size of datasets, movement of data is increasingly difficult. Database systems, such as the Johns Hopkins Turbulence Database (JHTDB) [22], provide public access via Web services to large datasets in a database cluster. Efficient and remote Visualization and analysis capabilities are crucially important for understanding such large-scale simulated datasets, to gain the desired value and scientific insight from the available data.

Compression for reducing bottleneck behavior in systems for data visualization purposes has been explored previously. Classic methods, such as Haar wavelets, are used in Computational Fluid Dynamics (CFD) for data transmission in Trott et al. [29]. Lippert et al. [17] applied wavelet splats, permitting lossy compression, to enable volume rendering of large datasets. Concerning large-scale and remote data visualization, Guthe et al. [12] proposed the use of hierarchical wavelets in a preprocessing step to reduce hardware requirements for volume visualization applications. While this technique reduces the cost of visualization on standard PCs, interactive data walkthrough leads to rendering times exceeding 1000 ms per frame, unless wavelet coefficients are cached. Woodring et al. [30] used a commercial wavelet standard, Jpeg2000, to perform compression over the network. Lakshminarasimhan et al. [13] described a new error-bound, B-spline-fitting-based method for lossy data compression and performed a direct comparison with wavelets. When comparing the method to traditional Haar and other linear wavelet methods, it has been shown to perform rather poorly for data analysis and visualization purposes [23]. Lindstrom et al. [16] developed a fixed-rate, near-lossless compression scheme targeting floating-point gridded data. While these compression schemes are useful for data storage purposes and transfer over networks, they do not enable the analysis of data for a band-/frequency-specific analysis, which is of great interest in the context of a scale-based analysis of phenomena exhibiting behavior at multiple scales. In our framework, we use cubic B-spline wavelets that significantly improve over linear wavelets, then only stream the coefficients required for a specific scale, without decompression. This approach supports fast adaptive refinement of coefficients to improve the level of detail of an approximation when requested.

Several visualization frameworks and systems have been developed over the past years. Ahrens et al. [1] devised Paraview, which is the underlying system we used for our efforts discussed in this paper. Childs et al. [7] developed another system, called Visit. Both systems are built on the Visualization Tool Kit (VTK) [25] and provide a most of the commonly used and needed data visualization methods. Cedilnik et al. [6] presented several remote visualization schemes and implemented them in Paraview. These schemes are based on streaming compressed images and geometrical data over a network, but they do not address the problem of processing large amounts of data at the location where the data physically reside. In our framework, we compress data at the database level and augment existing features of Paraview, including decompression capability to represent reduced datasets supporting concurrent users in a single node.

Finally, [28] provides an overview of visualization of large turbulence datasets, proposing the use of a GPU-based wavelet methods for data compression. Local large data can be processed quickly on a Desktop client's GPU by using a bricking scheme that reduces workloads into smaller compute blocks. Remote visualization, however, has to contend with the existing server limitations as it is not feasible to transfer all the data locally. Additionally, the existing JHTDB framework is limited by monetary cost to the usage of headless nodes without a GPU. Our wavelet approach is CPU-based by design, and many times does not require a full