Novel Visualization and Analysis for Extreme-Scale Wind Turbine Array Simulations

Kenny Gruchalla (PI)[1], Ryan Elmore[1], Matthew Churchfield[2], Patrick Moriarty[2], John Clyne[3]
[1]Computational Science Center, [2]National Wind Technology Center, [3]National Center for Atmospheric Research

Abstract / Concept

NREL is a world leader in large-scale simulation of multi-turbine arrays through the collaborative work of the NWTC and Computational Science Center. To maintain this leadership role, we need to develop new analysis strategies to address the massive data sets being generated by these simulations. Current wind field data being generated on RedMesa overwhelm traditional storage and analysis approaches, undermining our ability to perform interactive and exploratory analysis and visualization of these data in post processing. This situation is likely to get much worse, quickly, growing by orders of magnitude in the near future, as we dramatically increase NREL's HPC system in the ESIF, as we approach arrays with hundreds of turbines, and as we attempt to further resolve all the relevant scales of wind-energy production. These future models (with current practices) will strictly limit us to analyses designed *a priori* and performed *in situ*. The problem is fundamental to the trajectory of high-performance computing: computational capabilities are rapidly out pacing analysis techniques as well as I/O and storage capabilities. In an idealized setting, domain scientists would have full access to the simulation data after the simulation run to perform interactive queries enabling real-time hypothesis testing, knowledge extraction, and scientific discovery. However, for large-scale simulations this is becoming generally infeasible, and for many computational scientists, an inadequate analysis capability is the largest barrier to scientific discovery [1]. We propose to design and prototype a novel multi-resolution (hierarchical) analysis technique that identifies and extracts regions-of-interest *in situ*, providing post-processing access to data regions believed most salient in full fidelity, while providing access to less salient regions in lower fidelity. We believe through the novel combination of multi-resolution storage and region-of-interest identification and extraction, we can begin developing an infrastructure for the interactive analysis of extremely large-scale wind-turbine array fluid dynamic and aero-elastic simulations. This infrastructure would be directly applicable to other large-scale simulations where the enormity of the data poses an analysis problem (e.g., high-fidelity interfacial dynamics simulations).

This project would position NREL to compete in expected near-future DOE/SC/ASCR FOAs, as ASCR has identified novel data analysis strategies for massive data sets as a top research priority [1]. More generally, this project would be the best demonstration to date of our ability to operate on the research boundary of computational science, which is critically needed for us to compete for the largest and most difficult computational problems in renewable energy (e.g., the resubmission of *Center for Wind and Exascale System Technologies* proposal).

Background

The DOE Office of Science has released a ten-year vision, which mandates the development of "visualization and data management systems to manage the output of large-scale computational science runs and in new ways to integrate data analysis with modeling and simulation." [2] This mandate originates from several recent studies conducted by the DOE [1,3], and DARPA [4] that have concluded that a critical disparity is growing in the field of computational science: our ability to generate numerical data from scientific computations has in many cases exceeded our ability to analyze those data effectively. This situation has recently manifested itself at NREL in the large-scale wind-turbine array simulations being run on RedMesa. The issue is data analysis tools and the computational machinery that supports them have not been able to scale with the high-performance compute (HPC) systems that are generating the data. At NREL, the issue will soon be exacerbated with the much larger ESIF HPC system and more intensive wind-turbine array models. Likewise, the issue is only expected to worsen across the larger computational science community with the architectural changes expected at the exascale.

The imbalance of scale between numerical simulation and data analysis is largely due to their contrasting demands on computational resources. Large-scale numerical simulation is typically a *batch* processing operation that proceeds without human interaction on parallel supercomputers. Data analysis, in contrast, is fundamentally an *interactive* process with a human investigator in the loop, posing questions about the data and using the responses to progressively refine those questions [5]. Furthermore, the algorithmic performance of analysis on large-scale data is largely dependent on system memory and I/O bandwidth, while simulation performance is largely dependent on microprocessor performance. In the past thirty years processor performance has increased seven orders of magnitude, while data transfer rates have only increased by two [6]. The DOE Exascale Initiative Roadmap [3] quantifies the architectural disparities that are expect in the near-future: as system performance increases, the ratio of system memory to system performance and the ratio of system storage to system performance are expected to decrease by more than a factor of 10 (see Table 1).

Metric	**2010**	**2018**	**Factor Change**
System Peak Performance	2 PF/s	1 EF/s	500
System Memory	0.3 PB	10 PB	33
I/O Bandwidth	1.5 GB/s	50 GB/s	33
System Storage	15 PB	300 PB	20

Table 1: DOE Exascale Roadmap Design for 2018 and its relationship to current high-performance computer designs [3]

The team assembled by this proposal is in a unique position to address these challenges in a novel way. The vast majority of large-scale data analysis research in the DOE Office of Science laboratories and SciDAC institutes has been focused on *in situ* processing: the study of the data while it is being generated. An *in situ* approach conducts the visualization and analysis in tandem with the simulation on the same parallel supercomputer. This elevates many storage and data transfer issues; however, it comes at the cost of bias and unquantified uncertainty, as the analysis and visualization parameters must be specified *a priori*. The danger is unknown and unexpected patterns are likely to be overlooked in the data, because there is no opportunity to discover them. As Isaac Asimov (1920-1992) famously observed, "The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka' but 'That's funny…'." Without a mechanism to interactively explore the data from these large-scale simulations, we are much less likely to have those serendipitous moments in data analysis. Kenny Gruchalla (PI) and John Clyne on this proposal team have been focused on developing data reduction and compression strategies that provide schemes to interactively explore large-scale data [7-11]. And the most recent DOE workshop on data analysis as recognized, for the first time, that data reduction strategies are going necessary for scientific discovery at the exascale [1].

Approach

We propose to design and prototype a novel multi-resolution (hierarchical) analysis technique that identifies and extracts regions-of-interest *in situ*, providing post-processing access to data regions believed most salient in full fidelity, while providing access to less salient regions in lower fidelity. We will store the field data as wavelet transformed coefficients and vary the fidelity by the collection of coefficients that are retained. By judiciously saving heterogeneous blocks of wavelet coefficients one can perform various meaningful and contextualized interactive analyzes and visualizations of the data with dramatically reduced I/O and storage requirements. We intend to demonstrate the efficacy of this approach through its application to the large-scale wind-turbine array simulations being developed by the NWTC and the Computational Science Center. These simulations are well suited to this approach, since there are clear and inherent opportunities for multi-resolution analysis in these data. Specifically, understanding the impact of large turbulent structures forming in the atmospheric boundary layer (low-fidelity context) on the wake formation and structural loading of the turbine (full-fidelity detail).

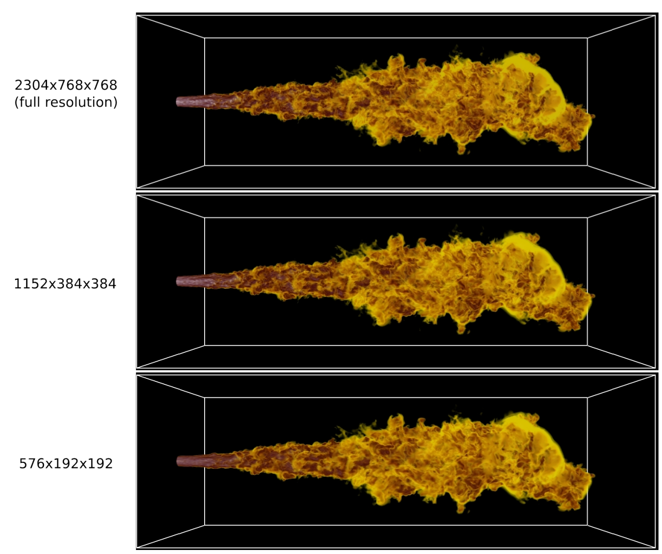

Figure 1: Three volume renderings of velocity magnitude from a turbulent atomizing jet simulation run on RedMesa. Three coarsenings are shown, note how even the small-scale structures

We have shown in previous work [7] that a multi-resolution access scheme based on storing a hierarchy of wavelet coefficients can effectively manage the analysis of terascale-sized data sets from a mass store. Each level of the wavelet coefficient hierarchy coarsens the spatial resolution in half. For a three-dimensional data set that provides an eight-fold reduction in the data volume and corresponding reduction in the demands of the visualization and analysis. The hierarchical nature of this scheme allows investigators to interactively refine the analysis by controlling the coarseness of the data. Many visualization and analysis operations are relatively insensitive to this type of data coarsening (e.g., see Figure 1) [11]. In addition, we have shown [8] that certain salient regions in turbulent flows can be isolated with a combination of topological and statistical techniques.

In this prior work, the visualization and analysis has always been applied to homogeneously coarsened data, and the original data has always been available without loss of information by retrieving all the coefficients. This project would build upon this prior work by allowing the storage and reconstruction of data regions comprised of blocks of heterogeneous wavelet levels, and storing that heterogeneous data in parallel directly from the wind-array simulation code. We will perform an initial validation of our results against complete full fidelity data to understand and quantify the error introduced by this approach, in the context of investigating the turbine wake and vortex dynamics.

We will also do some preliminary work to lay the foundation for an ASCR proposal, specifically, establishing a research plan for the dynamic extraction of regions-of-interest and for developing a comprehensive uncertainty model. Developing an *in situ* framework to dynamically identify salient regions of the data volume based on physical, topological, and statistical signatures would provide a great contribution to the wider computational science community, but will require a substantial effort. Likewise, a comprehensive uncertainty model that could bound the error introduced in the low-fidelity regions will greatly advance these techniques by providing measured level of confidence in the analysis, but will require a level of effort beyond the scope of this LDRD. Therefore, we plan to use these angles of attack to form the basis for follow-on funding proposal.

Measures of successful outcome

We aim to enable visualization and analysis on the largest-scale wind simulations being developed and run at NREL, and lay a foundation for follow-on funding establishing a research program to generalize the approach. Specific goals and measures:

- Store heterogeneous multi-resolution data in parallel directly from the wind simulation code, enabling three-dimensional visualization and analysis of very large-scale wind-turbine array simulations.
- Two peer-reviewed publications
- Submission of at least one proposal for direct funding from DOE ASCR or other sources.

Intellectual property and patent considerations

There are no intellectual property or patent considerations. All code and method development will be accomplished in an open-source framework.

Project management

Period of performance: October 1, 2012 through September 30, 2014

- Kenny Gruchalla (PI) is a Senior Scientist in the Computational Science Center (CSC). He has a PhD in computer science and over 17 years of professional experience in scientific visualization and scientific applications development. He will coordinate the project and work specifically on technical visualization and computational issues.
- Ryan Elmore is a Senior Scientist in the CSC. Dr. Elmore, a computational statistician, will work on the error and uncertainty quantification elements of this project.
- Matt Churchfield is a Research Engineer in the National Wind Technology Center (NWTC). Dr. Churchfield is computational fluid dynamics expert and an advanced OpenFOAM user and programmer. Matt will serve two roles. First, he will assist instrumenting the NWTC OpenFOAM wind-array research codes with the *in-situ* code described in the approach section. Second, he will drive and verify the domain-science analysis.
- Pat Moriarty is a Senior Engineer in the NWTC. Dr. Moriarty is currently leading the wind farm array simulation efforts at NREL. He will serve as the scientific advisor for the turbine-level research issues.
- John Clyne is a Computational Scientist at the National Center for Atmospheric Research (NCAR). John is the author of the open-source wavelet-based visualization suite, VAPOR, that we intend to utilize and extend for this research. John is funded through NCAR to support VAPOR, and he will act as an unfunded

collaborator. John's inclusion will lay the foundation for a joint NREL-NCAR proposal to DOE ASCR, as the program manager (Lucy Nowell) has indicated that larger multi-institution proposals are preferred.
- A graduate student intern will be hired to assist in the low-level programming and testing.

Staff Member	Title	Organization	Phone Number	FTE Months
Kenny Gruchalla	Senior Scientist	CSC (2C00)	x3713	10
Ryan Elmore	Senior Scientist	CSC (2C00)	x3191	4
Matt Churchfield	Research Engineer	NWTC (5000)	x7080	4
Patrick Moriarty	Senior Engineer	NWTC (5000)	x7081	1
TBD	Graduate-Student Intern	CSC (2C00)	TBD	N/A
John Clyne	Computational Scientist	NCAR	303-497-1236	N/A

Table 2: Project management data; project duration is 24 months

Cost

TBD

Milestones and dates

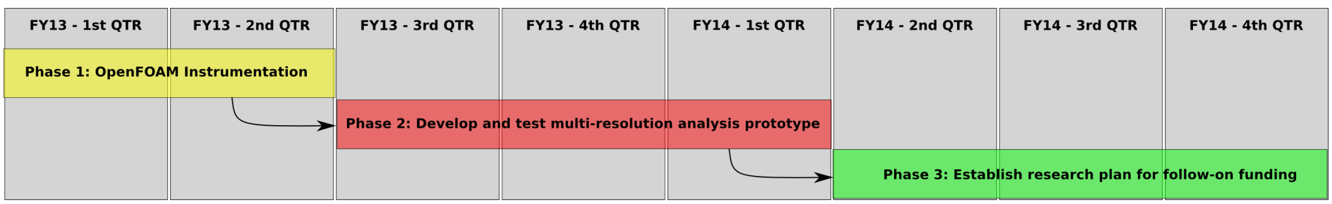

Figure 2: Project schedule

Phase 1: Store multi-resolution data in parallel directly from the wind simulation code.

- December 2012: Complete development of a C++ MPI VDF writer that can be called directly from the parallel OpenFOAM code
- March 2013: Complete the integration of the parallel writer with the NWTC wind-array OpenFOAM code.

Phase 2: Allow the storage and reconstruction of data regions comprised of blocks of heterogeneous wavelet levels.

- June 2013: Complete the development and testing of the ability to store heterogeneous blocks based on a static regions-of-interest geometry.
- September 2013: Complete the validation of the results against full-fidelity data providing a quantitative error measurement.
- March 2014: Submit first publication (top-tier journal, e.g., IEEE Transactions on Visualization and Computer Graphics)

Phase 3: Establish a research agenda for sizable ASCR proposal

- June 2014: Perform initial investigation and outline research plan to develop structural and statistical models to dynamically identify regions-of-interest *in situ*
- June 2014: Perform initial investigation and outline research plan to develop a comprehensive uncertainty model
- September 2014: Isolate and submit general code elements to VAPOR steering committee for inclusion in official VAPOR release.
- September 2014: Submit second publication and Final Report
- September 2014: Submit proposal to ASCR or other funding agency.

References

[1] Ahern, S., et al., 2011. Scientific Discovery at the Exascale. Technical Report. DOE ASCR 2011 Workshop on Exascale Data Management, Analysis, and Visualization.

[2] U.S. Department of Energy Office of Science, 2011. Simulation and Modeling at the Exascale for Energy, Ecological Sustainability and Global Security: An Initiative (DRAFT).

[3] ASCAC Subcommittee on Exascale Computing, 2010. The Opportunities and Challenges of Exascale Computing. Summary Report of the Advanced Scientific Computing Advisory Committee (ASCAC).

[4] Kogge, P., et al., 2008. ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems. Technical Report DARPA.

[5] Keim, D., Ward, M., 2000. Intelligent Data Analysis: An Introduction, 2nd edn. Springer, Heidelberg.

[6] Ross. R., et al. 2005. HPC File Systems and Scalable I/O: Suggested Research and Development Topics for the Fiscal 2005-2009 Time Frame, Technical Report. Los Alamos National Laboratory.

[7] Gruchalla, K., et al., 2009. Visualization-Driven Structural and Statistical Analysis of Turbulent Flows, *Lecture Notes in Computer Science*, **5772**, 321-332.

[8] Gruchalla, K., et al., 2011. Segmentation and Visualization of Multivariate Features using Feature-Local Distributions., *Lecture Notes in Computer Science*, **6938**, 619-628.

[9] Gruchalla, K. 2009. Progressive Visualization-Driven Multivariate Feature Definition and Analysis. PhD Thesis. University of Colorado at Boulder.

[10] Clyne, J., Gruchalla, K., Rast, M., 2010. VAPOR: Visual, Statistical, and Structural Analysis of Astrophysical Flows. In Proceedings of Numerical Modeling of Space Plasma Flows. Astronomical Society of the Pacific Conference Series, **429**, 323-329.

[11] Clyne, J., et al., 2005. A prototype discovery environment for analyzing and visualizing terascale turbulent fluid flow simulations. In Proceedings of Visualization and Data Analysis 2005, 284-294.