

In-Situ Data Transformations to Enable Exascale Analysis

Peer-Timo Bremer
Lawrence Livermore National Laboratory
University of Utah

Exascale computing promises unprecedented scientific breakthroughs assuming we are able to effectively use the next generation of computer hardware. In this context, we believe significant investments are required to provide the comprehensive and reliable data analysis necessary for a productive scientific workflow. In particular, we contend that we must move from the current post-processing environment and beyond in-situ analysis to novel in-situ data transformations that combine efficient online computation with flexible and unbiased offline data analysis and exploration. The primary challenge for existing data analysis techniques is data movement especially file I/O. Currently, the overwhelming majority of analysis is performed in post-processing using data saved during a simulation. However, as the spatial and temporal resolution of scientific codes increases and we simulate ever more complex phenomena, it is already becoming difficult to store sufficient information frequently enough to adequately describe the results. In practice, relatively fewer timesteps are stored each generation of simulations and already some applications are losing the ability to reliably track fast moving features [5, 23, 17]. At exascale, it is generally accepted that this approach will be unsustainable as too few snapshots of a process will be permanently stored to allow a dependable analysis. Instead, data analysis will have to be performed *in-situ* as the simulation is running.

This will be a formidable challenge as it is the first time many data analysis techniques will be applied at any scale beyond some tens or hundreds of cores of an analysis cluster. Additionally, the main simulation will likely impose strict limits on the available memory, the acceptable execution time, and the data layout. These challenges have been recognized, and a number of efforts to develop in-situ analysis capabilities have been started [1, 7, 15, 3]. Unfortunately, even if successful, these efforts will not solve the underlying problem. In fact, concentrating entirely on faster, more scalable versions of existing analysis approaches, while necessary and useful, may ultimately become a crutch with the potential to significantly delay the scientific process. The problem is that any analysis approach that decides which questions to ask and how to answer them before a simulation has started will inherently be limited by our current knowledge. For example, it is feasible and in fact highly anticipated that advancing to exascale computing will produce fundamentally new insights resulting in never before seen structures, processes, or events. Yet, a pre-set analysis is limited to finding expected or at least anticipated phenomena. Therefore, a breakthrough may well remain undetected since, by definition, it does not conform to the existing concepts that drove the analysis setup. Instead, we believe a long term, sustained research effort is required to move beyond traditional analysis approaches in order to re-enable the exploratory, post-process analysis crucial to the scientific process. In particular, we advocate to focus on *in-situ data transformations* that can drastically reduce the amount of data without effecting the results of subsequent analyses.

A well known yet often disregarded fact is that virtually all data analysis beyond simple statistics is currently performed in an iterative manner. Whether it is to semi-automatically pre-process data to remove noise or artifacts; to choose interesting subsets of the data to make an analysis approach tractable or highlight an effect; or to fine tune any number of parameters, rarely will the first attempt at an analysis produce the most useful (or any) results. However, in an in-situ setting no such iterations are possible, and instead the entire analysis pipeline must be defined before any data is produced. While one can imagine self-adapting algorithms and semi-automatic steering techniques, designing a single-pass analysis pipeline even for a new variant of a well known problem is a daunting task. To assume that, for a new mathematical model that represents previously unattainable physics, and simulating phenomena never before seen, it would be possible to setup, a priori, a comprehensive and fully automatic analysis pipeline producing all sought after insights seems unrealistic.

One potential approach would be to simply use existing algorithms, suitably adapted to in-situ and exascale, to repeatedly analyze the data using different thresholds and setups. However, to be effective and reliable, the range of parameters would have to be so large and the set of potentially interesting aspects so broad as to make the analysis infeasible. Another alternative might be traditional compression techniques. However, the compression ratios required to enable exascale analysis as envisioned here would likely result in a severe loss of information and potentially significant errors in the analysis. Instead, new data transformations should be developed that perfectly preserve the information required for a particular analysis while sacrificing most everything else. In this framework, rather than specifying the exact analysis to apply, a scientist would decide what aspects of what data, ie., temperatures, velocities, material concentrations, etc., are likely to be of most interest. For example, for indicator functions such as λ_2 in vortex core detection [14], the connected sets of high values, ie., the super-level sets, are of interest. Other features such as dissipation elements [18] are best described by the gradient flow and yet others may be encoded by ridges [21] or clusters [6]. Instead of aiming to extract a particular subset of such features in-situ, the data should be transformed into representations that encode all of them leaving the selection and exploration for post-processing. This will require an in-depth, theoretical understanding of the global arrangement of all potential features, algorithms to compute them, as well as data structures to encode them efficiently.

For a very restricted set of features, such representations exist and have been used to great effect. In particular, a topological structure called a merge tree encodes the number of all super-level sets [5] and, with some additional work, their location, shape, and integral properties [4]. These super-level sets describe features such as bubbles in a Raleigh-Taylor instability [16] or burning regions in a flame [5] in a parameter-independent manner. Most importantly, a merge tree is typically several orders of magnitude smaller than the original data, yet it encodes the identical information about threshold-based features. Similar capabilities are being considered in visualization, for example, by using intermediate representations for in-situ volume rendering to create explorable images [22] rather than static snapshots.

However, merge trees and explorable images cover only a small portion of the types of analysis and visualization that will be required at exascale. For gradient based features such as material core lines [10], the Morse-Smale complex [9] seems to be the appropriate representation. However, parallelizing even the basic algorithm has proven challenging [11], and it is unclear how to encode the resulting structure efficiently as it can easily become larger than the source data. For ridges, some theoretical work on their structure exists [19], yet current algorithms to compute even the most pronounced subsets are notoriously sensitive [20]. For other popular techniques such as Lagrangian Coherent Structures [13, 12], the fundamental structure is still under development and techniques to compute them [8, 2] are expensive enough to prevent a thorough exploration even in the current post-processing framework.

Replacing a direct analysis of these features by the corresponding data transformations will allow scientists to explore their results without the bias of a pre-determined choice of parameter and enable them to form and test new hypotheses as needed. Without such capabilities, it will be difficult to realize the full benefits of exascale computing as simulations will be run multiple times to adjust the analysis pipeline, unexpected results will be missed or obscured, and extraordinary pressures will be placed on the file systems in hopes of mitigating the problems. However, success in this area requires a concerted and long term effort in a number of areas in math and computer science including statistics, computational geometry, topology, data analysis, parallel computing, and systems research. Furthermore, given past experiences, the theoretical aspects of the research such as the fundamental understanding of the various feature spaces will require a multiple year head start. Therefore, strategic investments to foster such research are urgently needed to avoid situations in which ground breaking science results are achieved but not recognized for the lack of analytic capabilities.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-PROP-635912

References

- [1] Favre B. Whitlock, J.-M and J.S. Meredith. Parallel In Situ Coupling of Simulation with a Fully Featured Visualization System. In *Proc. of 11th Eurographics Symposium on Parallel Graphics and Visualization (EGPGV'11)*, 2011. i
- [2] S. Bakarat, C. Garth, and X. Tricoche. Interactive computation and rendering of finite-time Lyapunov exponent fields. *IEEE Trans. Vis. Comp. Graph.*, 18(8):1368–1380, 2012. ii
- [3] J. Bennett, H. Abbasi, P.-T. Bremer, R. Grout, A. Gyulassy, T. Jin, S. Klasky, H. Kolla, M. Parashar, V. Pascucci, P. Pebay, D. Thompson, H. Yu, F. Zhang, and J. Chen. Combining in-situ and in-transit processing to enable extreme-scale scientific analysis. In *Proc. ACM/IEEE Conference on Supercomputing (SC12)*, 2012. i
- [4] J. Bennett, V. Krishnamurthy, S. Liu, V. Pascucci, R. Grout, J. Chen, and P.-T. Bremer. Feature-based statistical analysis of combustion simulation data. *IEEE Trans. Vis. Comp. Graph.*, 17(12):1822–1831, 2011. ii
- [5] P.-T. Bremer, G. Weber, J. Tierny, V. Pascucci, M. Day, and J. B. Bell. Interactive exploration and analysis of large scale simulations using topology-based data segmentation. *IEEE Trans. on Visualization and Computer Graphics*, 17(9):1307–1324, 2011. i, ii
- [6] M. Davis, G. Efstathiou, C.S. Frenk, and S. White. The Evolution of Large Scale Structure in a Universe Dominated by Cold Dark Matter. *Astrophys.J.*, 292:371–394, 1985. ii
- [7] N. Fabian, K. Moreland, D. Thompson, A.C. Bauer, P. Marion, B. Gevecik, M. Rasquin, and K.E. Jansen. The paraview coprocessing library: A scalable, general purpose in situ visualization library. In *Proc. of IEEE Symposium on Large Data Analysis and Visualization (LDAV)*, pages 89–96, 2011. i
- [8] C. Garth, A. Wiebel, X. Tricoche, K. Joy, and G. Scheuermann. Lagrangian visualization of flow-embedded surface structures. *Computer Graphics Forum*, 27(3):1007–1014, 2008. ii
- [9] A. Gyulassy, P.-T. Bremer, V. Pascucci, and B. Hamann. A practical approach to Morse-Smale complex computation: Scalability and generality. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1619–1626, 2008. ii
- [10] A. Gyulassy, M. Duchaineau, V. Natarajan, V. Pascucci, E. Bringa, A. Higginbotham, and B. Hamann. Topologically clean distance fields. *IEEE Transactions on Computer Graphics and Visualization (TVCG)*, 13(6):1432–1439, 2007. ii
- [11] A. Gyulassy, T. Peterka, V. Pascucci, and R. Ross. Characterizing the parallel computation of morse-smale complexes. In *Proceedings of IPDPS '12*, Shanghai, China, 2012. ii
- [12] G. Haller. A variational theory of hyperbolic lagrangian coherent structures. *Physica D: Nonlinear Phenomena*, 240(7):574–598, 2011. ii
- [13] G. Haller and G. Yuan. Lagrangian coherent structures and mixing in two-dimensional turbulence. *Phys. D*, 147(3-4):352–370, 2000. ii
- [14] J. Jeong and F. Hussain. On the identification of a vortex. *J. Fluid Mech.*, 285(-1):69–94, 1995. ii

- [15] S. Lakshminarasimhan, J. Jenkins, I. Arkatkar, Z. Gong, H. Kolla, S.-H. Ku, S. Ethier, J. Chen, C.S. Chang, S. Klasky, R. Latham, R. Ross, and N.F. Samatova. Isabela-qa: Query-driven analytics with isabela-compressed extreme-scale scientific data. In *Proc. of the ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, pages 1–11, 2011. i
- [16] D. Laney, P.-T. Bremer, A. Mascarenhas, P. Miller, and V. Pascucci. Understanding the structure of the turbulent mixing layer in hydrodynamic instabilities. *IEEE Trans. Visualization and Computer Graphics (TVCG) / Proc. of IEEE Visualization*, 12(5):1052–1060, 2006. ii
- [17] A. Mascarenhas, R. W. Grout, P.-T. Bremer, E. R. Hawkes, V. Pascucci, and J.H. Chen. *Topological feature extraction for comparison of terascale combustion simulation data*, pages 229–240. Mathematics and Visualization. Springer, 2011. i
- [18] J.P. Mellado, L. Wang, and N. Peters. Gradient trajectory analysis of a scalar field with external intermittency. *Journal of Fluid Mechanics*, 626:333–365, 4 2009. ii
- [19] J. Miller. *Relative Critical Sets in R^n and Applications to Image Analysis*. PhD thesis, University of North Carolina, 2006. ii
- [20] R. Peikert and P. Sadlo. Height ridge computation and filtering for visualization. In *Proc. PacificVis 2008*, pages 119–126, 2009. ii
- [21] Shawn C. Shadden, Francois Lekien, and Jerrold E. Marsden. Definition and properties of lagrangian coherent structures from finite-time lyapunov exponents in two-dimensional aperiodic flows. *Physica D: Nonlinear Phenomena*, 212(3-4):271 – 304, 2005. ii
- [22] A. Tikhonova, H. Yu, C.D. Correa, J.H. Chen, and K.-L. Ma. A preview and exploratory technique for large-scale scientific simulations. In *Proceedings of Eurographics Parallel Graphics and Visualization Symposium (EGPGV)*, April 2011. ii
- [23] G. Weber, P.-T. Bremer, J. Bell, M. Day, and V. Pascucci. *Feature Tracking Using Reeb graphs*, pages 241–253. Mathematics and Visualization. Springer, 2011. i