High-level software frameworks to surmount the challenge of 100x scaling for biomolecular simulation science

Shantenu Jha[*1] and Peter M. Kasson[+2]
[*] *Electrical and Computer Engineering, Rutgers University*
[+] *Molecular Physiology and Biological Physics and Biomedical Engineering, University of Virginia*

Abstract

Next-generation exascale systems will fundamentally expand the reach of biomolecular simulations and the resulting scientific insight, enabling the simulation of larger biological systems (weak scaling), longer timescales (strong scaling), more complex molecular interactions, and robust uncertainty quantification (more accurate sampling). Since currently envisioned exascale hardware architectures are essentially larger versions of systems available today, it will be challenging to solve biological problems that require longer timescales, involve more complex interactions and robust uncertainty quantification without significant algorithmic improvements. We believe that high-level simulation algorithms incorporating high-level parallelism and leveraging the statistical nature of molecular processes can provide a means to address these challenges of scaling. Proof-of-concept simulation algorithms have yielded advanced sampling and adaptive control algorithms for efficient simulation of long timescales and complex behaviors. Novel dataflow and workflow systems are needed to implement these advanced algorithms in a way that is usable by the community in exascale systems. A middleware ecosystem that provides these in a robust, scalable, reusable, and extensible framework is a key requirement for exascale infrastructure investment to result in revolutionary biological insight.

In the past decade, substantial algorithmic and hardware advances have led to improvements in strong and weak scaling that permit millisecond-length simulations of moderate-sized biomolecular systems and short simulations for large assemblies [1]. Both of these have enabled direct comparison with experimental observables in ways not previously possible. However, most software development has focused on optimizing single-simulation performance, while many chemical and biological problems require solving a higher-level statistical problem such as the stochastic behavior of large ensembles of molecules or the statistical physics of a few molecules over longer timescales. Higher-level methods for solving these statistical problems have been the focus of many recent advances [2] in molecular simulation, but broad adoption has been limited by the lack of software frameworks to perform these calculations in a manner that is flexible, scalable, and provides a low barrier to entry.

[1] shantenu.jha@rutgers.edu
[2] kasson@virginia.edu

A 100x increase in computational capacity must be accompanied by a concomitant change in the sophistication and type of molecular simulations. A fundamental need is to support higher-level formulations that go beyond simply performing individual, isolated simulations faster. This is necessary to tackle biological and chemical problems that have thus far eluded accurate simulation. The ecosystem of cyberinfrastructure for molecular simulations must evolve to support simulation modes, which include but are not limited to:

- 100x Replicas/Ensembles: For long the supercomputing ecosystem has been tuned towards the effective execution of single large jobs (weak scaling). As important as this is, there is a need to support the requirements of molecular science applications that critically depend upon the effective execution of multiple simulations, either concurrently or in some coupled form. Advanced Sampling techniques such as Replica-Exchange and Transition Path Sampling are two prominent examples that require first-class support for many replicas/ensembles. The ability to support O(100,000) replicas, where each replica would itself be a parallel simulation, will be required to explore multi-dimensional ensembles such as protein mutation spaces or the binding of different small molecules to a drug target.
- Adaptive and Computationally Steered Workflows: Adaptive workflows can be thought of as internally steered, where real-time analyses are used to determine the next step in the workflow. Diffusion-map driven molecular dynamics and methods that determine collective variables "on the fly" require adaptive workflows [3]. Computational Steering [4] is the ability to dynamically change the course of a computational process in response to external stimuli e.g., human interaction or an external computational process that might be a consumer or producer of input/output data. Streaming-data steered simulations to integrate experiments, simulations, and analysis will be needed to take full advantage of experimental facilities such as high energy light sources [5]. Frameworks and libraries that will support the coupling of multi-physics high-performance simulations and hybrid computational processes will be required.

There is a need for a domain-optimized framework for expressing and implementing such high-level simulation strategies. Furthermore, scalable analysis capabilities are needed in conjunction with higher-level control algorithms and scalable simulations. Trajectory dataset sizes for large molecular systems are already stressing I/O subsystems; this trend will only increase. For example, critical experimental observables occur at widely divergent timescales (over million-fold different), and the cost of storing all data at the finest time resolution exceeds the cost of computation. This necessitates multi-scale and multi-modal analysis, viz., the ability to analyze data at different levels of granularity and resolution. Traditionally, the analysis of data was performed "offline", however, with large volumes of data generated as well as increasing sophistication, there is a need to support "online" or real-time analysis in conjunction with the generation of data. This results in a set of requirements ranging from "in-situ" analysis to advanced analytics platforms as supported by environments such as Spark, Flink etc. Additional important challenges include joint analysis of simulation and large-scale experimental data.

Recommendations: Progress will be stymied by the lack of a scalable, flexible, domain-optimized framework for expressing and implementing high-level simulation strategies and the right middleware and software ecosystem. Thus, functionally there is a need for:

- Workflow expressivity that can capture not just one "evolutionary" pathway but which supports statistical descriptions for multiple evolutionary pathways and adaptive control between them.
- Workflow platforms that provide a robust implementation of this rich expressivity are critically needed.
- Expressing simulation as a workflow can also provide reusable components that promote dissemination of new scientific methods and boost reproducibility.
- A middleware ecosystem that can support:
  - Efficient execution and management of large number of concurrent tasks.
  - Support for a range of adaptive execution and computational steering modes.
  - Inclusion of streaming data into the high-performance simulation modes.

Acknowledgements: We thank Michael Shirts (Colorado) and Thomas Cheatham (Utah) for early and helpful discussions about ensemble based methods.

References

[1] (i) Bowers, K. J., Dror, R. O., & Shaw, D. E. (2006). The midpoint method for parallelization of particle simulations. The Journal of chemical physics, 124(18), 184109., (ii) Shaw, D. E., Deneroff, M. M., Dror, R. O., Kuskin, J. S., Larson, R. H., Salmon, J. K., ... & Wang, S. C. (2008). Anton, a special-purpose machine for molecular dynamics simulation. Communications of the ACM, 51(7), 91-97., and (iii) Elsen, E., Houston, M., Vishal, V., Darve, E., Hanrahan, P., & Pande, V. (2006, November). N-Body simulation on GPUs. In Proceedings of the 2006 ACM/IEEE conference on Supercomputing (p. 188). ACM., and (iv) Zhao, G., Perilla, J. R., Yufenyuy, E. L., Meng, X., Chen, B., Ning, J., ... & Zhang, P. (2013). Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. Nature, 497(7451), 643-646.

[2] (i) Singhal, N., & Pande, V. S. (2005). Error analysis and efficient sampling in Markovian state models for molecular dynamics. The Journal of chemical physics, 123(20), 204909; (ii) Pronk, S., Larsson, P., Pouya, I., Bowman, G. R., Haque, I. S., Beauchamp, K., ... & Lindahl, E. (2011, November). Copernicus: A new paradigm for parallel adaptive molecular dynamics. In Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis (p. 60). ACM; (iii)NoÃľ, F., SchÃijtte, C., Vanden-Eijnden, E., Reich, L., & Weikl, T. R. (2009). Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. Proceedings of the National Academy of Sciences, 106(45), 19011-19016.

[3] (i) J. Preto, C. Clementi (2014). Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics. Physical Chemistry Chemical Physics, 16, 19181 (2014), (ii) C. A Laughton, M. Orozco and W. Vranken (2008). COCO: A simple tool to enrich the representation of conformational variability in NMR structures. Proteins 2009; 75:206-216.

[4] Computational Steering: `https://en.wikipedia.org/wiki/Computational_steering`.

[5] Several recent NSF/DOE workshops have examined the cross-section of computational requirements of experimental facilities: (i) http://streamingsystems.org (ii) http://extremescaleresearch.labworks.org/events/data-management-visualization-and-analysis-experimental-and-observational-data-eod-workshop (iii) http://s2i2.caltech.edu/main/, and (iv) http://idies.jhu.edu/symposium/cyberinfrastructure-ci-for-nsf-large-facilities-workshop/.