Distributed Workflows for Modeling Experimental Data

Vickie E. Lynch
Neutron Data Analysis & Visualization
Oak Ridge National Laboratory
Oak Ridge, TN
lynchve@ornl.gov

Jose Borreguero Calvo
Neutron Data Analysis & Visualization
Oak Ridge National Laboratory
Oak Ridge, TN
borreguero@gmail.com

Ewa Deelman
Information Sciences Institute
University of Southern California
Marina del Rey, CA
deelman@isi.edu

Rafael Ferreira da Silva
Information Sciences Institute
University of Southern California
Marina del Rey, CA
rafsilva@isi.edu

Monojoy Goswami
Center for Nanophase Materials
Oak Ridge National Laboratory
Oak Ridge, TN
goswamim@ornl.gov

Yawei Hui
Computer Science & Mathematics
Oak Ridge National Laboratory
Oak Ridge, TN
huiy@ornl.gov

Eric Lingerfelt
Computer Science & Mathematics
Oak Ridge National Laboratory
Oak Ridge, TN
lingerfeltej@ornl.

Jeffrey S. Vetter
Computer Science & Mathematics
Oak Ridge National Laboratory
Oak Ridge, TN
vetter@ornl.gov

Abstract— Modeling helps explain the fundamental physics hidden behind experimental data. In the case of material modeling, running one simulation rarely results in output that reproduces the experimental data. Often one or more of the force field parameters are not precisely known and must be optimized for the output to match that of the experiment. Since the simulations require high performance computing (HPC) resources and there are usually many simulations to run, a workflow is very useful to prevent errors and assure that the simulations are identical except for the parameters that need to be varied. The use of HPC implies distributed workflows, but the optimization and steps to compare the simulation results and experimental data are done on a local workstation. We will present results from force field refinement of data collected at the Spallation Neutron Source using Kepler, Pegasus, and BEAM workflows and discuss what we have learned from using these workflows.

Keywords—workflows, modeling, simulations, experiments

I. Introduction

When data is collected by an experiment, there is always some reduction procedure required before the results of the experiment can be viewed. Then if a simulation is needed to interpret those results, a similar reduction procedure is needed for the simulation output data. Ideally the simulation would run coincidently with the experiment with the results of both compared in the same computer window. In that case, the experiment could be changed using predictions from the simulation and/or the simulation input could be refined by the experiment.

With so many steps in the comparison of results from experiments and HPC simulations needed for a scan of values of the force field parameters, workflows are needed for modeling experimental data[1]. At the Spallation Neutron Source, there are many experiments that would benefit from comparing results to simulated data with a workflow. For the first attempts, quasi-elastic neutron scattering (QENS) data from the inelastic instrument, BASIS, was compared to data from simulations. Kepler[2], Pegasus[3] and BEAM[4] workflows were used to model this data.

II. Kepler Example

Molecular dynamics simulations of a concentrated aqueous solution of LiCl[1] were refined against thermodynamics data collected by the BASIS instrument at the Spallation Neutron Source (SNS) using a Kepler[2] workflow. The result of this work was an optimized water-model dipole moment that reproduced the dynamics of the experimental ionic solution at standard conditions. This refinement was done using a framework for optimizing simulations to model experimental data. Remote submission of the parallel NAMD[5] and Sassena[6] simulations and the reduction of the simulation results to calculate residuals with the experimental data were done using a Kepler workflow. Three parallel simulations were submitted at the same time to calculate local derivatives

which Dakota's[7] least squares optimization algorithm used to choose the next model parameters. The Dakota toolkit optimized the simulation and beamline-model scaling parameters using these residuals. This work was done to demonstrate a software suite that will enable simulation and modeling to be brought directly into the data analysis loop of experimental data taken at the SNS. The optimization process uses the Kepler GUI for input and the workflow was constructed graphically (Fig. 1) using Kepler components.

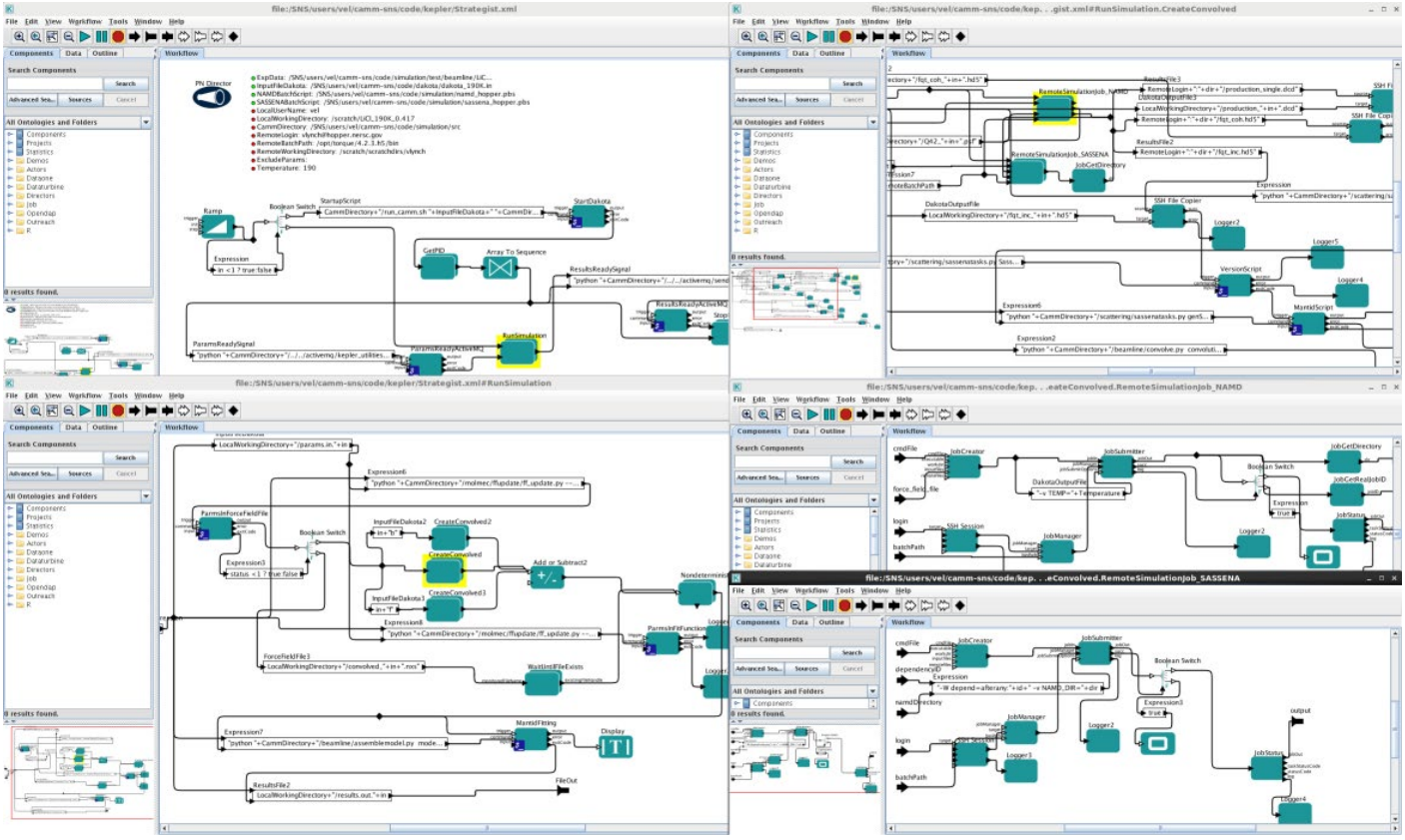

Fig. 1. Kepler actors in layers of components. Yellow actors are currently active.

Simulations are finite in time and length, leading to errors in the calculation of correlation functions such as the structure factor, S(**Q**,E). These errors lead to spurious minima (Fig. 2) in the goodness of fit when comparing to experimental QENS data. These local minima were found by the least squares optimization results from the Kepler workflow. An Interpolator plus smoothing procedure[8] is required to calculate structure factors that are smooth and derivable in the force field parameter that is sought to refine against QENS data. This optimization was redone by a manual submission of simulations varying the force field parameter which demonstrated the need for smoothing before optimization and the need for a workflow to manage the submission of the simulations. This parameter scan was later redone to demonstrate that the Pegasus workflow[3] was implemented correctly.

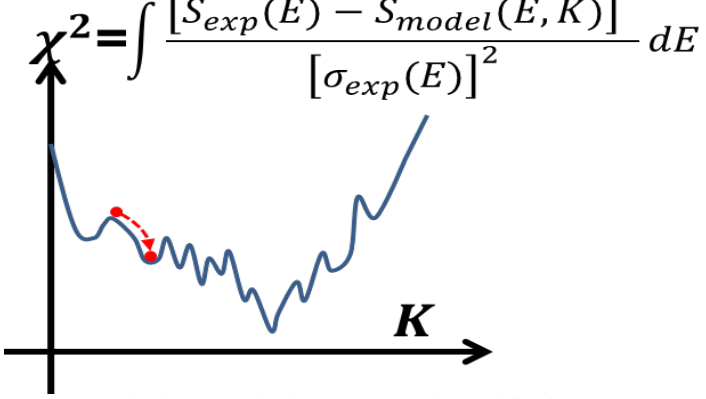

Fig. 2. Example of spurious local minima in goodness of fit when comparing quasi-elastic data to simulated data.

III. PEGASUS EXAMPLE

Our case that most illustrates the need for HPC simulations was exemplified in determining the accurate tRNA (transfer-RNA) Dynamics on Diamond Nanoparticles (Nanodiamond or ND)[9] which required 400,000 CPU hours of time on a Cray XE6. Simulations of tRNA and hydrophilic nanodiamonds in a deuterated water (D_2O) environment at 300K, 290K, 280K and 260K were done to understand the data obtained from quasi-elastic neutron scattering experiments at SNS. Due to the lack of availability of the proper force-field parameter, the simulations were done by assuming the interaction between oxygen of water and nanodiamond occurs via the Lennard-Jones (LJ) potential. The force field parameter between the hydrophilic nanodiamond and water, ε, (LJ ε) was optimized by comparing simulated and QENS scattering data. This ND-tRNA complex (Fig. 3) is a potential delivery method of foreign RNA into target cells.

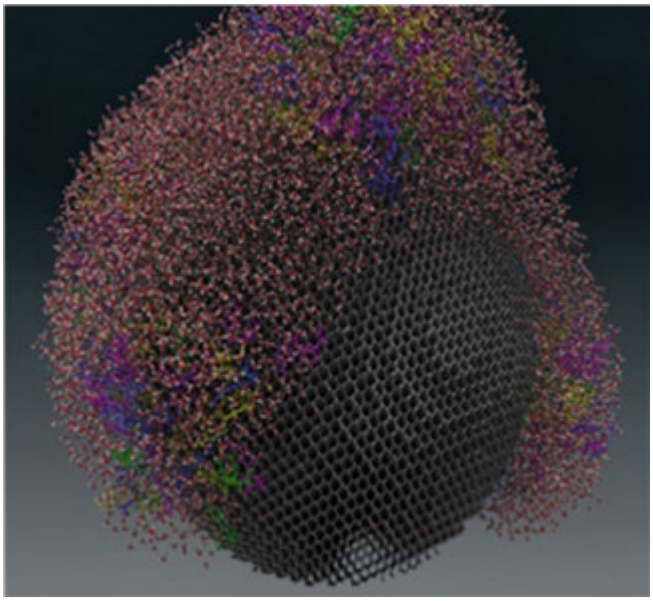

Fig. 3. Water is seen as small red and white molecules on large nanodiamond sphere. The colored tRNA can be seen on the nanodiamond surface.

The focus of Pegasus[3] workflows is data-aware workflow performance modeling, monitoring, and analysis of modeling that uses high performance computing resources. Pegasus submitted instances of this workflow varying ε in the NAMD and Sassena calculations, thus changing the affinity of RNA to the diamond nanoparticles. The software used for this optimization was Mantid's[10] algorithm for cubic spline interpolation of dynamics structure factors[8]. The optimal value found for all Q's (Fig. 4) was ε_{opt} = -0.01 Kcal/mol a 26% change from the non-optimized starting value ε_0 =-0.13 Kcal/mol.

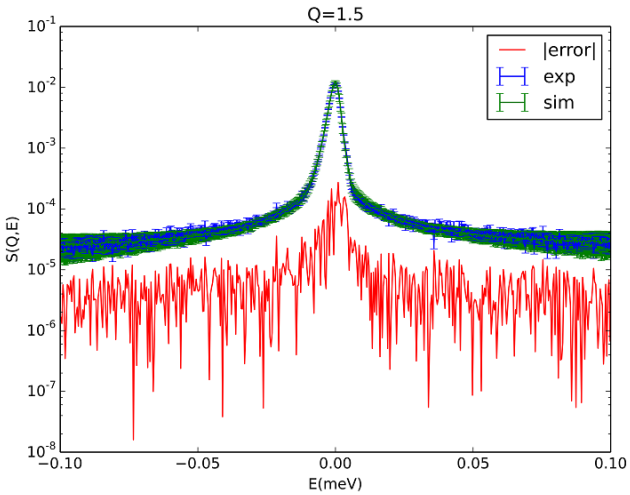

Fig. 4. Comparison of experimental and optimized simulation data for Q=1.5 $Å^{-1}$.

IV. BEAM EXAMPLE

The Bellerophon Environment for the Analysis of Materials (BEAM)[4] utilizes OLCF and CADES compute resources to automate workflows for parameter-refinement of force-fields used in molecular dynamics simulations by iterative optimization against QENS data. The workflow was tested on a full atom representation of the mPOSS molecule to refine the potential energy barrier to methyl rotations in octa-methyl silsesquioxane obtaining an optimal value for the activation energy. BEAM ran the NAMD and Sassena calculations, modified the force field parameter to the optimal value for the experimental data, and visualized the results with an interactive multidimensional data view of both the experimental data and optimal simulation results (Fig. 5).

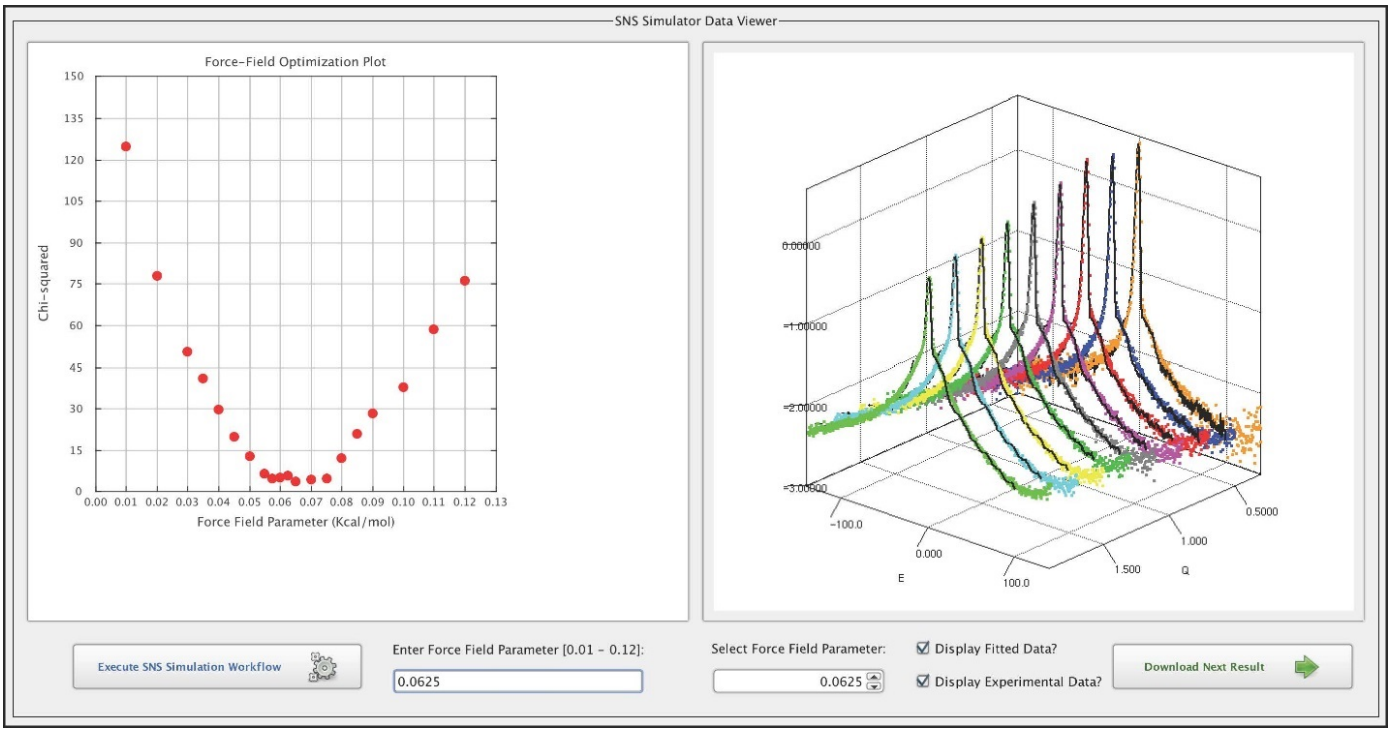

Fig. 5. BEAM display of the goodness of fit of the mPOSS molecule as the potential energy barrier to methyl rotations was changed.

V. CONCLUSIONS

The Kepler workflow used an optimization method that found local minima. After discovering that smoothing was required, the later Pegasus and BEAM workflows used Mantid[10] for smoothing and finding optimal force field parameter from an equally spaced parameter scan. We found that this smoothing is necessary to avoid local minima.

Workflow creation was simplest for the Kepler workflow where needed components were connected graphically. With Pegasus, a python code with a configuration file was written to generate the input for the workflow. Currently, new workflow tasks must be added manually to BEAM's backend workflow engine and database located within the web services tier.

All three workflows had monitoring tools to display progress of jobs. With the Kepler workflow, the Java GUI had to stay open while the workflow was running which would be difficult for workflows that wait for days or weeks for long runs on HPC resources. Both Kepler and BEAM use passwords for remote HPC authentication and ssh and scp for launching jobs and transferring data. Pegasus uses a grid certificate and gram and condor for launching jobs and transferring data. For displaying output, BEAM is the only workflow with a GUI that display 3-D visualization of results (Fig. 5). Pegasus has no GUI and the Kepler GUI just shows the execution of the components in the workflow.

All the parameter scans tested here were done for optimizing only one parameter. Simple equally spaced parameter scans will not work for cases where many force field parameters must be optimized simultaneously. Using a global optimization algorithm from Dakota[7] may result in fewer required simulations for optimizing many parameters.

Communication and moving files between the HPC and the local workstation are often complicated by computer security requirements. If the ssh connection is not persistent, a password is required for every HPC submission. Also, many firewall exceptions may be required for the grid software needed for Pegasus.

ACKNOWLEDGMENTS

The development workflows were supported by the U.S. Department of Energy (DOE), Office of Science, Basic Energy Sciences (BES), Materials Sciences and Engineering Division. The use of Oak Ridge National Laboratory's (ORNL) SNS was sponsored by the Scientific User Facilities Division, Office of BES. We thank Gurpreet K. Dhindsa and Xiang-Qiang Chu from Wayne State University for providing us with the Nanodiamond QENS data. The Pegasus workflow was funded by the DOE under contract number DE-SC0012636. This manuscript has been authored by ORNL, which is managed by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. DOE. This product includes software developed by and/or derived from the Globus project (http://www.globus.org/). This research used resources of the National Energy Research Scientific Computing Center (NERSC), a DOE Office of Scientific User Facility supported by the Office of Science of the U.S. DOE under Contract No. DE-AC02-05CH11231. Some NAMD simulations were performed on TITAN at the Oak Ridge Leadership Computing Facility at the ORNL, which is supported by the Office of Science of

the U.S. DOE under Contract No. DE-AC05-00OR22725. Part of this research was conducted at the Center for Nanophase Materials Sciences (CNMS), which is a DOE Office of Science User Facility. BEAM was partially supported by the Laboratory Directed Research and Development (LDRD) program at ORNL (E.J.L., Y.H) and utilizes the resources at ORNL's Compute and Data Environment for Science (CADES), which is managed by UT-Battelle, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC05-00OR22725.

REFERENCES

1 Borreguero, J.M., Campbell, S.I., Delaire, O.A., Doucet, M., Goswami, M., Hagen, M.E., Lynch, V.E., Proffen, T.E., Ren, S., Savici, A.T., and Sumpter, B.G.: ' Integrating Advanced Materials Simulation Techniques into an Automated Data Analysis Workflow at the Spallation Neutron Source' (John Wiley & Sons, Inc., 2014)

2 Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludascher, B., and Mock, S.: 'Kepler: an extensible system for design and execution of scientific workflows', Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on, 2004, pp. 423--424

3 Deelman, E., Carothers, C., Mandal, A., Tierney, B., Vetter, J.S., Baldin, I., Castillo, C., Juve, G., Król, D., Lynch, V., Mayer, B., Meredith, J., Proffen, T., Ruth, P., and Ferreira da Silva, R.: 'PANORAMA: An approach to performance modeling and diagnosis of extreme-scale workflows', The International Journal of High Performance Computing Applications, 2017, 31, (1), pp. 4-18

4 Lingerfelt, E.J., Belianinov, A., Endeve, E., Ovchinnikov, O., Somnath, S., Borreguero, J.M., Grodowitz, N., Park, B., Archibald, R.K., Symons, C.T., Kalinin, S.V., Messer, O.E.B., Shankar, M., and Jesse, S.: 'BEAM: A Computational Workflow System for Managing and Modeling Material Characterization Data in HPC Environments', Procedia Computer Science, 2016, 80, pp. 2276-2280

5 Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kale, L., and Schulten, K.: 'Scalable molecular dynamics with NAMD', Journal of Computational Chemistry, 2005, 26, pp. 1781-1802

6 Lindner, B., and Smith, J.C.: 'Sassena — X-ray and neutron scattering calculated from molecular dynamics trajectories using massively parallel computers', Computer Physics Communications, 2012, 183, (7), pp. 1491-1501

7 Adams, B.M., Ebeida, M.S., Eldred, M.S., Jakeman, J.D., Swiler, L.P., Stephens, J.A., Vigil, D.M., Wildey, T.M., Bohnhoff, W.J., Dalbey, K.R., Eddy, J.P., Hu, K.T., Bauman, L.E., and Hough, P.D.: 'Dakota, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 6.0 Theory Manual', in Editor (Ed.)^(Eds.): 'Book Dakota, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 6.0 Theory Manual' (2014, edn.), pp.

8 Borreguero, J.M., and Lynch, V.E.: 'Molecular Dynamics Force-Field Refinement against Quasi-Elastic Neutron Scattering Data', J Chem Theory Comput, 2016, 12, (1), pp. 9-17

9 Lynch, V.E., Borreguero, J.M., Bhowmik, D., Ganesh, P., Sumpter, B.G., Proffen, T.E., and Goswami, M.: 'An automated analysis workflow for optimization of force-field parameters using neutron scattering data', Journal of Computational Physics, 2017, 340, pp. 128-137

10 Arnold, O., Bilheux, J.C., Borreguero, J.M., Buts, A., Campbell, S.I., Chapon, L., Doucet, M., Draper, N., Ferraz Leal, R., Gigg, M.A., Lynch, V.E., Markvardsen, A., Mikkelson, D.J., Mikkelson, R.L., Miller, R., Palmen, K., Parker, P., Passos, G., Perring, T.G., Peterson, P.F., Ren, S., Reuter, M.A., Savici, A.T., Taylor, J.W., Taylor, R.J., Tolchenov, R., Zhou, W., and Zikovsky, J.: 'Mantid—Data analysis and visualization package for neutron scattering and μSR experiments', Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 2014, 764, pp. 156-166