A vision for a validated distributed knowledge base of material behavior at extreme conditions using the Advanced Cyberinfrastructure Platform

James Ahrens and Christopher M. Biwer

White paper response to Big Data and Extreme Scale Computing, 2nd Series, (BDEC2)
Workshop 1: Bloomington, Indiana

As materials age or undergo extreme conditions (eg. shock), their physical properties can leave acceptable ranges which jeopardizes their safety and effectiveness in applications. Beamline facilities are a unique tool to probe how these materials properties (eg. strength and ductility) vary as a function of composition, time, stress, and other conditions [1,2]. This is critical knowledge to inform models predicting the performance of materials in applications. Due to the importance of these experiments in understanding materials, Europe, Japan, and the United States are commissioning X-Ray Free Electron Laster (XFEL) facilities with increased beam intensity, shot repetition rates, and detector sizes that enable new experiments as well as higher throughput rates for experiments [3,4,5]. Technological advances at synchrotron X-ray and the emerging X-ray Free Electron Laser (XFEL) facilities have realized data collection frequencies at 100 Hz, and MHz repetition rates are anticipated in the next few years [4].

Despite the significant advancements at facilities, open questions remain how data analysis methods will be applied given the data rates at these facilities, and in particular, whether the feedback from a real-time analysis can improve experimental outcomes at the facility, when an experiment is typically a few days. Requirements for local computer systems to triage the raw data and super computers to process the resulting information have been identified, similar to the envisioned ACP architecture.

The XFEL community has only begun to grapple with these data analysis challenges, and therefore, now is the opportune time to direct the cyberinfrastructure development. The BDEC call for papers identifies three approaches that the scientific community should use to address their cyberinfrastructure needs, and the following three sections describe unique requirements and ideas emerging from the XFEL community in the context of the three BDEC approaches.

Novel and/or converged models of inquiry: Predictive modeling of materials in applications relies on accurately parameterized models that describe the material's strength and plasticity (ie. tendency to deform). The parameterization and validation of the constitutive strength and plasticity models requires comparing simulations of the experiment and experimental results. Manual efforts to perform this parameterization of the strength and plasticity models using experimental data (which may be indirect measurements of the model parameters) often lead to non-unique parameterizations without uncertainty. Machine learning techniques such as Gaussian process modeling [6] provides a statistical framework to infer model parameter values and to quantify the uncertainty of each parameter. Gaussian process modeling uses an ensemble of hydrodynamic simulations to construct an emulator (ie. a surrogate model) that mimics the outputs of the computationally intensive hydrodynamics simulations. The emulator is calibrated with experimental data to infer the strength and plasticity model parameters which are inputs to the hydrodynamics simulation of the experiment.

The distribution in the parameter space and quantity of experimental data have a direct impact on the uncertainty of model parameters in Gaussian process modeling. Therefore, a strategy for experimental design can be to perform new experiments that minimize the uncertainty of the inferred parameters, which improves the parameterization of the strength and plasticity models. This approach to experimental design maximizes the science capability of beamline experiments which have a limited allotment of time to use the facility.

Support for advanced data logistics: Due to the time constraints an experiment has at these facilities, any analysis designed to provide real-time feedback, to direct experimental design, must be executed quickly.

However, Gaussian process modeling uses a Markov-chain Monte Carlo [7,8] which can require time equal to a significant portion of the experimental time at the beamline facility. One strategy for reducing the analysis time in Gaussian process modeling is pre-computing components of the analysis prior to arrival at the beamline facility. For example, the step which uses an ensemble of simulations to construct an emulator could potentially be computed beforehand. Aside from expediting the parameterization analyses, an emulator can predict the output of the hydrodynamics simulation for a given set of parameters in less than a second. Therefore, the emulator could be used at the beamline facility to stage and execute approximations of the simulations at a significantly reduced computational cost. Further reductions in the analysis time could be achieving using approximate algorithms that trade computational time at the cost of increasing the uncertainty of model parameters.

Support for interfaces to commercial cyberinfrastructure: The concept of optimizing of the experimental facilities resources can be extended to optimizing the computing resources. Since an experiment could benefit multiple research groups, or sites, the flexible cost model of commercial cyberinfrastructure could be shared amongst sites. Sites could choose to contribute to the cost of analysis in order to improve metrics such as time to completion, accuracy (ie. uncertainty reduction), data transfer, and resources costs that align with their particular research goals. The data and simulations are shared between all the sites, and therefore the sites collectively improve the quality of the analysis as well as reduce the overall number of computing resources required.

Vision: Competitive/cooperative science teams build their own repositories for raw data, simulation models, and models of previous experiments at experimental and supercomputing sites on the ACP. There is a multi-resolution transfer of raw simulation and experimental data between these sites, focused by the teams, on improving local repositories of selected materials in physical regions where modeled predictions differ from the real-world results. Each team collectively improves their data to increase their understanding. Exchanges both pull and push based on complex cost metrics that optimize what each site wants to achieve. Open science strategies could lower the cost of data based on its age, rewarding teams with privileged access and exchange capabilities for a fixed window of time.

[1] H. M. Rietveld. J. Appl. Cryst. **2**, 65-71 (1969).
[2] L. M. Barker and R. E. Hollenbach. J. Appl. Phys. **43**, 4669 (1972).
[3] U. Zastrau et al. REPORT-2017-004. XFEL.EU TR-2017-001 (2017).
[4] J. Thayer et al. Adv. Struct. Chem. Imaging **3**(1):3 (2017).
[5] S. Owada et al. J. Synchrotron Rad. **25**, 282-288 (2018).
[6] D. Higdon, J. Gattiker, B. Williams, and M. Rightley, J. Am. Stat. Assoc. **103**, 570 (2008).
[7] N. Metropolis et al. J. Chem. Phys. **21**, 1087–1092 (1953).
[8] W. K. Hastings. Biometrika **57**, 97–109 (1970).