Modeling analysis computations and end-to-end simulation-analysis workflows

P. Malakar1, G. Agrawal2, T. Bicer1, V. Vishwanath1, T. Munson1, R. Kettimuthu1, I. Foster1

1Argonne National Laboratory 2Ohio State University

Data analysis enables scientists to glean insight from their simulations and experiments. Performance modeling of analysis, simulation-analysis workflows and experiment-analysis workflows helps scientists plan their simulation and experiment campaigns. In this work, we consider modeling the analysis computations of scientific simulations and experimental data. We present performance models for analysis computations of large-scale simulations and experiments, as well as approaches to model the overall end-to-end simulation-analysis workflows.

We present three models to predict the performance of analysis computations. The first approach is based on using the SKOPE performance-modeling framework. In this case, we modeled reduction-style computations using SKOPE and enhanced SKOPE to capture disk performance, cache hits/misses, and page faults. Our second approach is based on analytical models. The results demonstrate that both these approaches are quite effective, with advantages (and disadvantages) associated with both in certain cases. Our third approach is based on linear interpolation, using which we effectively modeled in situ analysis kernels for molecular dynamics simulations.

Next, we present a model for optimizing end-to-end simulation-analysis execution workflow. The resource constraints for performing analysis are different depending on the application and the mode of simulation-analysis workflow. The data movement time from simulation to analysis may be a dominant cost in the workflow. Hence, it is important to determine the optimal data transfer frequency without degrading the simulation throughput. We propose a linear program to determine the optimal frequency of data transfer for analysis based on the resource constraints at the user’s execution sites, such as the network bandwidth, disk space, available memory, and computation time. Our model outputs the feasibility of performing the analysis with the objective of maximizing the analyses frequencies in order to derive maximum insight from the simulations. We demonstrate the effectiveness of our approach using large-scale molecular dynamics simulations on two supercomputers.

Finally, we present models for optimizing the analysis of light source data on geographically distributed resources. We develop simple models for different stages of the analysis workflow including data movement and computation. We also simulate the batch queues of different HPC resources to estimate the wait times. Experimental evaluation on production distributed infrastructure shows that the proposed models and system can be used for selecting the optimum resources.

Our future work includes integrating models for experiments and simulations. We are also working on developing a collective model for co-analysis and in situ analysis of simulations.