LearnPack: Capturing Learning Patterns for Science-Oriented AI Infrastructure

Justin M. Wozniak
Data Science and Learning Division, ANL
woz@anl.gov

CHALLENGE

The preeminence of artificial intelligence (AI) is making radical changes in the commercial and consumer computing spaces, and an impetus is needed to transform scientific computing work cycles into a form compatible with the advances that have been made and will be made with AI. The post-exascale era will create a multitude of software development challenges and high-level tools will be necessary to get the widest possible range of workloads running quickly. Other projects have begun using AI (via deep learning (DL)) to address scientific problems. The ECP Cancer Deep Learning Environment (CANDLE), for example, has produced a range of benchmarks or mini-apps that exemplify the application of DL to three pilot areas, including drug response, protein folding, and text synthesis for clinical reports. The high-energy physics community has started using DL systems to analyze the immense data streams produced at large colliders like CERN, and experiment-in-the-loop workloads are starting to become a reality.

Exciting new scientific applications that are rapidly developed to attack new, critical, and dynamic application spaces are *predominantly* large composite applications (or workflows) that integrate a great deal of software together. New computational paradigms, such as the prevalence of machine learning (ML) techniques, uncertainty quantification, and design optimization add to the importance of programming at this level. Applications thus face challenges when integrating the significantly different paradigms of high-performance computing (HPC), big data analysis, and the ML toolboxes emerging today.

Rapid progress in disciplines that benefit rely on HPC is stymied by a bottleneck not addressed by the Exascale Computing Project or other scientific computing programs. The effort and length of time needed to interact with HPC machines via exotic programming techniques reserved for highly skilled developers is a bottleneck that can be greatly alleviated in most cases by allowing users to interact with the machine via reusable patterns captured in AI-based services designed for various use classes. These include: humans, both expert and non-expert, which interact with the machine via higher-level queries; experiments, such as light source science cases in which image classification can be readily utilized to modify the running experiment; sensors, such as telescopes or weather data sources, that provide data to be assimilated into running models; and other devices, such as experimental autonomous vehicles or other agents.

OPPORTUNITY

This fundamental re-architecture – both technological and conceptual – of the computing complex for AI raises multiple critical research and design questions. Novel query interfaces will have to be developed that allow scientific questions to be formulated for the system. These will be accessible to the traditional programmer, but will be developed with an eye toward the developing user interfaces in voice, vision, and other areas rapidly developing in consumer electronics. These queries will have to be mapped to emerging DL and ML technologies in meaningful and efficient (i.e., proper data representation) ways. Uncertainty quantification will be a key component; it will be used to determine when a query cannot be answered satisfactorily and must be converted to a simulation run (or other data acquisition action). Automatic workflow control and deployment will be used to manage these simulations in support of the user query. These subsystems will manage the execution of the simulations, feed results to the storage hierarchy,

and provide additional training data back up to the DL interface, which will ultimately produce the response. The patterns that build up this computing model will be highly reusable.

These include patterns that address sampling, multiple-fidelity experiments, data integration and assimilation, convergence, anomaly detection, and other such high-level coordination concepts. Relevant techniques for sampling include automating management of parameterized task queues, asynchronous input and output of algorithm parameters, handling of partial sample results with handling of stragglers or missing results. Techniques for data integration include connecting matching tasks for assimilation, asynchronous/partial task-based reductions, and access to efficient I/O for assimilation with external data. Techniques for convergence include abstractions for task cancellation, termination, or modification if convergence is detected and task prioritization (e.g., for samples that are predicted to be closer to an optimum). Techniques for anomaly detection include retry strategies, management of conflicting results, and coupling of integrated testing modules.

Existing approaches to implementing outer loops algorithms are typically either monolithic codebases or build on existing workflow language/runtime solutions. Neither approach, however, is capable of achieving what is proposed here. Directed acyclic graph (DAG), block-synchronous parallel (BSP), and dataflow abstractions cannot be used to optimally implement these patterns. An important research direction will be to *research the primitives needed to support these patterns, build them upon a generic messaging abstraction, and present them as a system of coordination patterns for AI-driven studies*.

The ECP Cancer Deep Learning Environment (CANDLE) prototyped aspects of this approach in its reusable "Supervisor" workflow framework [1]. While this approach was originally deployed for hyperparameter sweep and optimization workflows, it more recently demonstrated reusability as applied to other workflow studies in data analysis [2] and decision boundary analysis [3]. What we are proposing here, however, is a more fundamental re-evaluation of how progress happens in an AI-based system and how to produce the underlying components that can make rapid AI-based studies possible and scalable.

TIMELINESS

Our approach – re-architecting computational experiments around reusable patterns in an AI-based system – has the capability to revolutionize an extremely broad range of applications, and is particularly appropriate for a range of DOE-relevant applications. This problem is not well-studied in other computer systems research. We propose that this re-architecture may be broken down into a reasonable set of programming systems research milestones and technology developments with regards to advanced workflows, data management, and ML.

A recent motivator for this model of development is the study of COVID-19 transmission through a large city population via multi-objective optimization [4], a case in which ML dominated the computation time and workflow complexity of the study. This project demonstrated adaptability in response to changing scientific questions in the early days of the pandemic, and showed that AI-driven workflows can be rapidly re-formulated and scaled to address dynamic problems.

This level of programming will become critically important as complex model exploration studies and AI-infused workflows are deployed on exascale systems. In the absence of a comparable systems and methodologies, researchers will have difficulty developing and deploying such complex applications.

REFERENCES

[1] J. M. Wozniak, R. Jain, P. Balaprakash, J. Ozik, N. Collier, J. Bauer, F. Xia, T. Brettin, R. Stevens, J. Mohd-Yusof, C. G. Cardona, B. V. Essen, and M. Baughman, "CANDLE/Supervisor: A workflow framework for machine learning applied to cancer research," *BMC Bioinformatics*, vol. 19, no. 18, p. 491, 2018. [Online]. Available: https://doi.org/10.1186/s12859-018-2508-4

[2] J. M. Wozniak, H. Yoo, J. Mohd-Yusof, B. Nicolae, N. Collier, J. Ozik, T. Brettin, and R. Stevens, "High-bypass learning: Automated detection of tumor cells that significantly impact drug response," in *Proc. Machine Learning in High Performance Computing Environments (MLHPC) @ SC*, 2020.

[3] R. Jain, A. Shah, J. Mohd-Yusof, J. M. Wozniak, T. Brettin, F. Xia, and R. Stevens, "Probing decision boundaries in cancer data using noise injection and counterfactual analysis," in *Computational Approaches for Cancer Workshop @ SC*, 2021.

[4] J. Ozik, J. M. Wozniak, N. Collier, C. M. Macal, and M. Binois, "A population data-driven workflow for COVID-19 modeling and learning," *International Journal of High Performance Computing Applications (Finalist for Gordon Bell COVID-19 Special Prize)*, 2021.