White Paper Submitted to NSF Workshop on Smart Cyberinfrastructure, Feb 2020

The Role of Machine Learning in Scientific Workflow Management on Distributed Cyberinfrastructure

Anirban Mandal (anirban@renci.org), Ewa Deelman (deelman@isi.edu)
10 February, 2020

Scientific workflows are key to today's computational science, enabling the definition and execution of complex applications in heterogeneous and often distributed cyberinfrastructure (CI). There are several current challenges in managing scientific workflows in distributed systems: composing the workflows, provisioning the resources for the workflows, and executing the workflows efficiently and reliably. Promising machine learning (ML) techniques can be applied to meet these challenges thereby enhancing the current workflow management system capabilities [1]. We foresee that as the ML field progresses, the automation provided by workflow management systems will greatly increase and result in significant improvements in scientific productivity and reproducibility. We herein describe a few representative possible applications of ML technologies that can enable efficient and reliable use of NSF-supported CI by scientific workflows.

Integrity Introspection for Scientific Workflows using ML: Data-driven science workflows often suffer from unintentional data integrity errors when executing on distributed national scale CI. However, today, there is a lack of tools that can collect and analyze integrity-relevant data from workflows and thus, many of these errors go undetected jeopardizing the validity of scientific results. In the context of the IRIS project [2], we are developing methods to automatically detect, diagnose, and pinpoint the source of unintentional integrity anomalies in scientific workflows executing on distributed CI. The approach is to develop an appropriate threat model and incorporate it in an integrity analysis framework that collects workflow and infrastructure data and uses ML algorithms to perform the needed analysis. Our goal is to integrate our solutions into the Pegasus workflow management system [3], which is used by a wide variety of scientific domains. It is also important to engage with science application partners (e.g. gravitational-wave physics, earthquake science, and bioinformatics) to deploy the ML analysis framework for their workflows, and to iteratively fine tune the threat models, ML model training, and ML model validation in a feedback loop.

Testbed Experimentation and ML Model Validation: An important aspect of developing ML based analysis techniques is the appropriate use of testbed infrastructures (NSFCloud [4][5] and other NSF-supported testbeds, e.g. ExoGENI [6]) to simulate realistic infrastructure conditions and error scenarios to train the ML models. For example, we are currently simulating aspects of the Open Science Grid [7] data distribution system in a testbed scenario to research on introspection and diagnosis of data integrity errors. The goal is to build an analysis framework that is powered by novel ML-based methods developed through experimentation in a controlled testbed, and then validated in and made broadly available on NSF production CI. Future NSF-supported infrastructures like the FABRIC [8] mid-scale research infrastructure can also be brought to bear to develop ML models for workflows with realistic production scenarios.

Performance Analysis for Workflows: ML techniques can also be utilized to address many of the challenges faced in managing scientific workflows in distributed systems. Several ML techniques are being used today to analyze the behavior of workflows at various levels of abstraction (workflow, task, and infrastructure) using different processing modalities (online and offline) and techniques to train the ML models [1]. Although there are promising initial results in using ML algorithms for scheduling, anomaly detection and provisioning resources for efficient and reliable workflow executions, the community is just at the beginning of exploring ML techniques in the scientific workflow space. Significant research advances need to be made in applications of ML technologies to the area of workflows to truly automate the analysis of the workflow behavior, understand the sources of anomalies, and make adaptation decisions to efficiently support the entire workflow lifecycle.

References

[1] E. Deelman, A. Mandal, M. Jiang and R. Sakellariou, “The role of machine learning in scientific workflows,” in *International Journal of High Performance Computing Applications*, doi: https://doi.org/10.1177/1094342019852127, May 2019.
[2] IRIS website. https://sites.google.com/view/iris-nsf/home
[3] E. Deelman, K. Vahi, G. Juve, M. Rynge, S. Callaghan, P. J. Maechling, R. Mayani, W. Chen, R. Ferreira da Silva, M. Livny, and K. Wenger, “Pegasus: a workflow management system for science automation.” *Future Generation Computer Systems*, vol. 46, pp. 17–35, 2015.
[4] NSF Chameleon Cloud. https://chameleoncloud.org/
[5] NSF CloudLab. https://cloudlab.us/
[6] I. Baldin, J. Chase, Y. Xin, A. Mandal, P. Ruth, C. Castillo, V. Orlikowski, C. Heermann, J. Mills. “ExoGENI: A Multi-Domain Infrastructure-as-a-Service Testbed.” *The GENI Book*, pp. 279--315, 2016.
[7] Open Science Grid. https://opensciencegrid.org/
[8] FABRIC. https://whatisfabric.net/

Contact author: Anirban Mandal (anirban@renci.org)