Cybersecurity and Privacy for Instrument-to-Edge-to-Center Scientific Computing Ecosystems

Ryan Adamson (adamsonrm@ornl.gov), Christian Engelmann (engelmannc@ornl.gov)
Oak Ridge National Laboratory, Oak Ridge, TN, USA

Challenge: The DOE's Artificial Intelligence (AI) for Science report [1] outlines the need for intelligent systems, instruments, and facilities to enable science breakthroughs with autonomous experiments, "self-driving" laboratories, smart manufacturing, and AI-driven design, discovery and evaluation. The DOE's Computational Facilities Research Workshop report [2] identifies intelligent systems/facilities as a challenge with enabling automation and eliminating human-in-the-loop needs as a cross-cutting theme. Autonomous experiments, "self-driving" laboratories and smart manufacturing employ machine-in-the-loop intelligence for decision-making. Human-in-the-loop needs are reduced by an autonomous online control that collects experiment data, analyzes it, and takes appropriate operational actions in real time to steer an ongoing or plan the next experiment. DOE laboratories are currently in the process of developing and deploying federated hardware/software architectures for connecting instruments with edge and center computing resources to autonomously collect, transfer, store, process, curate, and archive scientific data. These new instrument-to-edge-to-center scientific ecosystems face several cybersecurity and privacy challenges.

Computing systems from different administrative domains with different cyber security policies are interconnected with each other. This may involve instrument control systems, laboratory robots/automation/sensors, edge computing devices for real-time processing, Cloud-like computing for design of experiments and DOE's Leadership computing systems for scientific data analyses and digital twins. Scientific data, including experiment setup data, control data, and results, is then transferred between and processed in these different administrative domains. Resource orchestration across administrative domains ensures that required resources, such as network and compute, are available when needed and at the required capacity. The involved administrative domains may be different network enclaves within a DOE laboratory, or involve the experimental facilities of outside organizations such as other laboratories, universities, and industry. Some of these instrument-to-edge-to-center scientific ecosystems may even cross country boundaries to connect unique experimental facilities with unique computing capabilities.

The specific cybersecurity and privacy challenges for such instrument-to-edge-to-center scientific ecosystems are multifold and include bridging the differences in cyber security policies of several administrative domains, ensuring operational safety and security of experimental facilities and guarding the privacy and integrity of scientific data. Ecosystem computing can be represented as a set of the composable building blocks (system of systems) used by scientific workflows. Unfortunately, no single organization has the authority, responsibility, or capability to secure multi-organizational interconnected systems. ***Thus, the wholistic application of cybersecurity and privacy to these interconnected systems must ultimately be owned by the scientific workflow operators themselves***. The workflow itself is the only layer of this model that interacts with the complete set of systems, which often do not expose security, trust, and assurance primitives to scientists.

Today, the individual systems that make up autonomous experiments, "self-driving" laboratories, and smart manufacturing are *already* secured to appropriate standards set by their managing organizations. Yet, security and privacy concerns are not being wholistically addressed for scientific workflows. The NIST risk management framework [3] is commonly used within DOE for this purpose, but the resulting Confidentiality, Integrity, and Availability assurance levels are incredibly coarse; systems receive just one of three ratings: Low, Moderate, or High! These categories are insufficient since organizations rated at the same level are rarely able to interconnect systems without some normalization. What can be done to facilitate trust between various experimental facilities, networks, edge devices, and supercomputer centers?

The most effective security tools available today are inadequate for a multi-organization system of systems. Mandatory Access Control (MAC) systems such as AppArmor and SELinux only operate at the single node level within a single organization. Firewall rules are largely implemented to police IP addresses instead of data content because the very nature of inter-organization communication is endpoint-based and not data-centric. Recent research in Zero Trust Architectures (ZTA) has led to successful implementation and best practice development, but implementation of ZTA is still typically limited in scope to single applications, protocols, or organizations. ***How can we ensure privacy and security of distributed systems when tools are fundamentally based on a localized security model?***

Opportunity: Several research areas exist regarding the security of scientific computing ecosystems:

1) Security and privacy risk bounding techniques such as Uncertainty Quantification (UQ) have not yet been applied wholistically to systems, such as edge sensors, network interconnects, backing storage, and computational systems. ***Fine-grained UQ may enable scientific workflow operators to make strategic decisions about which systems to use and which ones to avoid in order to attain an acceptable measure of scientific data privacy and integrity assurance***.

2) Mandatory access control primitives are not present in distributed, highly-scalable ecosystem computing. Research and development is needed to discover, define, and integrate these primitives into scientific workflows and systems. ***Distributed mandatory access control methods may need to be integrated by workflow operators and subsequently enforced by experimental facilities to the satisfaction of all organizations spanned by autonomous workflows.***

3) Network security techniques have developed around the fact that communication *endpoints* are easier to protect than the data itself, but *data* protection at the network layer is needed. New data-centric communication protocols such as Content Centric Networking (CCN) and Named Data Networking (NDN) promote data sets to first class citizen status and replace the Internet Protocol (IP) layer of the network stack. NDN in particular requires cryptographically signed data packets at the lowest levels of the networking stack. ***When data trustworthiness is provided at the network layer, the application layers above become much more secure and much less complex***.

Timeliness or maturity: The instrument-to-edge-to-center scientific ecosystem is extremely complex. Today, there are already 300 independently maintained workflow solutions [4]. Tools to reduce and bound complexity and risk are urgently needed to meet the privacy and cybersecurity requirements of all the stakeholders that participate in this new scientific ecosystem. Without techniques to assess and trust the security of distributed workflows, many organizations will not be able to provide resources to enable "self-driving" laboratories and machine-in-the-loop workflows. Research into these opportunities is timely; UQ research has been applied recently to quantify AI model accuracy, and a recent Executive Order [5] has highlighted the need for implementing Zero Trust networking models across the federal government. MAC is a key component of Zero Trust, and implementation of NDN can essentially enable Zero Trust networking for free (for all layers of the networking stack) since data provenance is inherently necessary to the protocol.

References:

[1] **AI for Science Report.** March 2020. URL https://www.anl.gov/ai-for-science-report
[2] **DOE National Laboratories' Computational Facilities – Research Workshop Report.** ANL/MCS-TM-388. February 2020. URL https://publications.anl.gov/anlpubs/2020/02/158604.pdf
[3] **NIST Risk Management Framework.** 2018. URL https://doi.org/10.6028/NIST.SP.800-37r2
[4] **List of Computational Data Analysis Workflow Systems**. URL https://s.apache.org/existing-workflow-systems
[5] **Executive Order on Improving the Nation's Cybersecurity.** May 2021. Executive order 14028