NSF Campus Cyberinfrastructure and Cybersecurity Innovation for Cyberinfrastructure PI Workshop

September 23 – 25, 2019 | Minneapolis, MN

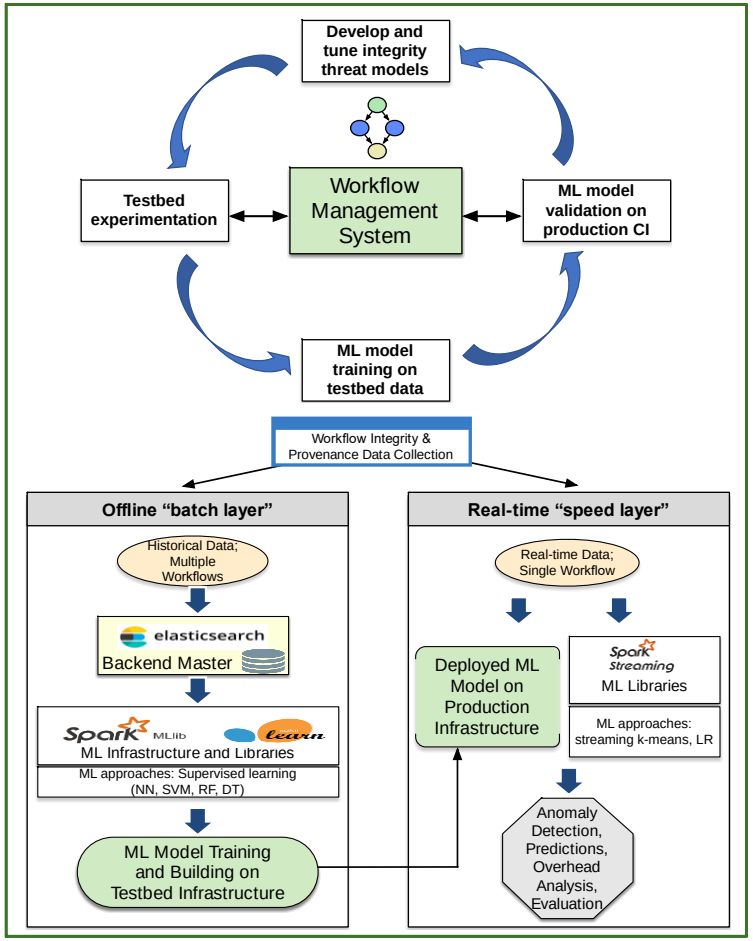
Quad Chart for: *CICI SSC: Integrity Introspection for Scientific Workflows (IRIS)*

Challenges:

- Scientific workflow processing sometimes suffers from data integrity errors when executed on distributed cyberinfrastructure (CI).
- Lack of tools that can collect and analyze integrity-relevant data and hence, errors go undetected and corrupt data becomes part of the scientific record.

Deliverables:

- Develop an integrity introspection framework that collects integrity data and utilizes ML algorithms to automatically detect, analyze and pinpoint source of integrity errors.
- Train ML algorithms on controlled testbeds and validate on national CI by integrating framework with Pegasus.
- Engage with science application partners in gravitational-wave physics, earthquake science, and bioinformatics to deploy the analysis framework.



Broader Impact:

- The IRIS integrity introspection framework will be available to a broad range of domains that rely on Pegasus.
- IRIS will contribute to the discussions on reproducibility since integrity is essential to supporting reproducibility.
- Integrity-relevant data collected in IRIS and ML algorithms developed can be used by students and researchers.

Metadata tag:

- *<https://sites.google.com/view/iris-nsf/>*
- *<Application of ML for integrity analysis>*
- *<Looking for different sources of integrity data: infrastructure and application>*
- *<Builds on CICI SWIP project, by adding detection and analysis of integrity errors>*

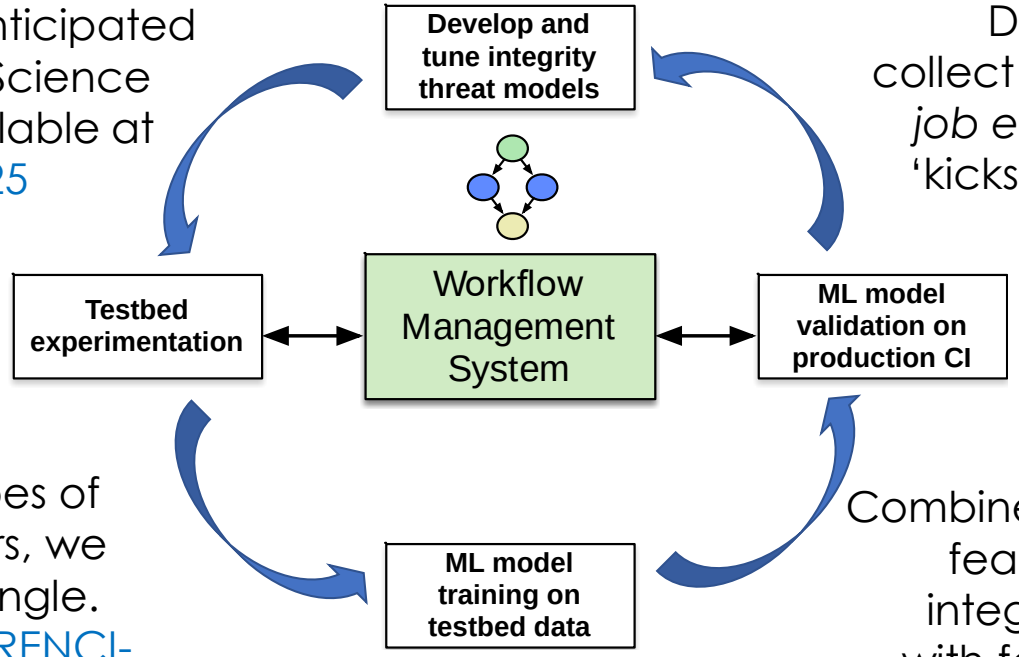
PI/co-PIs: Anirban Mandal, Ewa Deelman, Von Welch
Contact: anirban@renci.org

renci
USC Viterbi School of Engineering Information Sciences Institute
CENTER FOR APPLIED CYBERSECURITY RESEARCH

IRIS Approach and Accomplishments

Goal: *Detect, diagnose, and pinpoint* *the source of unintentional integrity anomalies in scientific workflow executions on distributed cyberinfrastructure.*

Threat Model development
Developed threat models for anticipated integrity issues based on Open Science Cyber Risk Profile (OSCRP). Available at http://hdl.handle.net/2022/23225



Integrity Data Collection
Developing new capabilities in Pegasus to collect integrity-related metrics: new *composite job event* that triages information from (a) job 'kickstart' output, (b) job stdout and stderr, (c) job transfers, and (d) per file checksum.

Testbed Experimentation
In order to simulate different types of sources of storage integrity errors, we added a new utility in Chaos Jungle. Github link: https://github.com/RENCI-NRIG/chaos-jungle/tree/storage/storage

Integrity Data Analysis
Combine workflow and incomplete infrastructure features: ML inference from end-to-end flow integrity checks with training data generated with fault injection. Cast it as a ***network system Root Cause Analysis (RCA)*** problem.

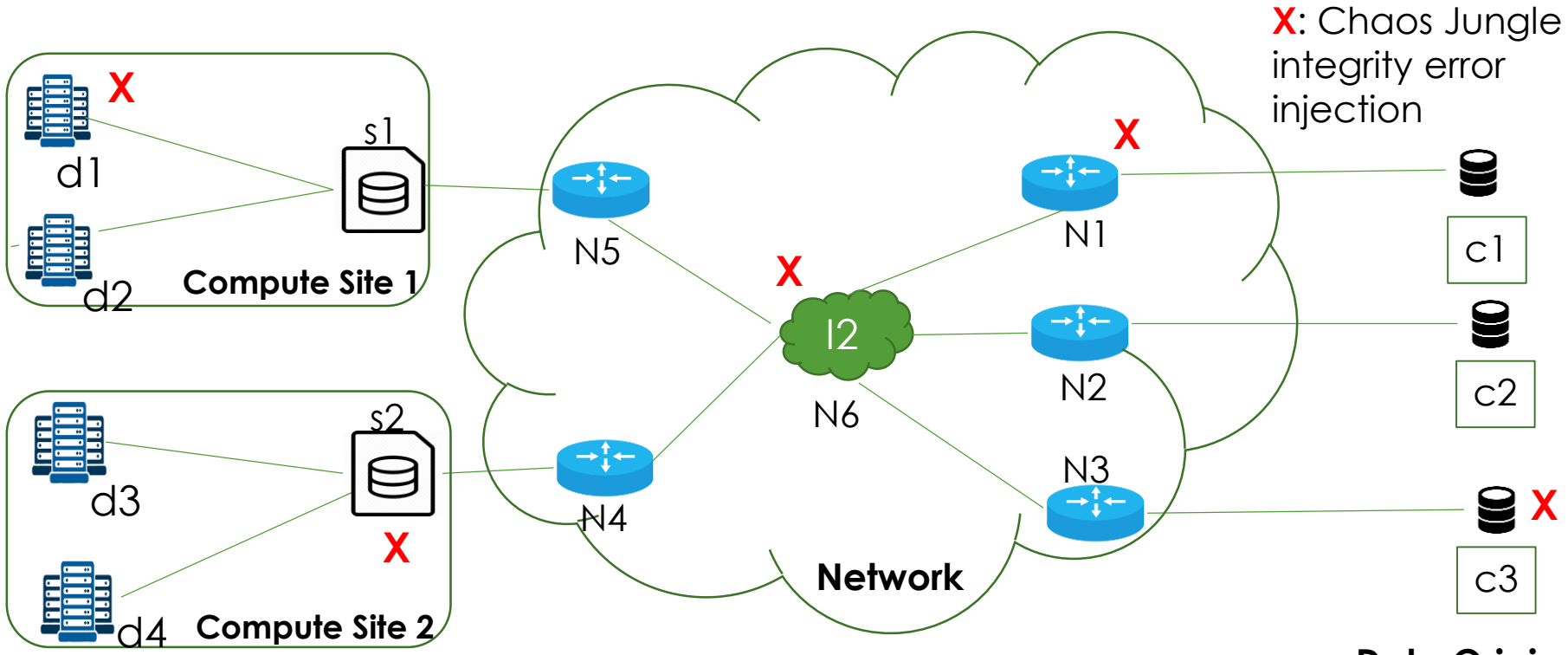
renci

USC Viterbi School of Engineering
Information Sciences Institute

CENTER FOR APPLIED CYBERSECURITY RESEARCH

Network System Root Cause Analysis

Work in Progress: *Emulating content distribution networks (CDN) on ExoGENI and injecting integrity errors with Chaos Jungle. The emulated system is inspired by science data flows/distribution on the Open Science Grid (OSG).*



- Gathering training data with several features.
- Labels are infrastructure entities (links, compute nodes, caches, storage nodes) responsible for root cause of integrity error.
- ML analysis of data and model development are ongoing.

renci

USC Viterbi School of Engineering Information Sciences Institute

CENTER FOR APPLIED CYBERSECURITY RESEARCH