CICI:TCR:SAFARI: Scientific Analytics, Forensics, and Reproducibility for Workflows in Cyberinfrastructure (CI)

PI: Michela Taufer (U. of Tennessee, Knoxville) Co-PI: Ewa Deelman (U. Southern California)
Team: Kin Ng (UTK), Ty Anderson (UTK), Rajiv Mayani (USC), Jack Marquez (UTK), Jay Lofstead (Sandia)
https://globalcomputing.group/safari

Goal	Robust and Resilient Open Science
Objectives	Trustworthiness | Reusability | Reproducibility
Services	Provenance, documentation, and verification | Decomposition and abstractions | Annotation, provenance, and validation
Outputs	Practices | Artifact commons | Automation
Outcomes	Tools and Methodologies for Data Reliability and Integrity Applied to Earth Science
Impacts	End-to-End Robust and Resilient Open Science

SAFARI integrates forensic data analytics into the Pegasus Workflow Management System to enhance trustworthiness, reusability, and reproducibility (TR&R) of scientific workflows. By modularizing workflows, embedding provenance and verification, and automating integrity checks, SAFARI ensures results are reliable, secure, and explainable across high-throughput computing (HTC) platforms

Workflow Building Blocks	Trustworthiness	Reusability	Reproducibility
Artifacts	Deployment of vulnerability scanners & integrity checks	Containerization, data Wrangling & virtualization	Integration of provenance layers & execution metadata
Toolchains	Tooling for code & data integrity & provenance	Layered software systems & high-level interfaces	Code variability reduction & verification
Automation	Automated attestations & paired verification	Identification & mapping of workflow patterns	Intermediate restarting, generative testing, specifications
Practices	Procedures to improve artifact trustworthiness	Characterization & taxonomy of interface labels	Standard formats & practices from data generation to validation

Cells: Research-heavy with software aspects, software heavy but including research, and balanced between research and software

Cybersecurity Innovation

- Embed *forensic data analytics* directly into workflow systems (Pegasus); ensure *trustworthiness, reusability, and reproducibility (TR&R)* of scientific workflows; and automate *data provenance, validation, and integrity checks* for end-to-end security
- Modularize workflows into *containerized components* for explainability and reuse; and establish *forensic analytics as a core CI service*, setting new standards for open science

Approach For Transitioning the Innovation

- Provide *interoperable building blocks* (artifacts, toolchains, automation, practices)
- Deliver *containerized workflow modules* for portability and reusability
- Integrate *automated forensic verification frameworks* directly into Pegasus workflows
- Train students and researchers through *hands-on forensic CI services*
- Engage with communities (ACCESS, SC, PASC, AGU, AAAS) to build adoption pathways

Evaluating and Demonstrating Transiton

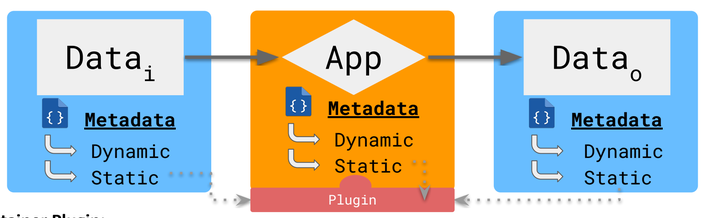
- Num. of workflows with embedded forensic analytics with trustworthiness (provenance integrity, validation accuracy) and reproducibility across HPC/HTC platforms. (*Metrics*)
- Adoption and reuse of components by Earth science and cross-domain users, with training outcomes measured by num. of students and early-career researchers engaged. (*Metrics*)
- Open-source forensic workflow modules via Pegasus WMS GitHub, with documentation, tutorials, and training through ACCESS affinity groups and workshops. (*Community Access*)
- Dissemination via SC, PASC, AGU, AAAS, and community platforms to ensure broad access and reproducibility. (*Community Access*)

Programmatic Details

- 3-year project started on October 2025
- Led by U. of Tennessee, Knoxville with U. of Southern California (ISI)
- Unfunded collaborations with Earth science research groups, ACCESS communities, and Sandia National Laboratories

Outcomes of Year 1: Capturing Provenance in Scientific Workflows

Apptainer plugin to enable fine-grained containerization of workflows, ensuring provenance, traceability and reproducibility



Apptainer Plugin:
- Builds workflow, executes workflow, and annotates provenance metadata into corresponding containers

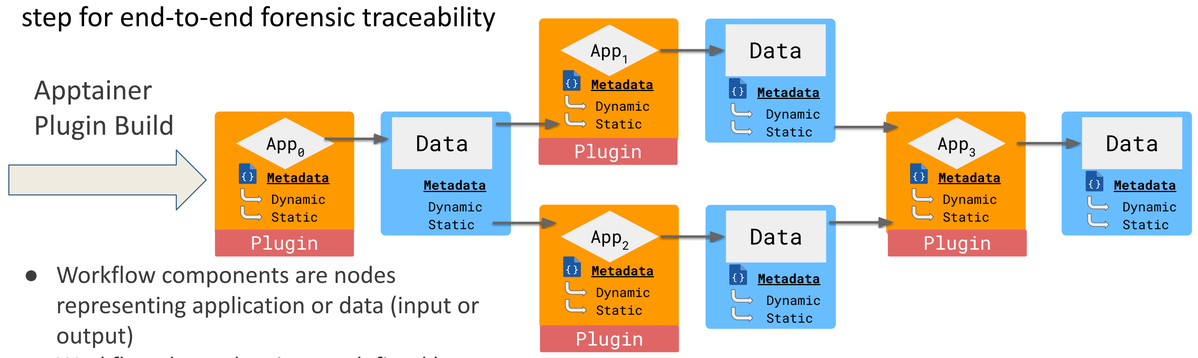
Provenance Metadata:
- Dynamic: Annotated at container build time
- Static: Annotated at runtime

Data Container
Application Container

```
"WorkflowName": "simple_wf",

"Nodes":
[
  {
   "ID": "node_id",
   "Type": "input"OR "app"
    OR "output",
   "Container":
    {
        "Name": "name",
        "InPath":
"sif/path",
        "Size": 0,
    },
   "Inputs": [],
   "Outputs": []
  }
    .
    .
    .
]
```

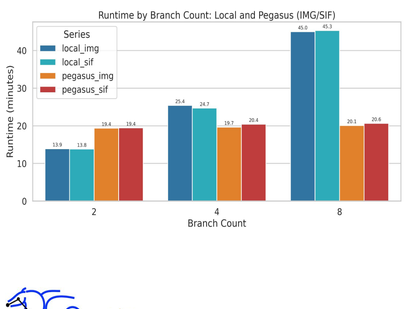
Building arbitrary fine-grained containerized DAG workflows with embedded provenance at every step for end-to-end forensic traceability

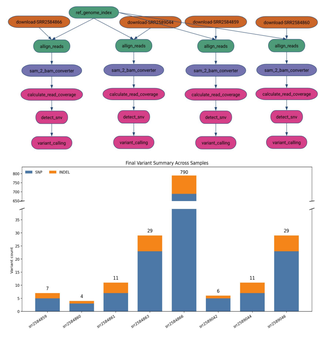


- Workflow components are nodes representing application or data (input or output)
- Workflow dependencies are defined by Inputs[]/Outputs[]

Scaling to Distributed CI: Workflow Execution via Pegasus

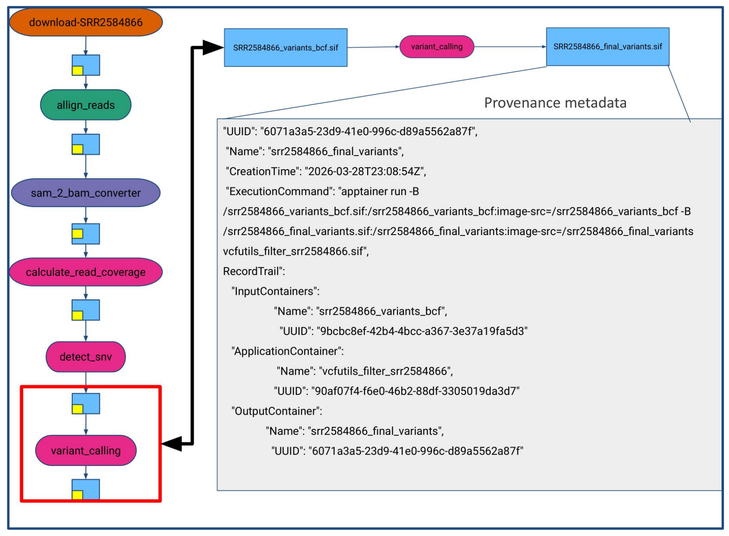
Integrating fine-grained containerized workflows with the Pegasus Workflow Management System enables *scalable, parallel execution across distributed cyberinfrastructure*. We validate our approach on a real bioinformatics variant-calling workflow (E. coli, 8 sequencing runs) deployed on Purdue Anvil via NSF ACCESS, demonstrating that Pegasus achieves near-constant runtime as workflow parallelism increases





8 sequencing runs produce **highly variable** final variant counts from heterogeneous read datasets

Provenance Metadata in Action: Forensic Traceability at Every Workflow Step



Benefits to Scientific Cyberinfrastructure

- Impacts scientific workflows across diverse domains — including bioinformatics, Earth science, materials science, and physics — where verifiable data lineage and result integrity are essential for trustworthy, reproducible research
- Establish trust in workflows by ensuring data integrity, reproducibility, and reusability—raising the standard for open and explainable science across HPC/HTC platforms.

Risks Versus Potential For Advances

- Risks: Integration of forensic analytics may introduce overhead or complexity; adoption requires cultural and technical shifts in workflow communities
  - Payoff: Sets a new benchmark for trustworthy, reproducible scientific workflows; enables cross-domain reuse, reduces costs; and builds workforce expertise in secure and explainable CI.