

From Scripts to Workflows: Orchestrating Science with Pegasus

Ewa Deelman

University of Southern California
Information Sciences Institute

deelman@isi.edu



SciTech Activities

1. Develop technologies to support scientific workflows
2. Provide user support on national cyberinfrastructures and clouds
3. Research new avenues in the workflow management space
4. Provide advice on the design, implementation and use of Cyberinfrastructure (CI) for the Research Infrastructure lifecycle
5. Create awareness and understanding of opportunities in CI for Science

USC/ ISI



Marina del Rey, CA

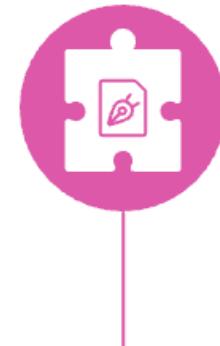
Technology Development

Focus on creating tools for scientific workflows.



Research Avenues

Explore new possibilities in workflow management.



Awareness Creation

Promote understanding of CI opportunities in science.



User Support

Assist users with national cyberinfrastructures and cloud services.



Design Advice

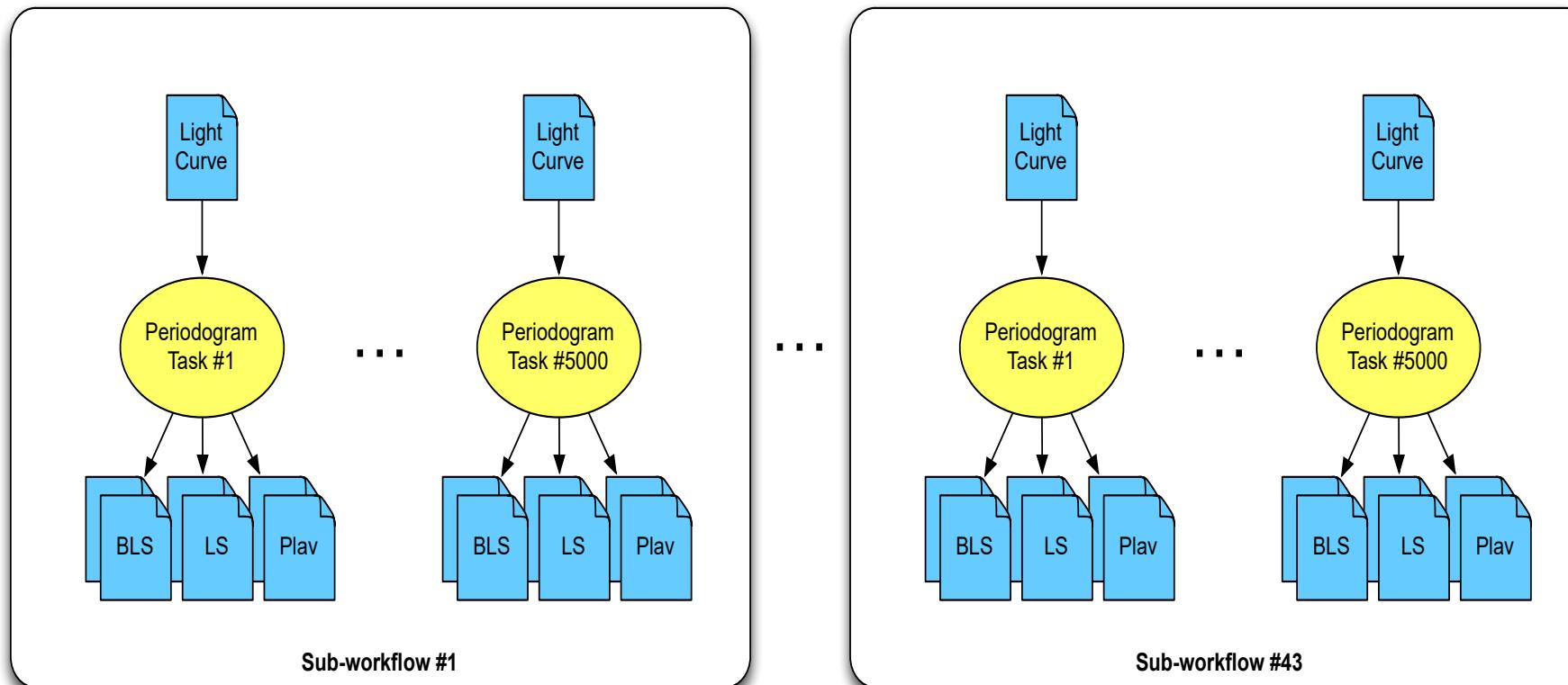
Offer guidance on CI design and implementation.

Most projects are collaborative across 4-6 organizations

Some workflows are simple!



Bag of Tasks



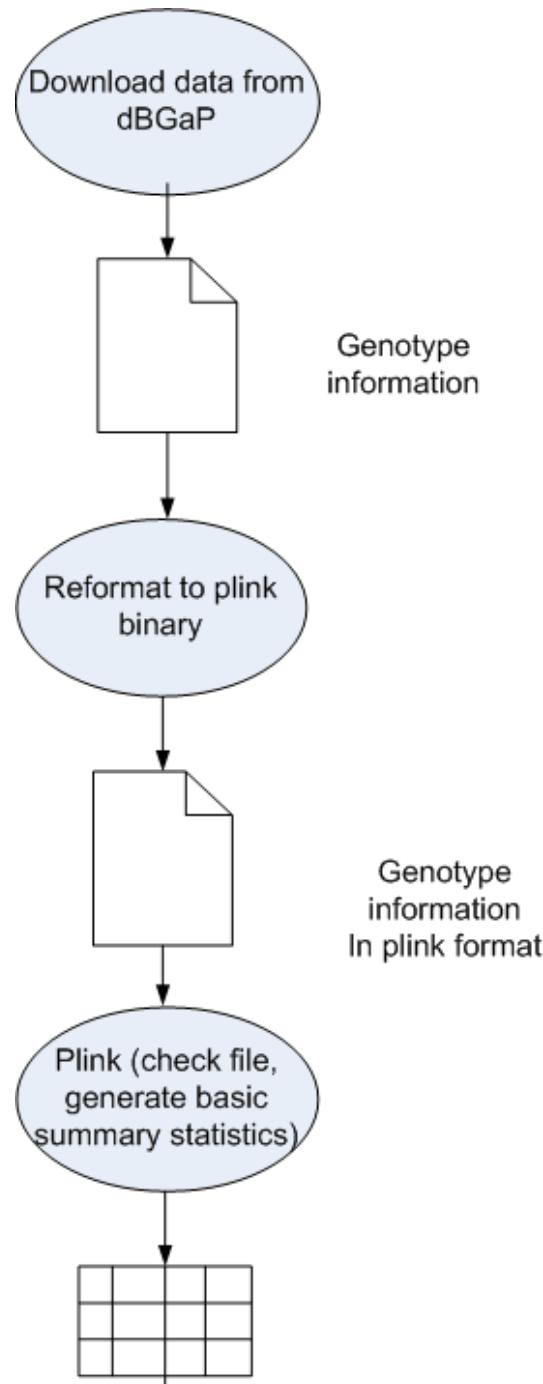
Kepler periodogram workflow

Scripts

- Quick to code and deploy
- Quick to learn and tailor to the local campus resource
- Not very scalable, hard to move between resources, hard to maintain, hard for others to understand, often error-prone

Computational workflows

- Help express multi-step computations in a declarative way
- Can support automation, minimize human involvement
 - Makes analyses easier to run
- Can be high-level and portable across execution platforms
- Keep track of provenance to support reproducibility
- Easier to maintain
- Foster collaboration—code and data sharing
- Take some time investment to learn



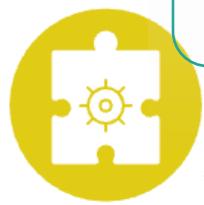


Workflow Challenges Across Domains

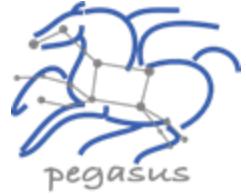
- Describe complex workflows in a simple way
- Access distributed, heterogeneous data and resources (heterogeneous interfaces)
- Deal with resources/software that change over time
- Ease of use. Ability to monitor and debug large workflows

Our Focus

- Separation between workflow description and workflow execution
- Workflow planning and scheduling (scalability, performance)
- Task execution (monitoring, fault tolerance, debugging, web dashboard)
- Workflow optimization, restructuring for performance and fault tolerance.



Submit locally, run globally



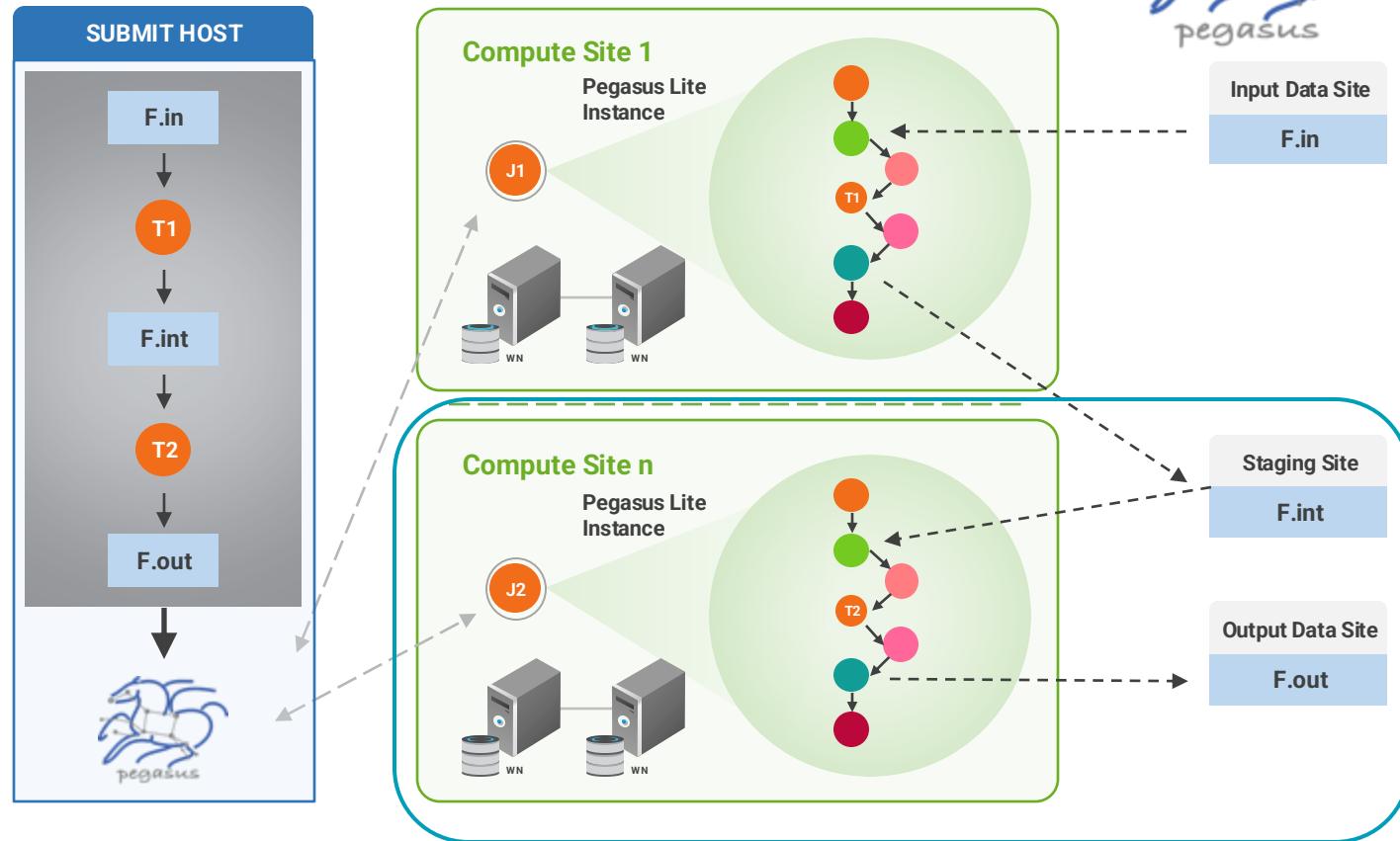
► Pegasus WMS == Pegasus planner (mapper) + DAGMan workflow engine + HTCondor scheduler/broker

- Pegasus maps workflows to target infrastructure (1 or more resources)
- DAGMan manages dependencies and reliability
- HTCondor is used as a broker to interface with different schedulers

► Planning converts an abstract workflow into a concrete, executable workflow

- Planner is like a compiler
- Optimized performance
- Provides fault tolerance

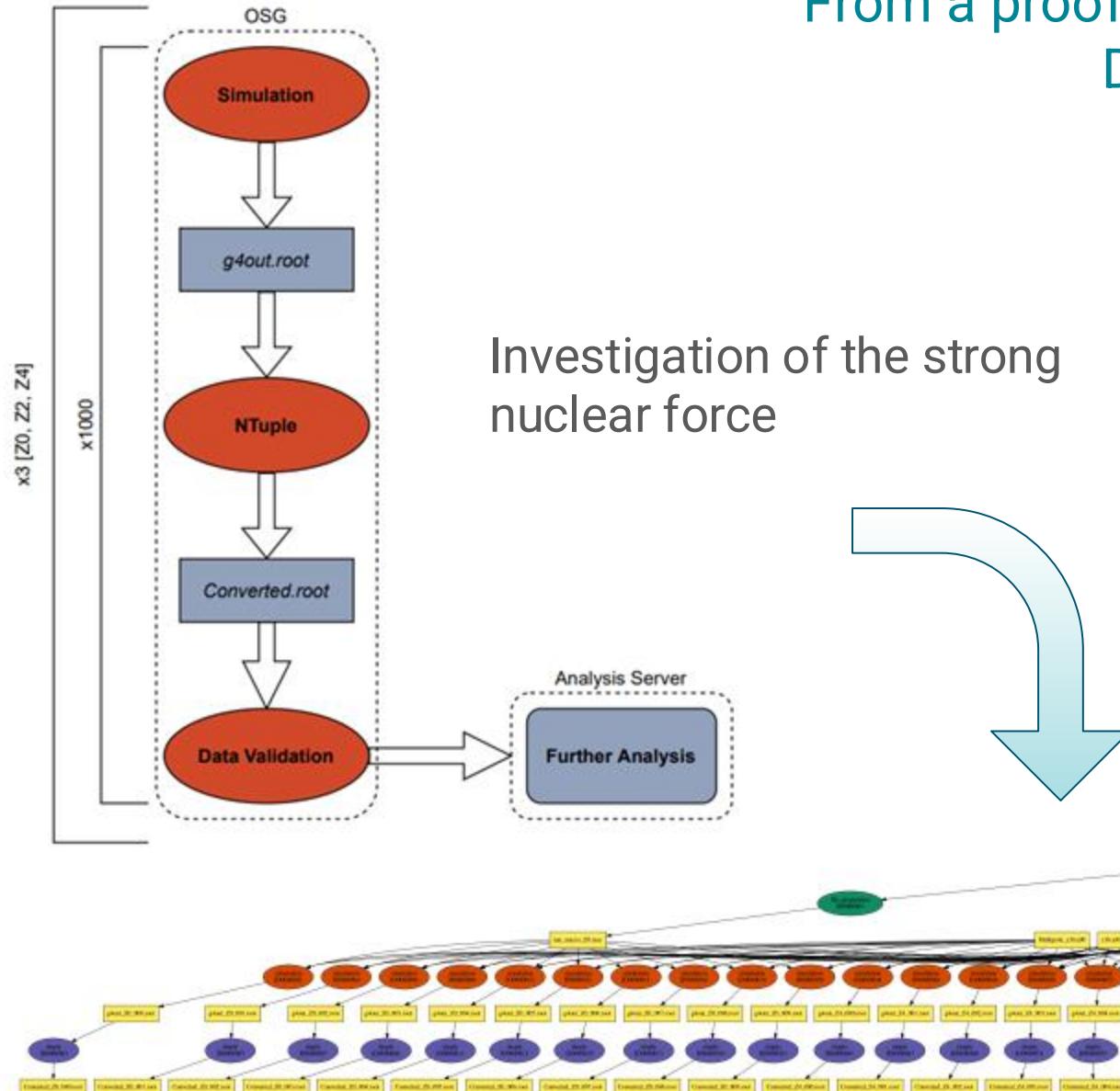
► Can leverage distributed and heterogeneous CI



From a proof of concept to winning the 2022 David Swanson Memorial Award



Connor Natzke,
Physics Student
Colorado School of Mines



Investigation of the strong
nuclear force

Pegasus-based
Monte-Carlo
simulation, 590,000
jobs, 15 years total
wall time, 4 hours wall
time on Open Science
Grid

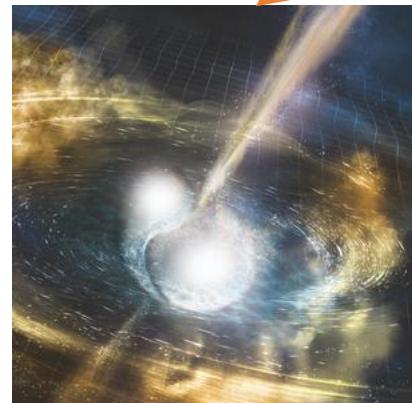
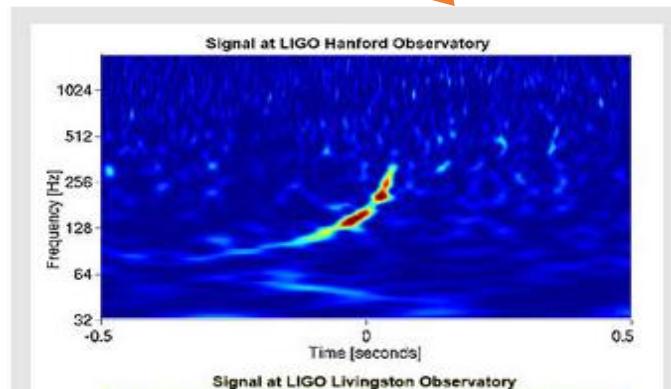
- Pegasus provided:
- 1) Automation and scaling
 - 2) Automatic job retries
 - 3) Automation of file transfers
 - 4) Managed disk space at execution sites

Workflows Can Move with Time

Pegasus use in the Laser Interferometer Gravitational Wave Observatory



Nobel
Prize



First Pegasus
prototype



Blind injection detection

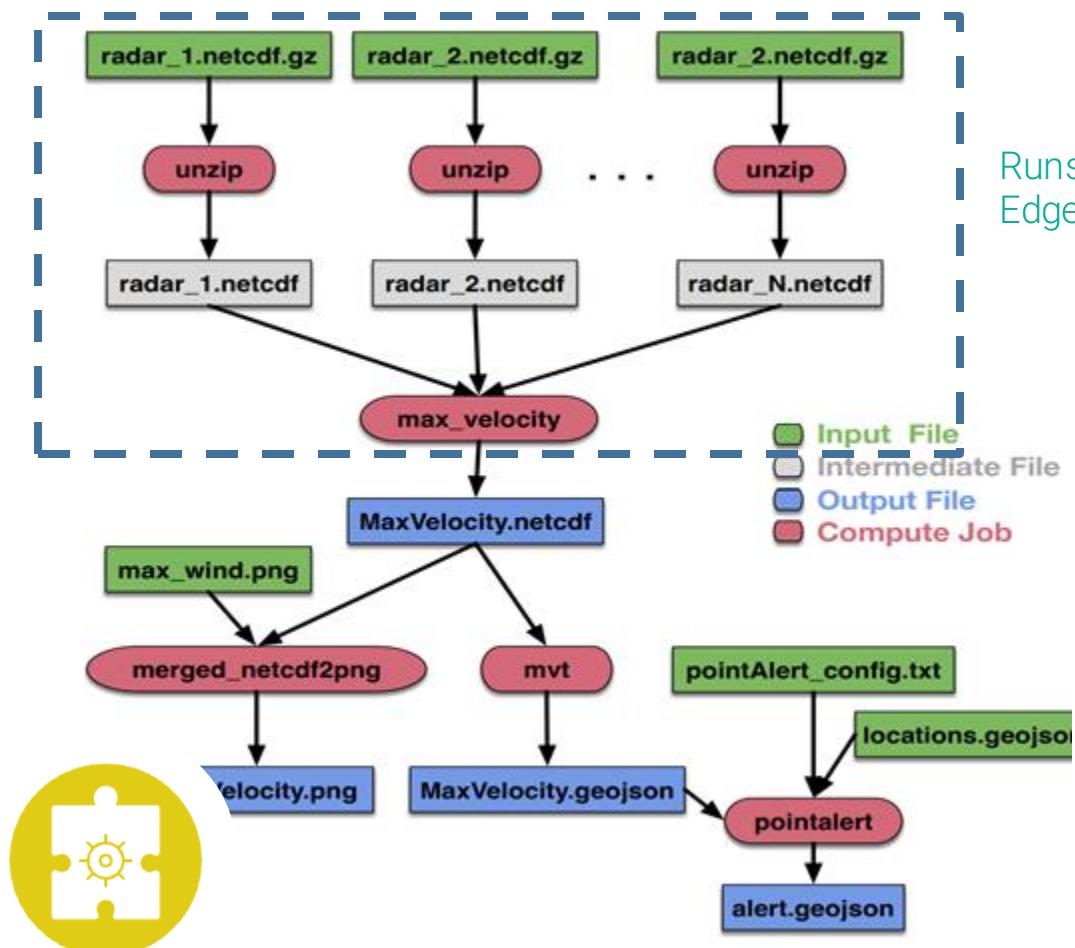
First detection of
black hole collision

Multi-messenger
neutron star
merger
observation

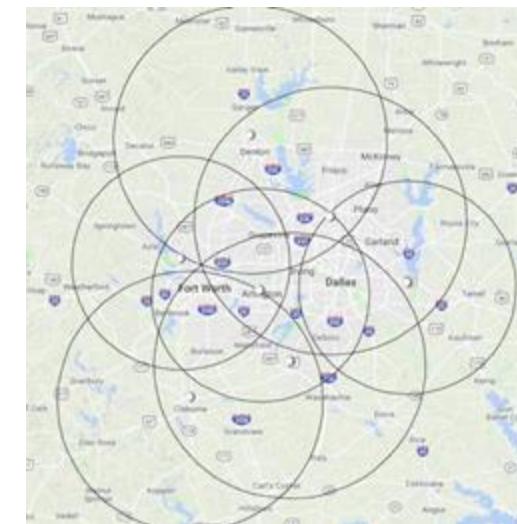
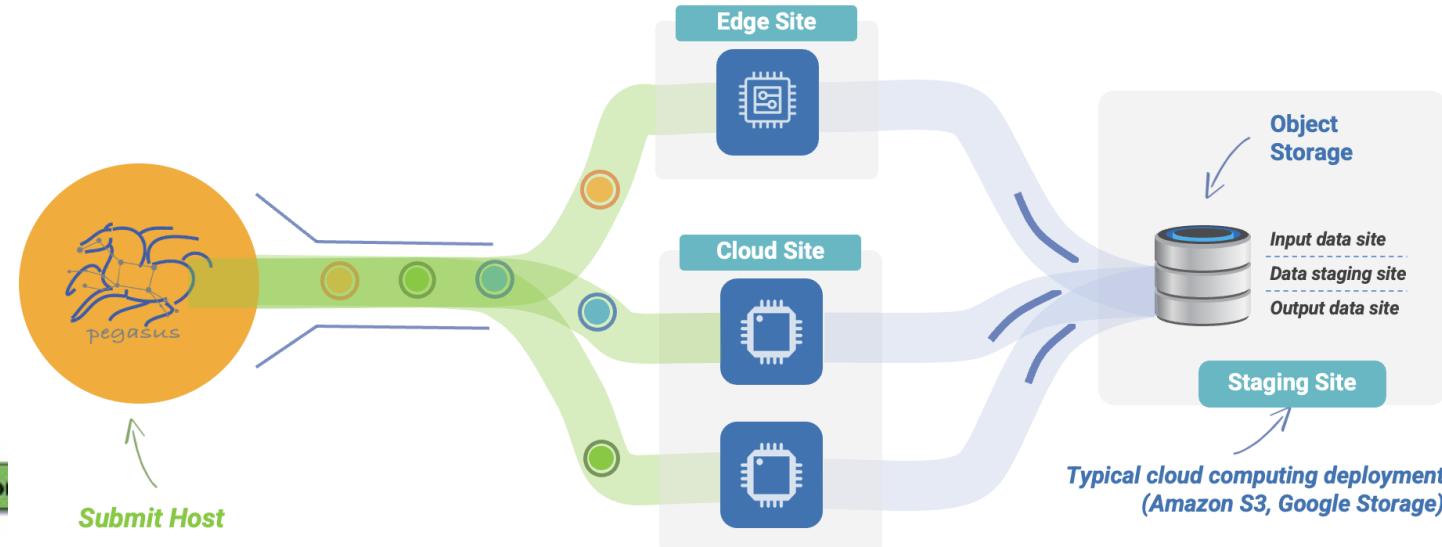
Edge-2-Cloud Applications

CASA: Collaborative and Adaptive Sensing of the Atmosphere

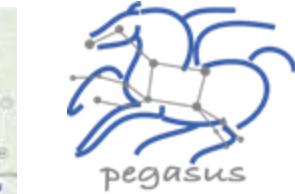
- Has deployed a network of short-range Doppler radars
- Compute and data repositories at the edge, close to the radars
- Use on demand cloud resources to scale up their computations



Runs at the Edge

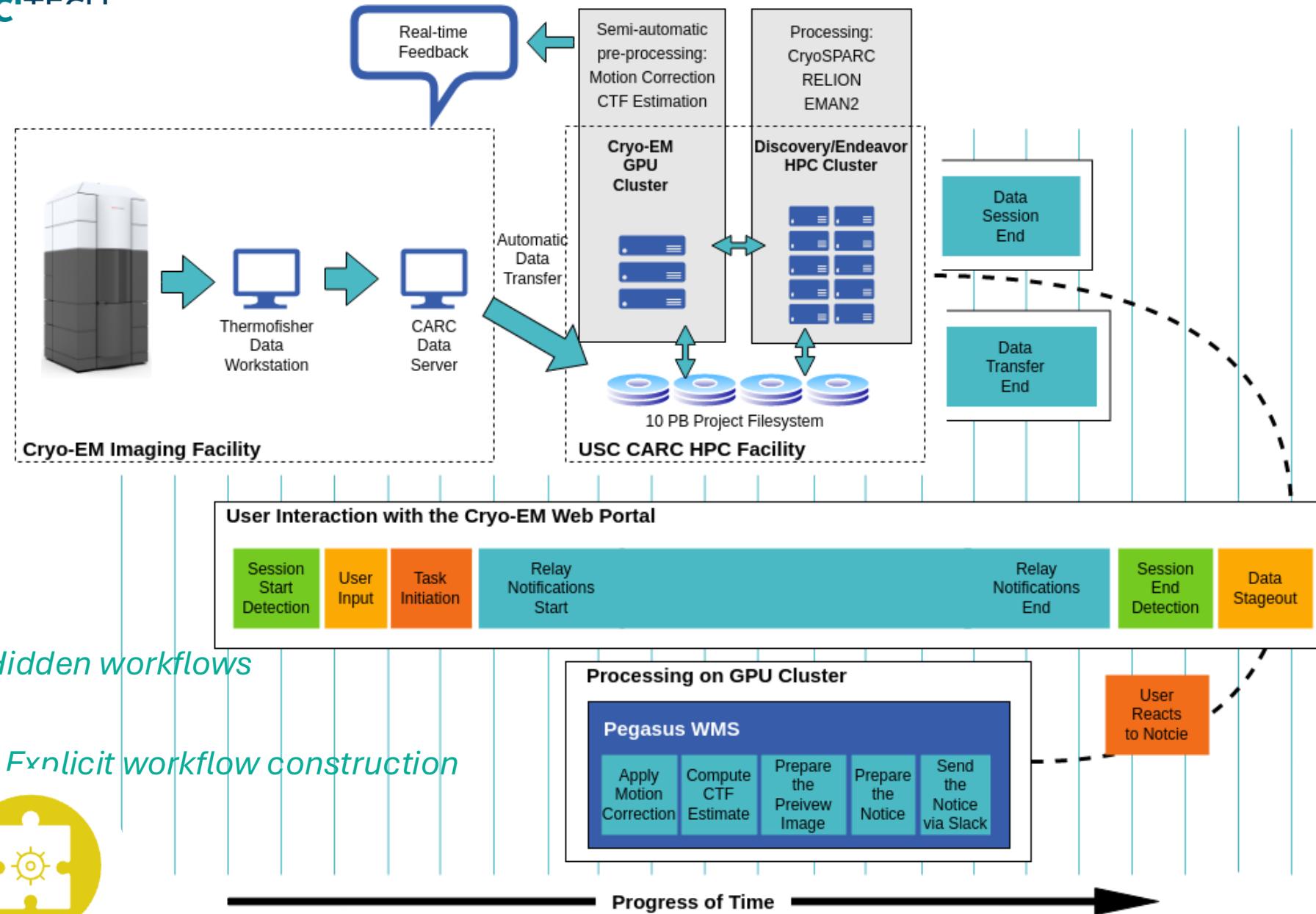


<http://www.casa.umass.edu/>



Decoupling of compute and storage

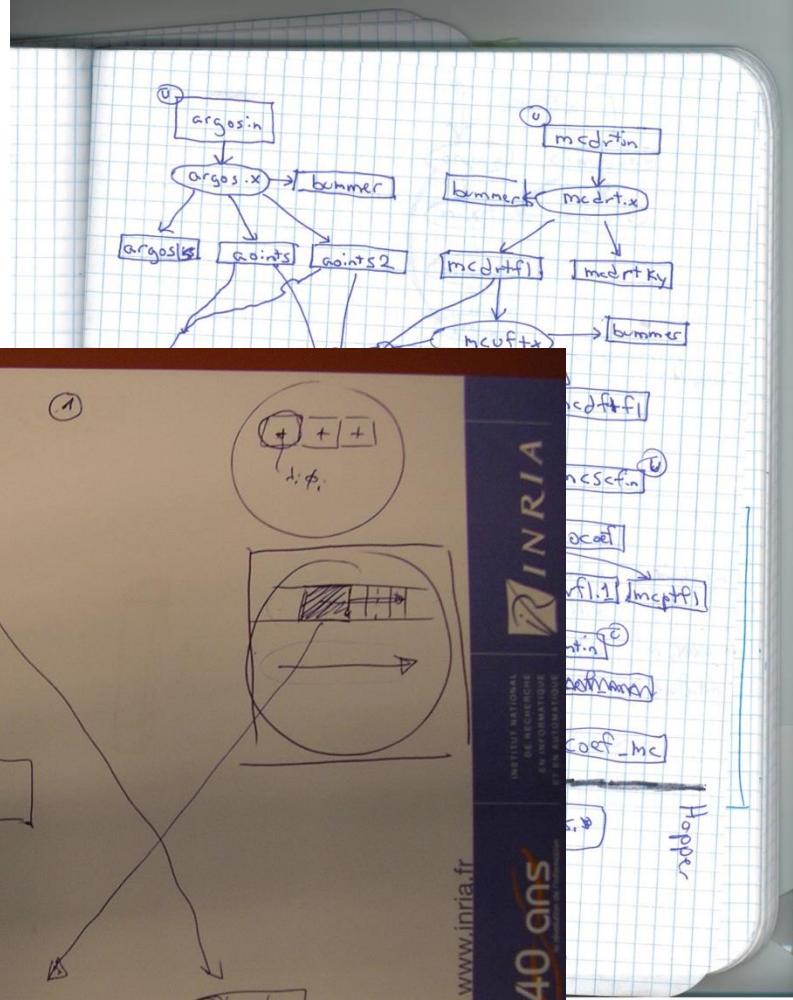
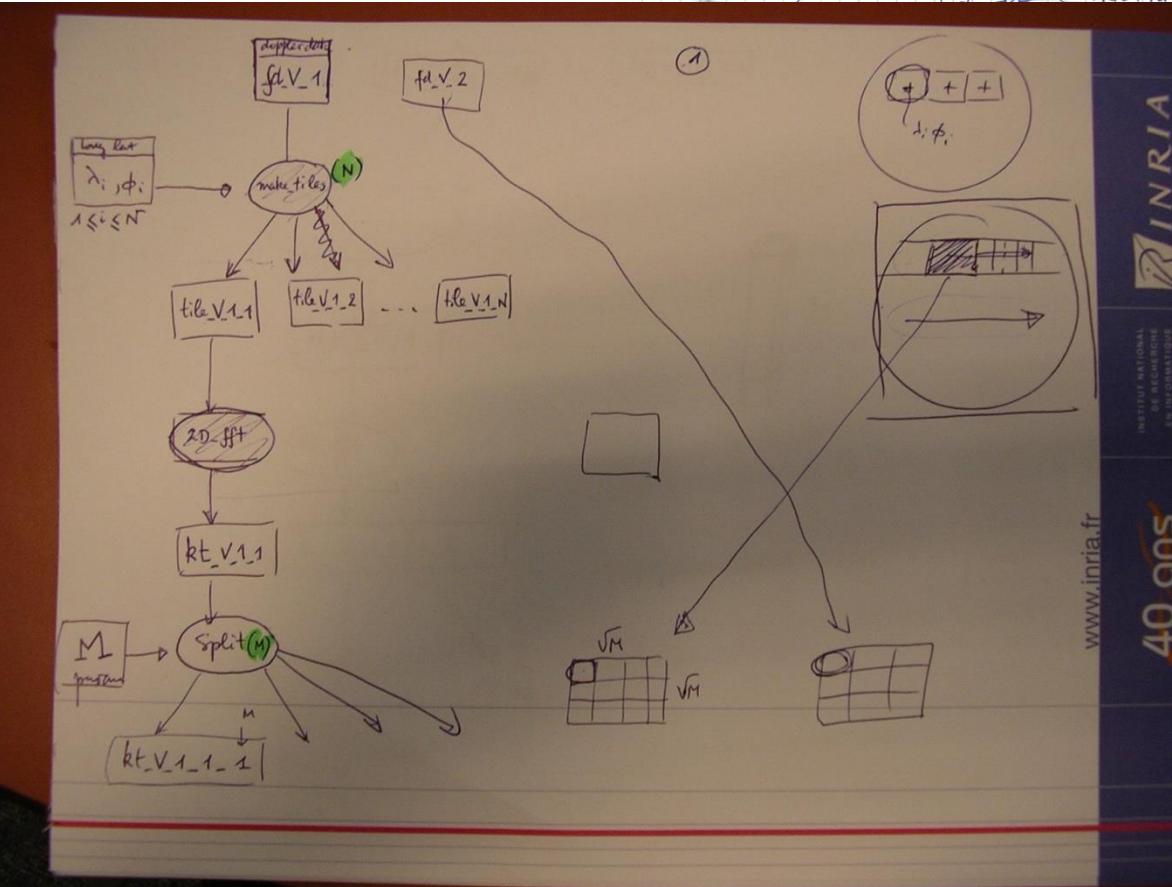
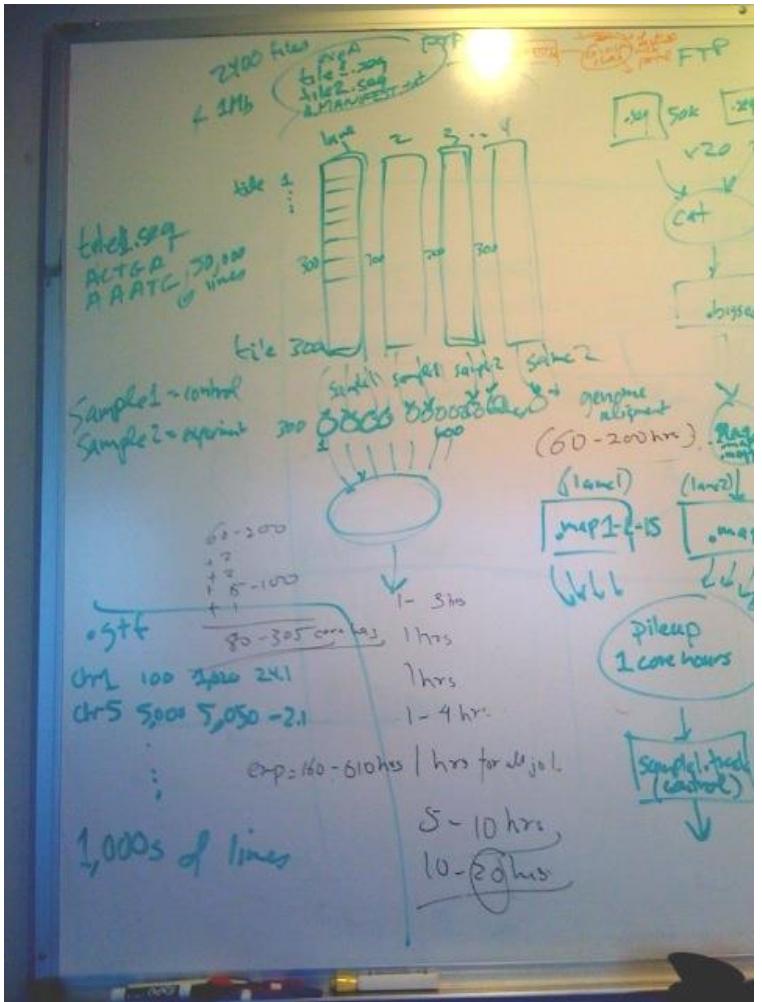
Processing instrument data in real time



- **Totally hidden from the user**
- Curated, pre-defined workflow
- Automated data transfers
- Automation of pre-processing
- Quick feedback during experiments
- Used in production at USC

Pegasus
Workflow Management System
<https://pegasus.isi.edu/>

How do workflows start?



Pegasus provides tools and APIs to create workflows using Python, Java, R and support Jupyter Notebooks

Challenge	Pegasus' Solution
Staging data	Automated data transfer to and from computations
Different storage systems	Pegasus can talk a number of protocols, including HTTP, FTP, Globus Online, HTCondor and others
Small workflow tasks	Pegasus can cluster tasks together for more efficient execution
Limited storage (edge)	Pegasus analyzes the workflow and cleans up data no longer needed
Failures during execution	Job retries, trying different data sources, workflow-level checkpoint, rescue DAGs
Have a full workflow, but some data was already computed	Pegasus can re-use that data and run only the necessary jobs
Don't know what happened during the execution	Pegasus has tools for analyzing workflow performance and help debug them, pinpointing errors

¹ Pegasus keeps track of how the result was obtained: Full Provenance, support for containers

ACCESS CI provides researchers with:

High-Performance Computing (HPC) : NCSA Delta, Purdue Anvil, TACC Frontera, SDSC Expanse, PSC Bridges2, ...

High-Throughput Computing (HTC): OSPool (Open Science Pool)

Cloud & Interactive Platforms: IU Jetstream2 for on-demand VMs and analysis

Data & Storage Services: Shared storage, data movement, OSN (Open Storage Network)

Networking & Middleware: Fast research network connections and software services

Expert Support: Consulting and training to optimize codes, workflows, and data



SCITECH



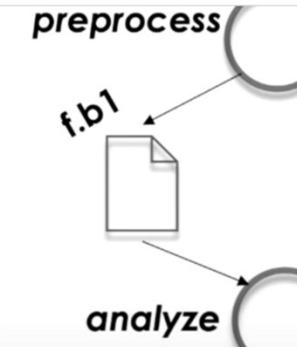
ACCESS
■ Support

Quick Links ▾ Community ▾ Knowledge Base ▾ MATCH Services ▾ Office Hours ▾ Tools ▾

⌂ / SUPPORT / TOOLS / PEGASUS WORKFLOWS

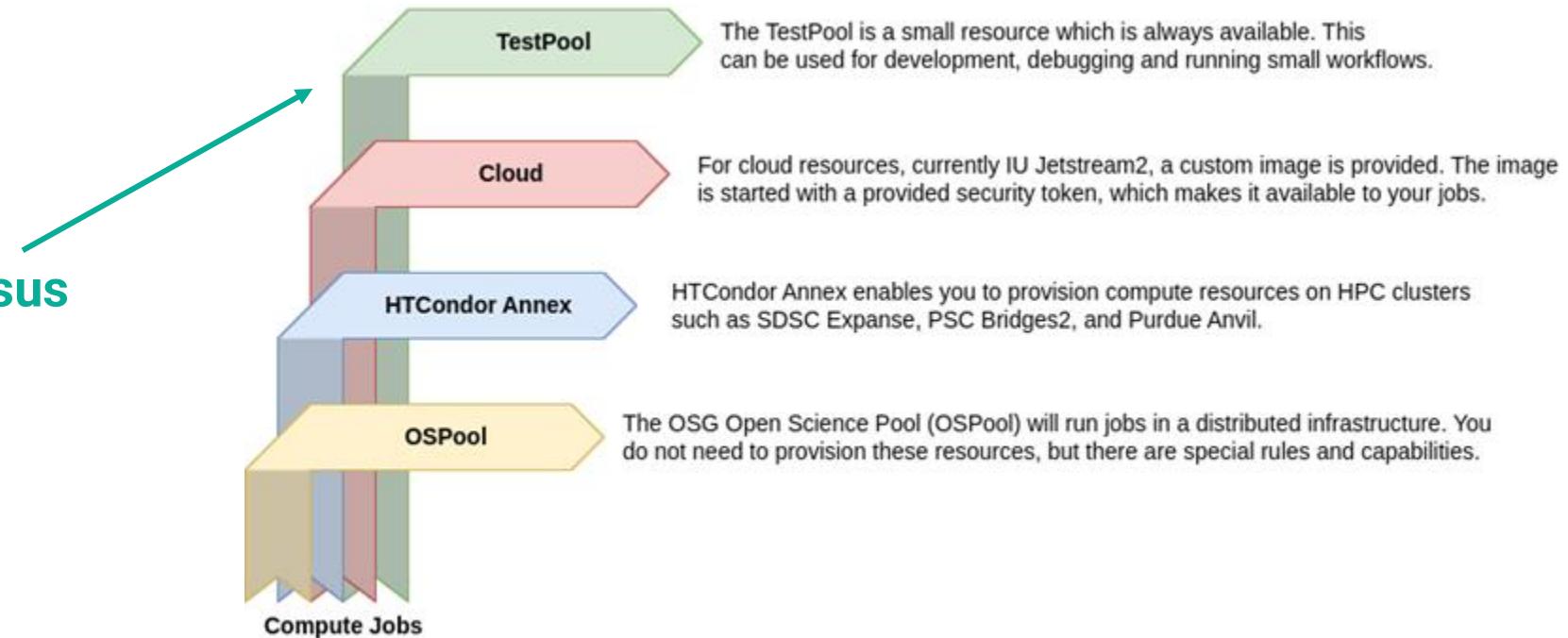
The screenshot shows a code editor with a snippet of Python code for a Pegasus workflow. Below the code is a banner for "Pegasus Easy to use hosted workflow system" with a "TRY PEGASUS" button.

```
5 fa = File("f.a")
6 fb1 = File("f.b1")
7 fb2 = File("f.b2")
8
9 preprocess_job = Job("preprocess")
10 .add_arg("-i", fa)
11 .add_input(fa)
12 .add_output(fb1)
13
14 fc = File("f.c")
15
16 analyze_job = Job("analyze")
17 .add_arg("-i", fb1)
18 .add_inputs(fb1)
19 .add_outputs(fc)
```



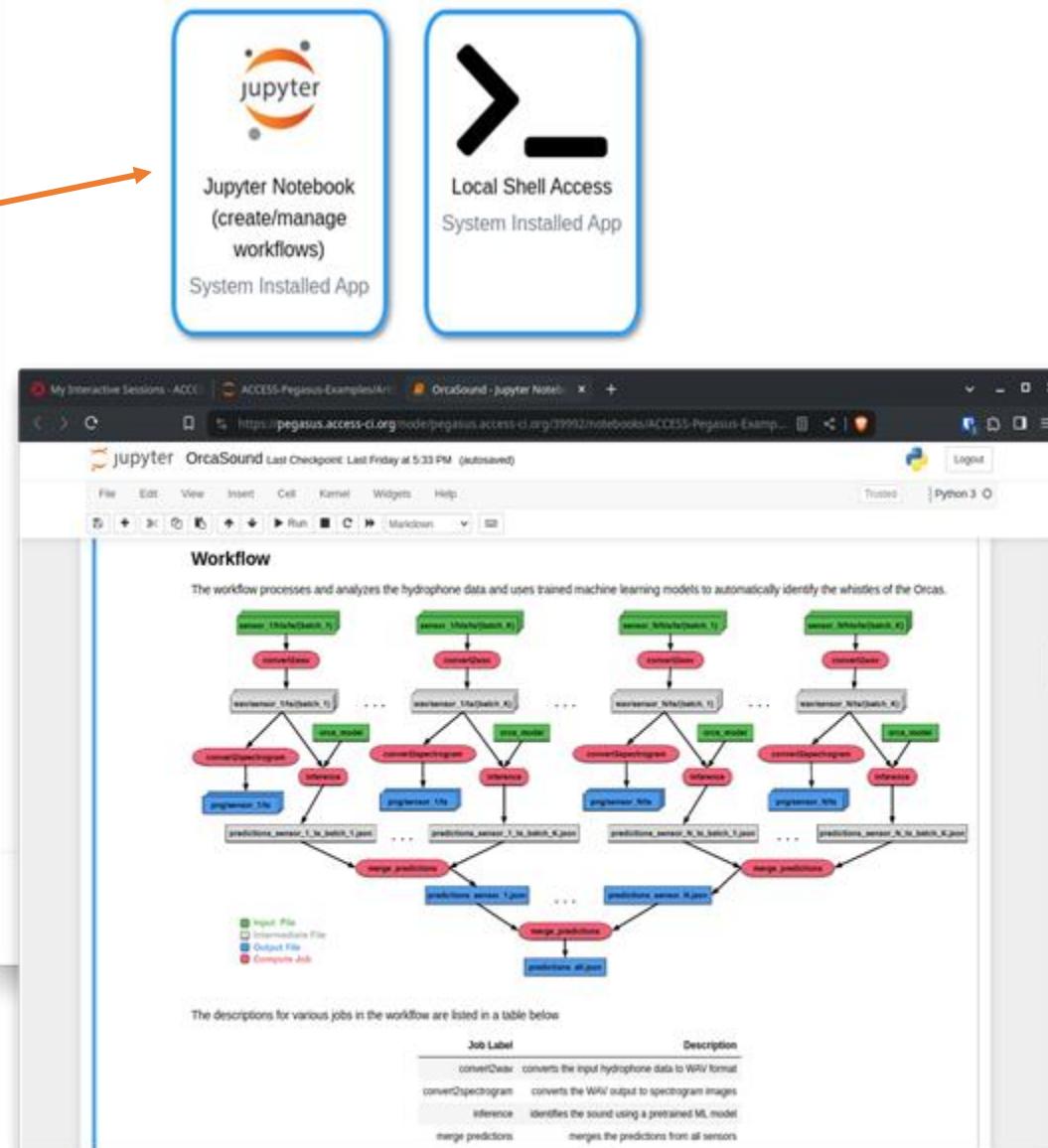
Get an account on
ACCESS and visit the
Support Page

You don't need an
allocation to try Pegasus



ACCESS Pegasus

Leveraging Open OnDemand and Jupyter Notebooks to broaden access to Pegasus' capabilities



Getting Started

Documentation

- ACCESS Pegasus Overview
 - Detailed ACCESS Pegasus documentation
 - Pegasus User Guide

Quickstart

1. Start an interactive Jupyter session. For the sample workflows, 24 hours runtime is fine. For production workflows, you might need to start a longer Jupyter session.
 2. Sample workflows are pre-installed in your home directory under the [ACCESS-Pegasus-Examples](#) directory. These include a set of self guided tutorials, as well as a set of domain specific examples.
 3. Workflows in the tutorials can be run without an allocation.
 4. For the domain examples, or your own workflows, provision resources under your allocation with HTCondor HPC Annex.

OnDemand version: 3.0.3

Jupyter-Notebook Self-Paced Tutorials and Examples

ACCESS Pegasus Apps Files Clusters ACCESS

Home / My Interactive Sessions / Jupyter Notebook (create/manage workflows, run tutorials)

Pegasus Apps

- Jupyter Notebook (create/manage workflows, run tutorials)** (circled in orange)
- ZZZ - DEVELOPERS Jupyter Notebook / Pegasus 5.1.2-dev
- ZZZ - DEVELOPERS Jupyter Notebook / Pegasus 5.2.0-dev

Jupyter Notebook (create/manage workflows, run tutorials)

This app will launch a Jupyter Notebook server

Number of hours

2

Launch

* The Jupyter Notebook (create/manage workflows, run tutorials) session data for this session can be accessed under the [data root directory](#).



Files Running Clusters

Select items to perform actions on them.

<input type="checkbox"/> 0	<input type="checkbox"/> / ACCESS-Pegasus-Examples
<input type="checkbox"/>	..
<input type="checkbox"/>	01-Tutorial-Running-a-Complete-Workflow
<input type="checkbox"/>	02-Tutorial-API
<input type="checkbox"/>	03-Tutorial-Catalogs
<input type="checkbox"/>	04-Tutorial-Debugging-Statistics
<input type="checkbox"/>	05-Tutorial-Provisioning
<input type="checkbox"/>	Artificial-Intelligence
<input type="checkbox"/>	Astronomy
<input type="checkbox"/>	Bioinformatics
<input type="checkbox"/>	images
<input type="checkbox"/>	tools
<input type="checkbox"/>	LICENSE
<input type="checkbox"/>	README.md

Jupyter-Notebook based Self-Paced Tutorials and Examples



Running a Complete Workflow

Objective: Familiarize users with the Pegasus workflow structure using an end-to-end LLM-RAG book summarization example.

Welcome to the first Pegasus tutorial notebook, which is intended for new users who want to get a quick overview of running a Pegasus workflow.

In this tutorial, a full workflow is provided. In later tutorials, we will learn how to use the API, the provided debugging/statistics tools, and how to provision resources for the workflow to execute on. The outline of those tutorials is:

- 01 - Running a Complete Workflow (this one)
- 02 - API
- 03 - Catalogs
- 04 - Debugging / Statistics
- 05 - Provisioning

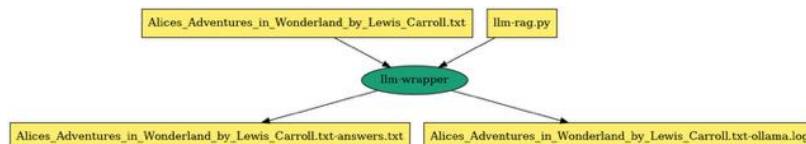
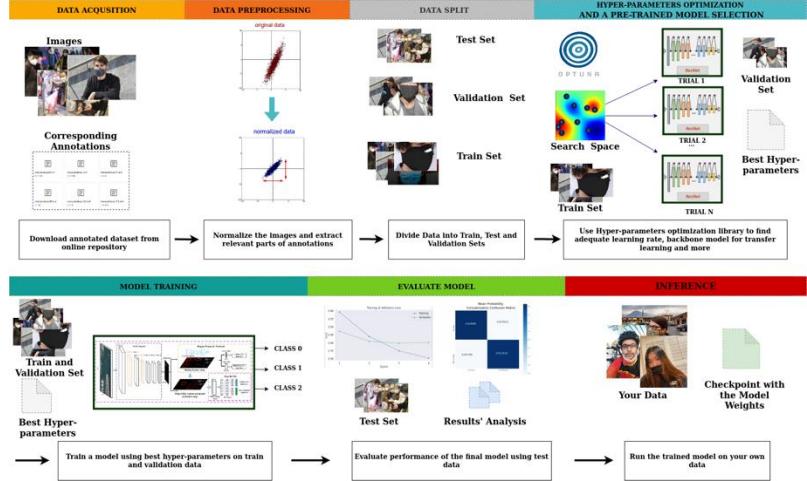
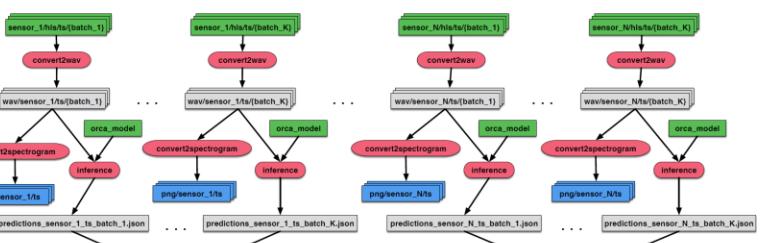
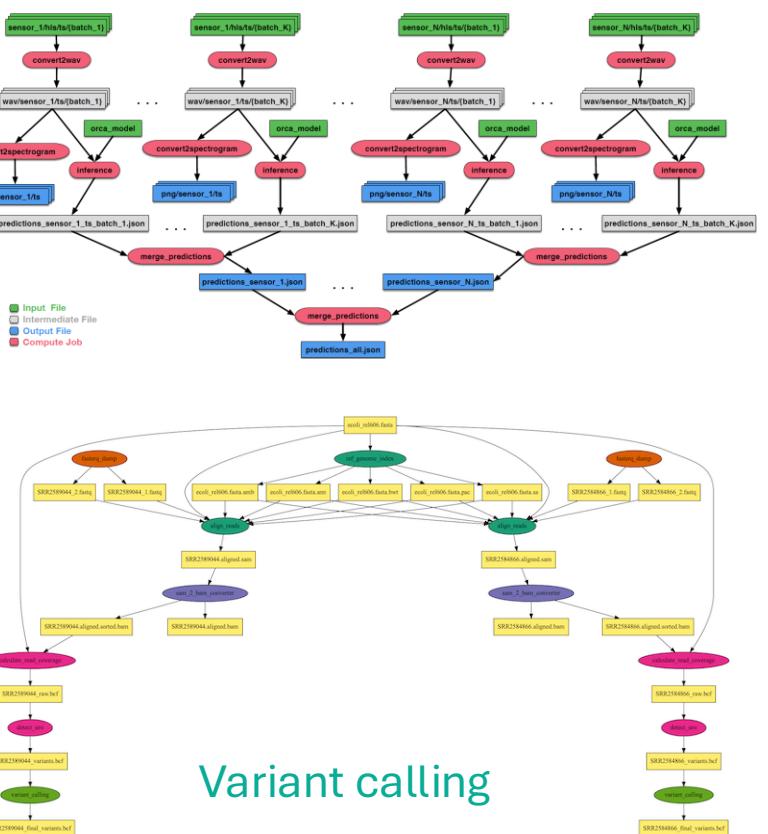
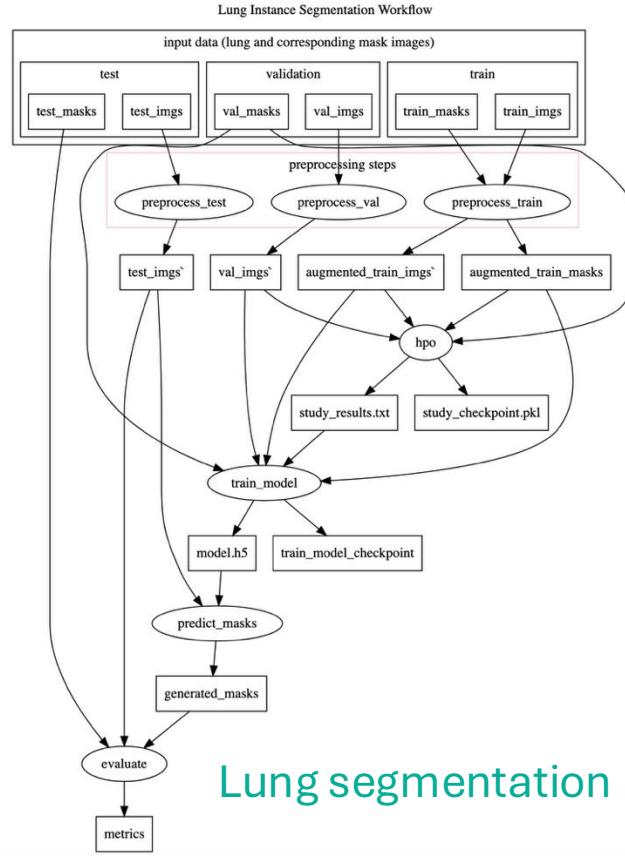
To get started, just step through the following steps.

Defining the Workflow

Pegasus workflows are created using an API, making it easy to build, manage, and run workflows in a flexible and scalable way. While it might feel unnecessary for small workflows, this approach works well for creating workflows dynamically based on data, parameters, or triggers, which is essential for automating tasks and handling large projects.

The following example is organized as a Python class. While this isn't strictly necessary, it helps keep the different parts of the workflow well-structured.

Pegasus workflows are portable, meaning you can execute the same workflow on different infrastructures at different times. To enable this portability, Pegasus uses an abstract workflow model and relies on "catalogs" to describe the execution environment, software, and input data. The Abstract Workflow description that you specify to Pegasus is portable, and usually does not contain



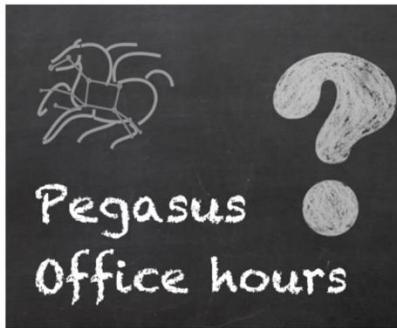
Engage with us!

- Slack channel
- Email: pegasus-support@isi.edu
- Office hours every Friday

Office Hours

Join the Pegasus team every Friday for virtual office hours at 11 AM Pacific / 2 PM Eastern.

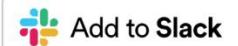
Do you have questions about workflows or need guidance on organizing and implementing them? Join our weekly office hours – designed to support both new and experienced users in learning and engaging with Pegasus. Here's what to expect:



- Tutorial walkthrough First Friday of the month
- <http://pegasus.isi.edu>

We can help you get started!

Slack



Add to Slack

We encourage you to join the Slack Workspace as it is an on-going, open forum for all Pegasus users to share ideas, experiences, and talk out issues with the Pegasus Development team. Please click the button above or ask for an invite by trying to join pegasus-users.slack.com in the Slack app, or send an email to pegasus-support@isi.edu to request an invite.

Mailing Lists

pegasus-users@isi.edu

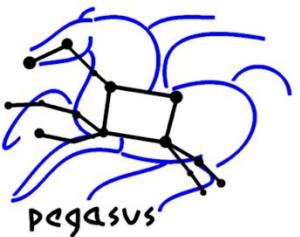
This list is the main support vehicle for Pegasus. Please note that the following subscription link requires a Google-linked email address. If you want to subscribe with non-linked email address, please contact us at pegasus-support@isi.edu and we will add you.

- [Subscribe / unsubscribe / archive](#)

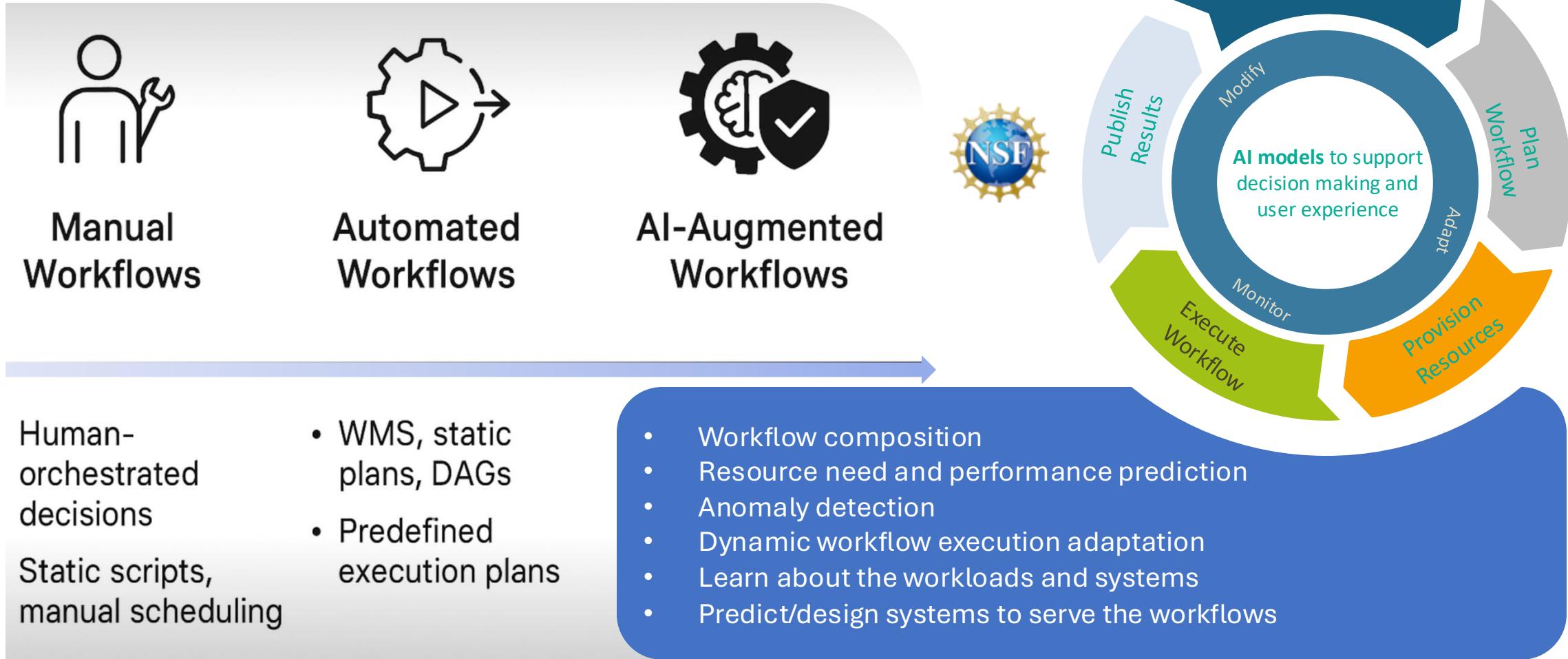
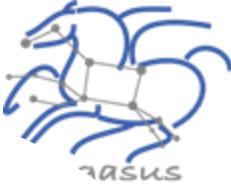
pegasus-announce@isi.edu

Messages about new releases and updates. Low traffic. Please note that the following subscription link requires a Google-linked email address. If you want to subscribe with non-linked email address, please contact us at pegasus-support@isi.edu and we will add you.

- [Subscribe / unsubscribe / archive](#)



2025 - 2030





PegasusAI Plans



- **Intelligent Resource Planning:** Uses machine learning models to predict resource needs and optimize workflow execution.
- **Adaptive Workflow Management:** Detects anomalies and performance issues in real time, automatically adjusting plans or alerting users.
- **Human-in-the-Loop Design:** Guides researchers through AI-augmented tools for workflow creation, monitoring, and debugging.
- **Scalable Across CI:** Supports execution on HPC, cloud, and edge platforms, enabling flexible deployment and broad applicability.
- **AI-Ready Data Generation:** Provides curated datasets and trained models to advance AI for scientific computing and CI research.

Next major version of Pegasus will have AI
assistant

Pegasus Analyzer

Current version of analyzer pinpoints and highlight failures. Example:

```
2025-09-25 19:10:16,985 INFO: /bin/cp -f -R -L '/home/rynge/ACCESS-Pegasus-Examples/04-Tutorial-Debugging-Statistics/bin/llm-rag.py' '/home/rynge/ACCESS-Pegasus-Examples/04-Tutorial-Debugging-Statistics/scratch/rynge/pegasus/llm-rag-books/20250925T190853+0000./llm-rag.py'  
2025-09-25 19:10:16,991 ERROR: Expected local file does not exist: /home/rynge/ACCESS-Pegasus-Examples/04-Tutorial-Debugging-Statistics/inputs/Alices_Adventures_in_Wonderland_by_Lewis_Carroll.txt
```

We are adding a Pegasus AI assistant to help interpret the errors.

Next major version of Pegasus will have AI assistant

===== Pegasus AI Analysis =====

The workflow failed due to a missing input file. The job `stage_in_local_local_0_0` encountered an error:

****Expected local file does not exist: /path/to/Alices_Adventures.txt****

****Root Cause:****

- The required input file is missing from the specified path.
- This prevents the transfer process from completing, causing the workflow to fail.

****Next Steps:****

1. Verify the file exists at the specified path.
2. Ensure the file path in the workflow configuration matches the actual location.
3. Resubmit the workflow after resolving the file issue.

The remaining unsubmitted jobs (7 total) likely depend on this staged file, so fixing this error will enable further execution.



Pegasus Workflow Management System



Automates the execution of scientific workflows across CI

- Provides a portable workflow description
- Automates repetitive and time-consuming tasks
- Reduces human error
- Increases productivity



Data Management



Error Recovery



Provenance Tracking



Heterogeneous Environments

Pegasus handles data transfers, input data selection and output registration by adding them as auxiliary jobs to the workflow

Pegasus handles errors by retrying tasks, workflow-level checkpointing, remapping and alternative data sources for data staging

Pegasus allows users to trace the history of a workflow and its outputs, including information about data sources and softwares used

Pegasus can execute workflows in a variety of distributed computing environments such as HPC clusters, Amazon EC2, Google Cloud, Open Science Grid or ACCESS

CI Compass Fellowship Program (CICF)



For Undergrads! and Faculty mentors

CI Compass
459 followers
2w ·

Applications for the Spring 2026 NSF CI Compass Fellowship program are open!

Learn more about what the program is about and offerings ...more



CICompass
Fellowship Program

"The CI Compass program not only introduced me to advanced computing tools, but also showed me how essential cyberinfrastructure is for solving the big questions in research."

Naomi Kolodisner, CICF 2025, student at the University of Arizona

Applications are open
for the Spring 2026 Program

Deadline to apply is October 17, 2025.
ci-compass.org/student-fellowships

with Naomi Kolodisner and 8 others

Anshuraj Sedai and 22 others

2 reposts



Like



Comment



Repost



Send

Goal: Broaden student participation in CI research, development, deployment, and operations

Virtual Spring Program

- Free to undergraduate students. Possibility of course credit.
- *Technical Skills Component:* Students are taught technical skills relevant to CI.
- *Data Lifecycle Component:* Students research MFs and the data lifecycle to understand the importance and context of MFs, and the related data and CI. They present their results at the end of the Spring Program.

(Optional/Invited) Summer Program

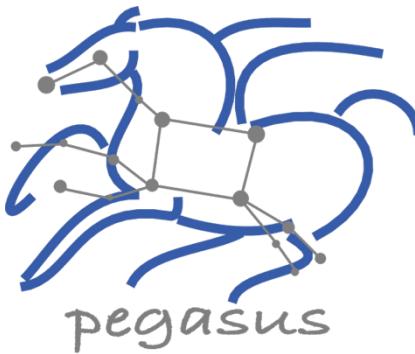
- We collaborate with MFs to provide CI-related summer projects for some of our student fellows.
- In-person or virtual, depending on the MF/project.
- Students are paid for their participation.

<https://ci-compass.org/>

**Major Facilities are large-scale NSF-funded science projects:
Telescopes, Research Vessels, Ecological and Ocean Observatories**



More Information About the Projects:



Pegasus coming to Laguna

- Pegasus: <http://pegasus.isi.edu>
- Pegasus support: <https://pegasus.isi.edu/contact/>
- PegasusAI: <http://pegasusai.io>
- ACCESS Support: <https://support.access-ci.org/>
- CI Compass: <https://ci-compass.org>

NAIRR AI Workshop at USC January 22, 2026

<https://nairrpilot.org/>

pegasus@isi.edu



Big Thanks to DOE and NSF for the support!

