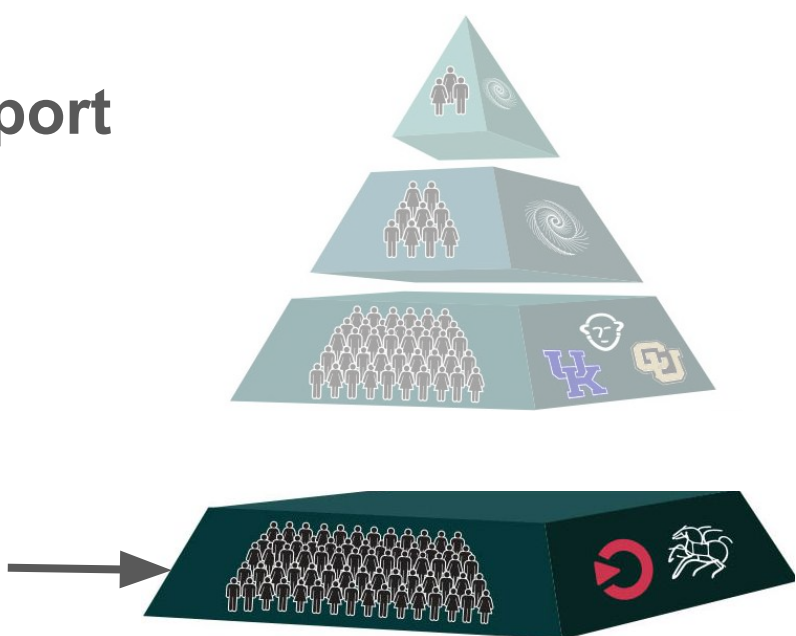


ACCESS Pegasus Goals

Provide an intuitive, accessible scientific workflow environment, catering to all user levels, with easy-to-use workflow design tools, provisioning across ACCESS resource providers, and collaborative features.

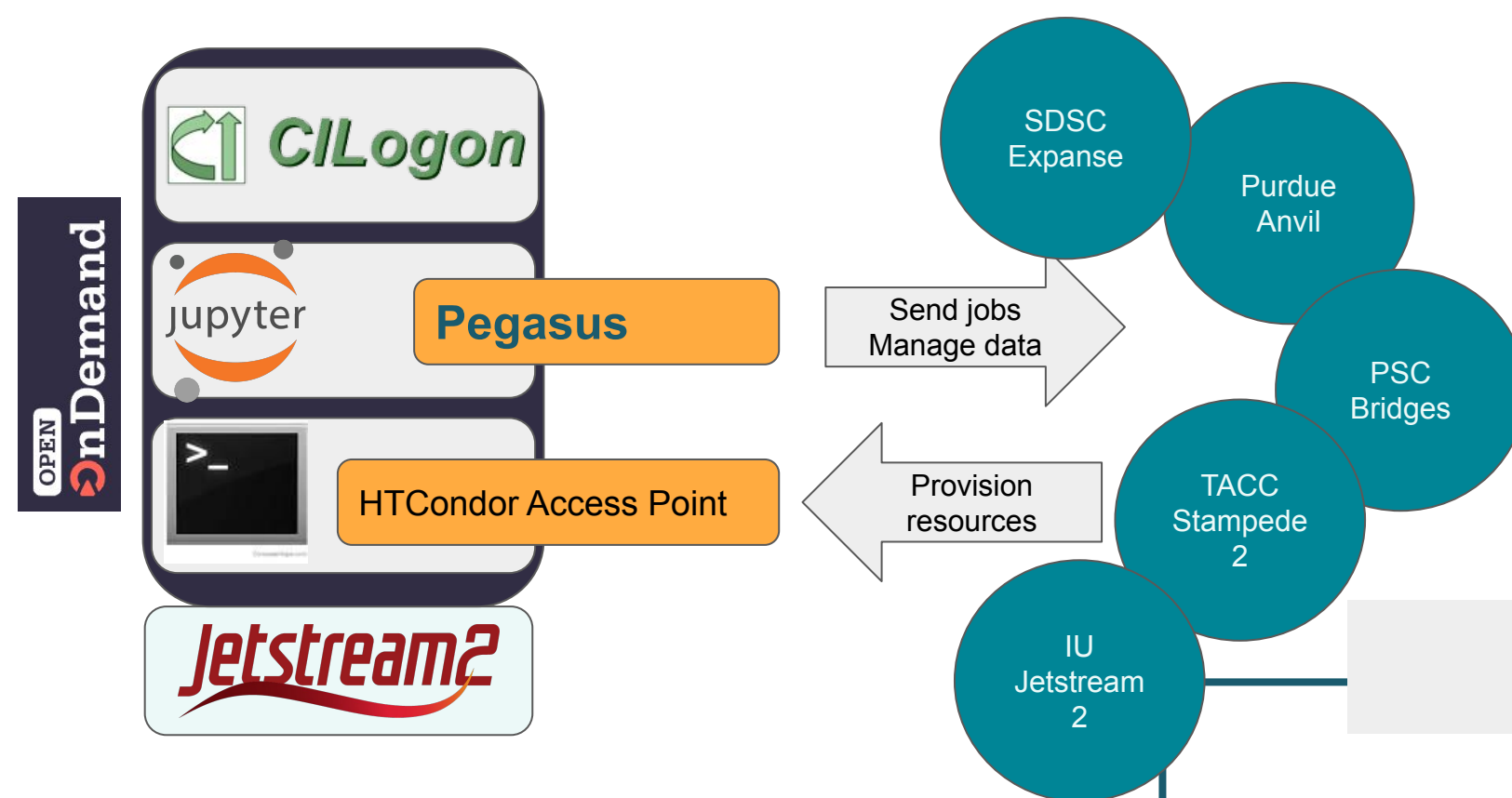
- **Democratize access:** Broaden ACCESS to sophisticated systems and tools
- **Empower scientists:** Enable individual users, in the same way as large-scale collaborations
- **Promote open science and team science:** Sharing workflows is as easy as sharing Jupyter Notebooks
- **Promote reproducibility:** Use someone else's workflow with your data and your ACCESS allocation
- **Part of ACCESS Support**

Intuitive Easy-to-Use User Interfaces (OOD, Pegasus)

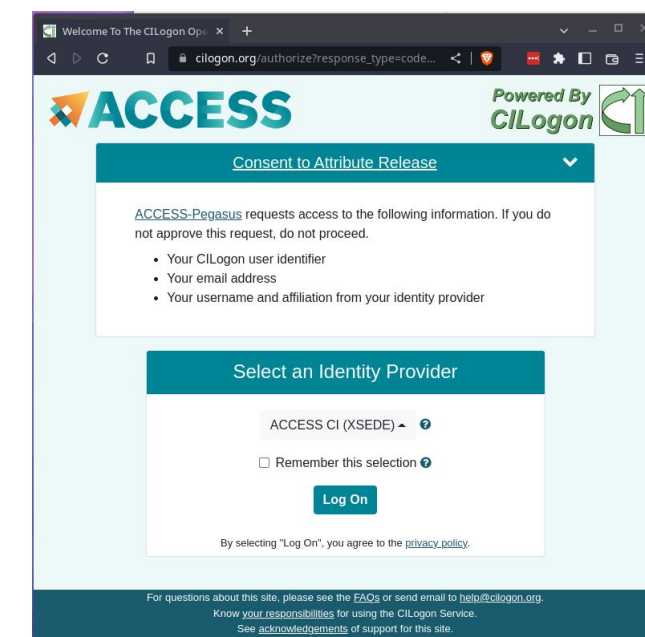


Architecture

- **Open OnDemand:** Centralized instance for users to login with their CILogon credentials. Provides Hosted Jupyter Notebooks that users use for submitting workflows.
- **Pegasus:** Simplifies complex data workflows on ACCESS resources. Provides a powerful Python API for workflow management.
- **Access Point:** A workflow submission node that includes Pegasus and HTCondor. Includes tools to provision resources.

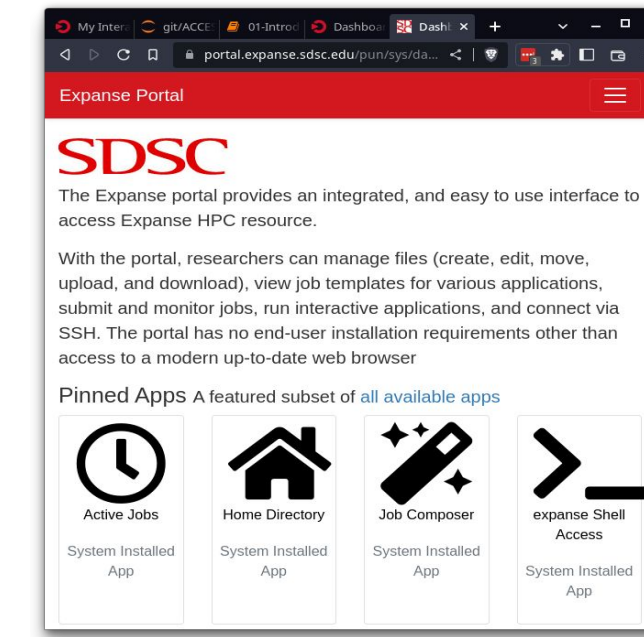


Account / Allocations



Single Sign On

All registered users with an active allocation automatically have an ACCESS Pegasus account

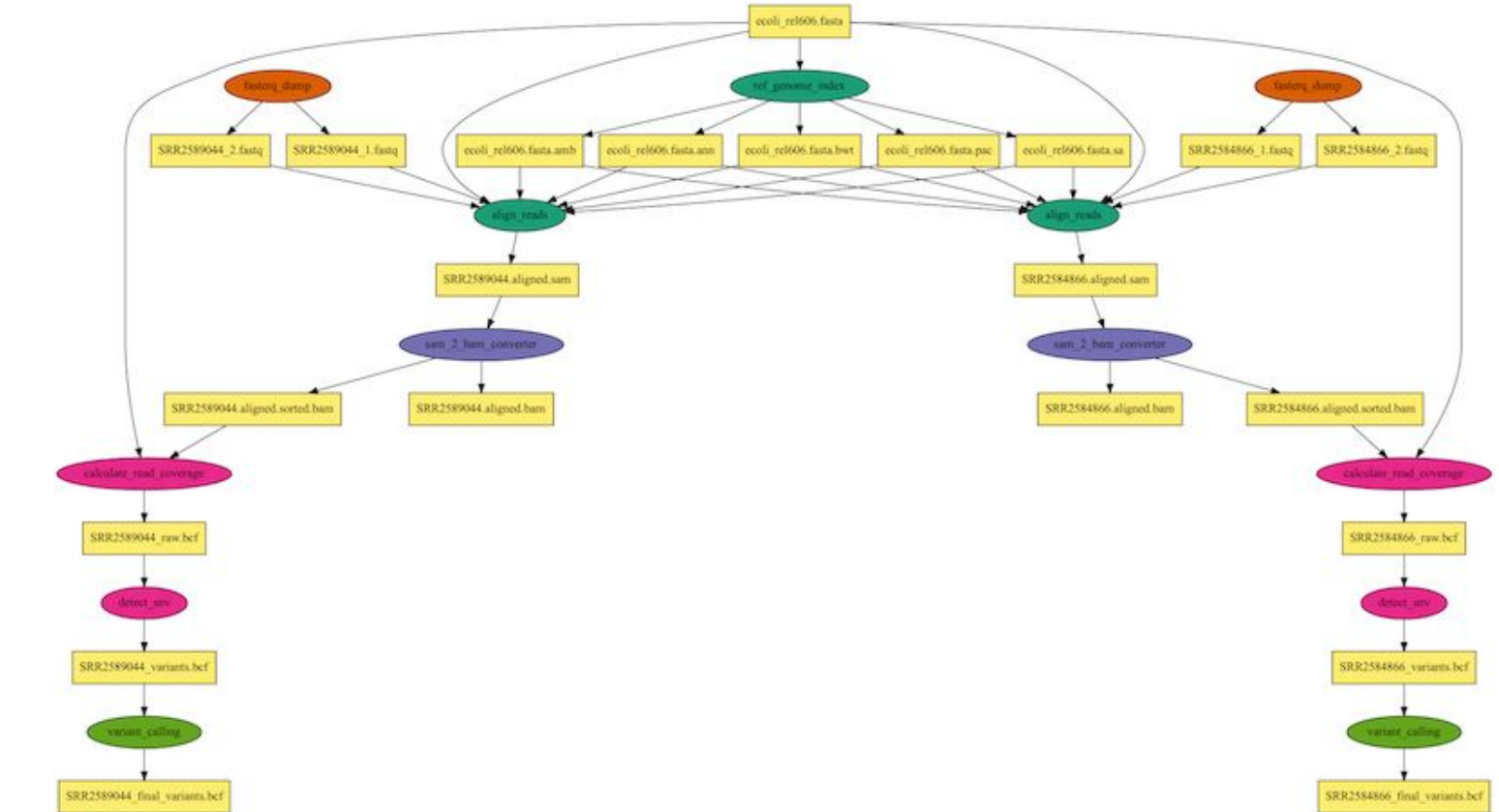


Map Allocations

One time setup for configuring the resources: Use Open OnDemand instances at resource providers to install SSH keys and determine local allocation IDs

Jupyter Training Notebooks

- **Self Guided Tutorial:** introduction to API, debugging, CLI and a complete workflow example
<http://access.pegasus.isi.edu>
- **Additional Example Workflows - Variant Calling:** Adapted from the Data Carpentry Lesson on Data Wrangling and Processing for Genomics
<https://datacarpentry.org/wrangling-genomics/04-variant-calling/index.html>
Can run simultaneously across multiple ACCESS resources, and uses OSN for data storage

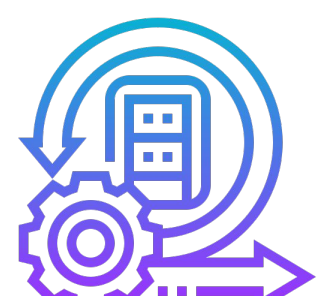


Why Pegasus for Workflows?



Heterogeneous Environments

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



Data Management

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



Provenance Tracking

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



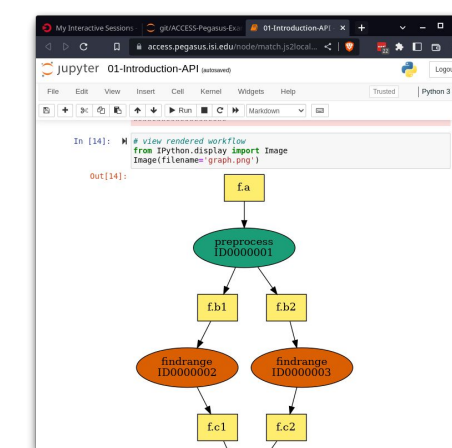
Error Recovery

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

Pathway

1. Create the workflow

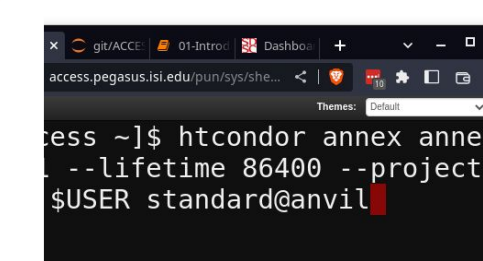
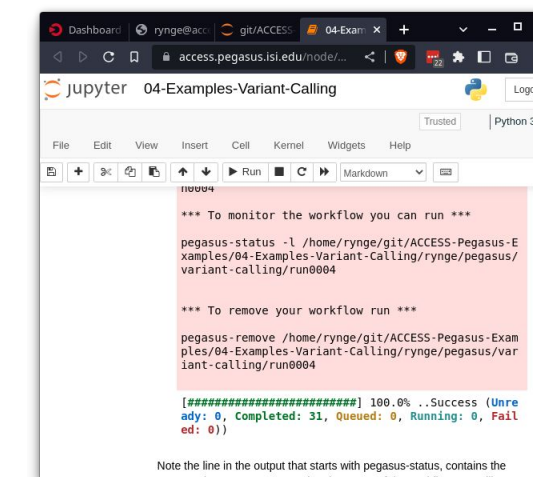
Use Pegasus API in Jupyter Notebook or use our examples



Submit your workflow for execution

3. Staying Informed

You can follow the workflow execution within the notebook or in the terminal

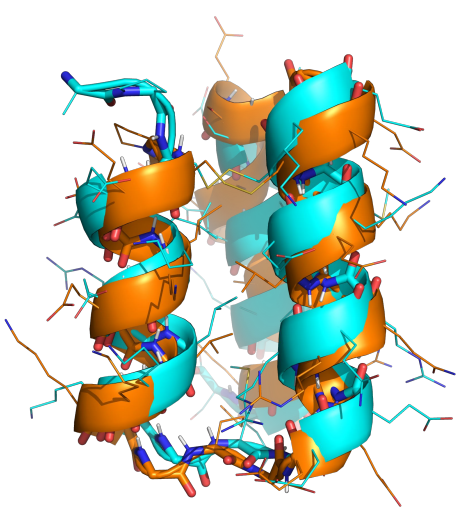


2. Provision compute resources

Use HTCondor Annex tool to provision pilot jobs on your allocated ACCESS resources

Recently Added

Rosetta - modeling and analysis of protein structures



AlphaFold - accurately predicts 3D models of protein structures

Orca Sound - sound identification of Orca whales

Mask Detection - detecting face masks in a group of people

Lung Segmentation - predicting lung masks using Chest-X rays