Pegasus: Distributed Workflow Management Research and Software in Support of Science

*Ewa Deelman, **Karan Vahi**, Mats Rynge, Rajiv Mayani*

USC Information Sciences Institute

USC Viterbi School of Engineering

Pegasus WMS

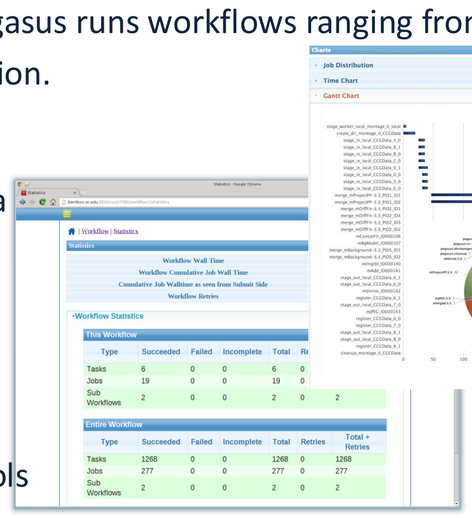
Pegasus is a system for mapping and executing abstract application workflows over a range of execution environments.

The same abstract workflow can, at different times, be mapped different execution environments such as ACCESS, PATh, OSG OS Pool, commercial and academic clouds, campus grids, and clusters.

Pegasus can easily scale both the size of the workflow, and the resources that the workflow is distributed over. Pegasus runs workflows ranging from just a few computational tasks up to 1 million.

End to End automatic checksumming of workflow data to ensure data integrity.

Stores static and runtime metadata associated with workflow, files and tasks. Accessible via command line tools and web-based dashboard.



Real-time Monitoring

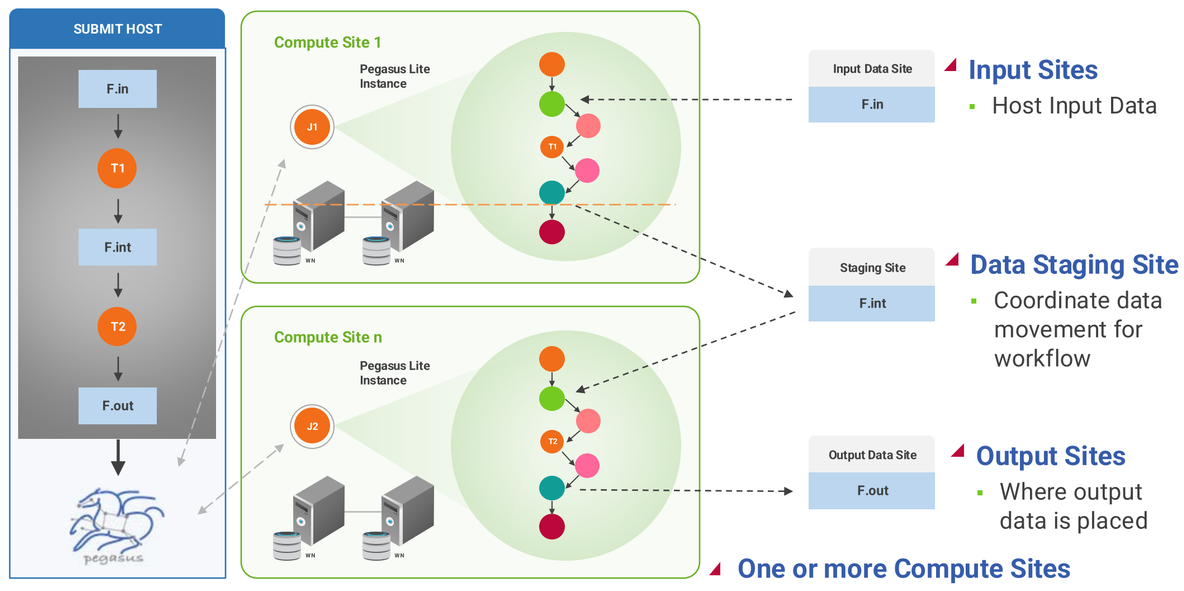
Reporting

Debugging

Troubleshooting

RESTful API

Pegasus Deployment



Input Sites

- Host Input Data

Data Staging Site

- Coordinate data movement for workflow

Output Sites

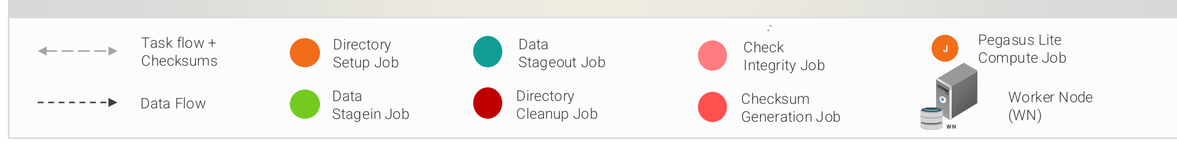
- Where output data is placed

Workflow Submit Node

- Pegasus WMS
- HTCondor

One or more Compute Sites

- Compute Clusters
- Cloud
- ACCESS CI
- PATh, OSG OS Pool



Applications using Pegasus

Astronomy and Physics

- Pegasus powered workflows help detect gravitational waves.
- XENONnT workflows for searching Dark Matter.
- Event Horizon Telescope for creating images of Black Holes.
- Galactic Plane workflow generates mosaics for astronomy surveys.

Seismology

- USC CyberShake workflows for seismic hazard analysis of LA basin.

Ecology

- Integrated Assessment Models to project impact of policy scenarios On socio environmental systems.
- Predicting Flash Floods in Dallas FortWorth Metroplex.
- Producing Titan2D hazard maps that display the probability of a volcanic flow depth reaching a critical height following a premonitor volcanic eruption event.

Microscopy

- Investigation of Strong Nuclear Force using gamma ray spectroscopy
- Cryo-EM Electroscopy for 3D reconstruction of biological samples

Bioinformatics

- Quality control workflows for data submissions to NRGR repository
- Genomic Variant Calling Workflow

Others http://pegasus.isi.edu/applications

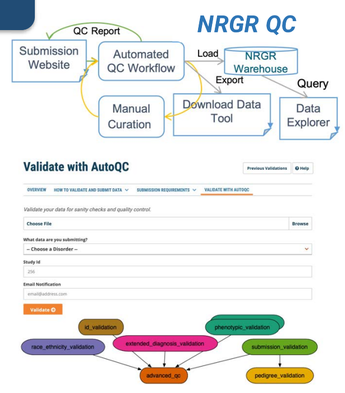
XENONnT

Two Workflows

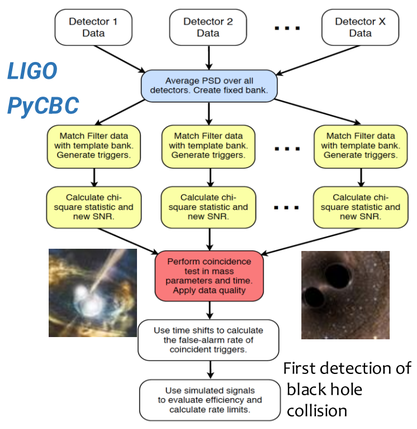
Monte Carlo simulations and the main processing pipeline.

- Workflows execute across Open Science Grid (OSG) & **European Grid Infrastructure**
- Rucio for data management
- MongoDB instance to track science runs and data products.

NRGR QC

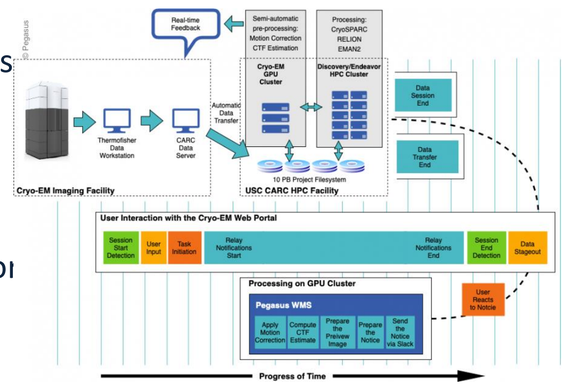


LIGO PyCBC

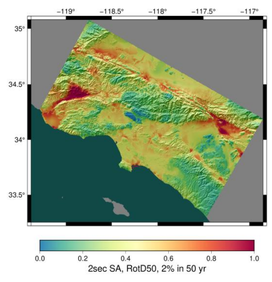


First detection of black hole collision

CryoEM Processing at USC



Southern California Earthquake Center Cybershake



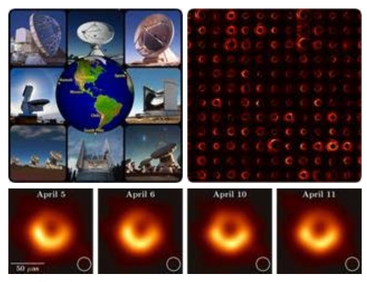
Mix of MPI and single-core jobs, mix of CPU, GPU codes.
Recent CyberShake Study 24.8, which used Pegasus to submit approx. 28,000 jobs across DOE Leadership class systems Frontier and Frontera .
Pegasus automatically managed 1 PB of data and staged 9 million output files totaling 36 TB back to archival to USC.
Supported since 2005: changing CI, x-platform execution

Event Horizon Telescope

Bringing Black Holes into Focus

8 telescopes: 5 PB data

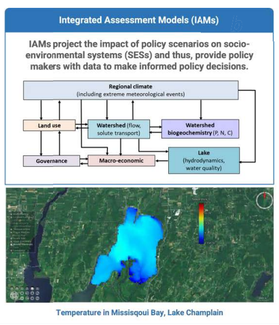
60 simulations: 35 TB data



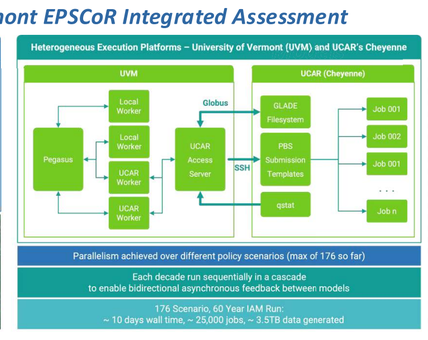
First images of black hole at the center of the M87 galaxy

Improve constraints on Einstein's theory of general relativity by 500x

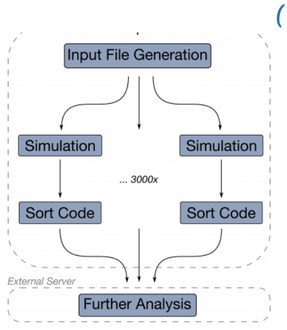
Vermont EPSCoR Integrated Assessment



Temperature in Missisquoi Bay, Lake Champlain



Gamma Ray spectroscopy (



Access to National CI

Pegasus is supported workflow tool on important National CI that provides researchers easy and free access to computing services.

ACCESS CI

https://access-ci.org

ACCESS is a program established and funded by the National Science Foundation to help U.S. based researchers and educators, with or without supporting grants, to utilize the nation's advanced computing systems and services – at no cost. Single entry point for over 20 compute, cloud, storage, and networking systems.

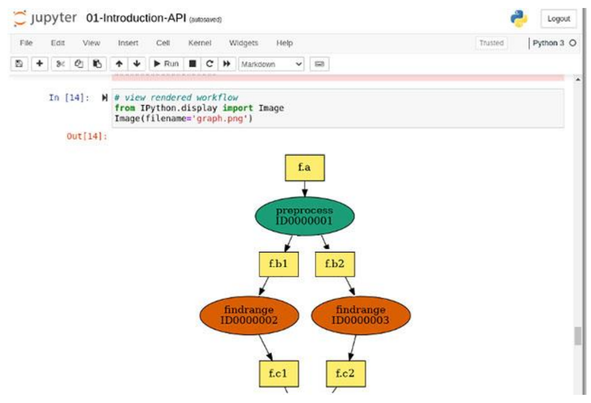
PATh and OSG Open Science Pool

https://path-cc.io

The PATh Facility funded by the NSF Office of Advanced Cyberinfrastructure and provides free access to large-scale distributed High Throughput Computing systems for US based open science projects.

ACCESS Pegasus

ACCESS Pegasus is a hosted workflow management system which allows users to construct, run, and debug workflows from a Jupyter Notebook. Perform simple interactions on the command line. Work across environments and get started quickly with sample workflows using a Python API.



Some reasons you should consider using a system like Pegasus WMS:

REPRODUCIBILITY
Document and reproduce your analyses, ensuring their validity, with scientific workflows.

AUTOMATION
Automate repetitive and time-consuming tasks and reduce your workload.

SCALABILITY
Handle large data sets and complex analyses and take on bigger research problems.

REUSABILITY
Build libraries of reusable code and tools that can be adapted by other researchers.

Support and Downloads

Documentation: https://pegasus.isi.edu/documentation/

Tutorials: Jupyter Notebooks in Docker containers

Email: pegasus-support@isi.edu

Pegasus Users Slack: https://pegasus.isi.edu/contact/

Office Hours: https://pegasus.isi.edu/office-hours/

Downloads & Usage Since 2013

Workflows: 2,700,552 **Jobs:** 2,028,572,901

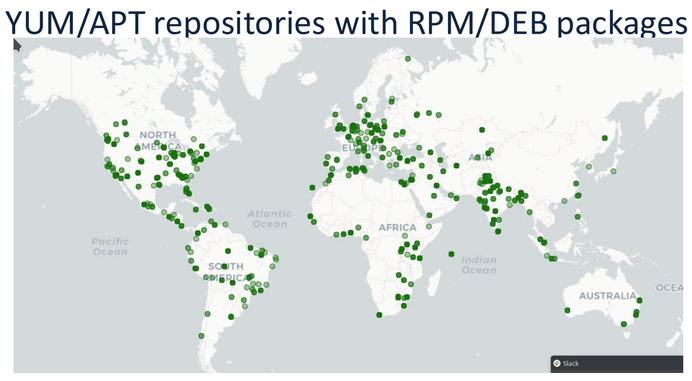
Tasks: 7,310,416,875

Release Schedule

- Major Release every 9 months. Minor releases every 4 months.
- Continuous Integration Testing with GitLab

Download Options

- Source Code and Issues publicly hosted on GitHub
- Binary packages for Linux and MAC
- YUM/APT repositories with RPM/DEB packages



Downloads

Supported by NSF Office of Advanced Cyberinfrastructure Grant: 2138286

NSF · ACCESS