Title: Scientific Data Processing with the Pegasus Workflow Management System

Presenters: Karan Vahi (vahi@isi.edu)
USC Information Sciences Institute, Marina Del Rey, CA, USA
Mats Rynge (rynge@isi.edu)
USC Information Sciences Institute, Marina Del Rey, CA, USA

Expected Duration: 2 hours

Target Audience: Scientific Domain Application Developers, Application Scientists, System Architects doing large-scale scientific analysis.

Tutorial Format: Talk and guided hands on exercises

General description of tutorial content

Workflows are a key technology for enabling complex scientific computations. They capture the interdependencies between processing steps in data analysis and simulation pipelines as well as the mechanisms to execute those steps reliably and efficiently. Workflows can capture complex processes, promote sharing and reuse, and also provide provenance information necessary for the verification of scientific results and scientific reproducibility.

Pegasus (https://pegasus.isi.edu) is being used in a number of scientific domains doing production grade science. In 2016 the LIGO gravitational wave experiment used Pegasus to analyze instrumental data and confirm the first detection of a gravitational wave. The Southern California Earthquake Center (SCEC) based at USC, uses a Pegasus managed workflow infrastructure called Cybershake to generate hazard maps for the Southern California region. In 2019, SCEC completed the largest CyberShake study to date, producing the first physics-based PSHA maps for the Northern California region. Using Pegasus, they ran CyberShake workflows on three systems: HPC at the University of Southern California (USC), Blue Waters at the National Center for Supercomputing Applications (NCSA), and Titan at the Oak Ridge Leadership Computing Facility (OLCF), consuming about 120 million core hours of compute time. Pegasus orchestrated the execution of over 18,000 remote jobs using Globus GRAM, rvGAHP, and Condor Glideins, and transferred over 150 TB between the three systems. Pegasus is also being used in astronomy, bioinformatics, civil engineering, climate modeling, earthquake science, molecular dynamics and other complex analyses.

In 2020, we released Pegasus 5.0 that is a major improvement over previous releases. Pegasus 5.0 provides a brand new Python3 workflow API developed from the ground up so that, in addition to generating the abstract workflow and all the catalogs, it now allows you to plan, submit, monitor, analyze and generate statistics of your workflow. Since 2022, Pegasus has been a key part of the ACCESS support strategy (https://support.access-ci.org/pegasus).

Primary learning objectives

The goal of the tutorial is to introduce application scientists to the benefits of modeling their pipelines in a portable way with use of scientific workflows with application containers. We will examine the workflow lifecycle at a high level and issues and challenges associated with various steps in the workflow lifecycle such as creation, execution and monitoring and debugging. Through hands-on exercises in a hosted Jupyter notebook environment, we will describe an application pipeline as a Pegasus workflow using Pegasus Workflow API and execute the pipeline on distributed computing infrastructures. The attendees will leave the tutorial with knowledge on how to model their pipelines in a portable fashion using Pegasus workflow and run them on varied computing environments. The tutorial will also cover how to bundle application codes into a container and use them in workflows.

Why the topic is relevant to conference attendees

We believe this tutorial would be relevant to ADASS'23 as we address the importance of formal descriptions of the computations, data management, and control over the execution environment (containers), in a world where both computations and data grow in complexity. The training will serve to provide foundational knowledge for researchers looking to do large scale data processing on local HPC clusters, or resources part of national cyberinfrastructure such as OSG and ACCESS.

Brief outline of the topics to be covered

The tutorial involves hands-on exercises in a Jupyter notebooks environment. The tutorial presented will be a mix of presentation slides and hands on tutorial exercises using Pegasus.

We will give an overview of scientific workflows, followed by an introduction to Pegasus WMS covering basic concepts and system architecture. We will then have hands-on exercises, walking attendees through composing, submitting and monitoring their workflows using the Pegasus Python Workflow API. Users will also learn how to debug workflows when failures happen. By means of exercises, we will illustrate why it makes sense to package application code into containers and then have them compose and execute an astronomy mosaicing workflow using Montage toolkit and containers. We will end the tutorial with slides on advanced issues in scientific workflows just such as data integrity etc.

Detailed agenda of the tutorial (Total Time 2 hours)

1. **Introduction (10 minutes)**
    a. Scientific Workflows
    b. Pegasus Overview
    c. Success Stories
2. **Pegasus Overview (10 minutes)**
    a. Basic Concepts
    b. Features
    c. System Architecture
3. **Hands-on Tutorial - (30 minutes)**
   Jupyter notebooks:
   https://github.com/pegasus-isi/pegasus/tree/master/tutorial/docker/notebooks
    a. Composing and Submitting a Simple Workflow - *15 minutes*
    b. Debugging a failed workflow - *10 minutes*
    c. Command line tools for monitoring - *5 minutes*
4. **Break (10 minutes)**
5. **Hands-on Tutorial Continued - (30 minutes)**
   Jupyter notebooks:
   https://github.com/pegasus-isi/pegasus/tree/master/tutorial/docker/notebooks
    a. Using Containers to package application and their dependencies - *10 minutes*
    b. Compose an astronomy mosaicing workflow using Montage toolkit and containers - *20 minutes* ***
6. **Advanced Topics - (25 minutes)**
    a. Data Staging
    b. Data Integrity
    c. Hierarchical Workflows
    d. Checkpointing
    e. Metadata
7. **Conclusion and Summary (5 minutes)**

*** *5.b will be based on our ADASS21 demo notebook*
https://github.com/pegasus-isi/ADASS21-montage-docker-image/blob/main/PegasusMontage/PegasusMontage.ipynb

Tutorial Materials

- Jupyter notebooks:
  https://github.com/pegasus-isi/pegasus/tree/master/tutorial/docker/notebooks
- Tutorial Slides -
  https://pegasus.isi.edu/tutorial/escience22/escience22-pegasus5-tutorial-final.pdf
- Self Guided Tutorial - https://pegasus.isi.edu/documentation/user-guide/tutorial.html

Pegasus Information

- Pegasus Website - https://pegasus.isi.edu
- Documentation - https://pegasus.isi.edu/documentation

Previous Tutorials

We have given many workflow tutorials over the years and at major conferences such as PEARC, EScience and CCGrid. Pegasus training is regularly included in various training workshops for both OSG and ACCESS. In the past, Pegasus has been presented as a focus demo at various ADASS conferences, with most recent being ADASS 2021 where a Montage pipeline was parallelized using Pegasus (https://github.com/pegasus-isi/ADASS21-montage-docker-image) .

A comprehensive list of previously presented tutorials and venues can be found at: https://pegasus.isi.edu/documentation/tutorials/

Participant Requirements

The participants will be expected to bring in their own laptops with the following software installed: Web Browser, PDF reader. We assume familiarity with working in a Linux environment. The laptops should be able to connect to the internet over WI-FI. Attendees will use a hosted Jupyter notebook environment, and thus only a web browser is required.

Infrastructure Requirements

Presenter Laptops

The presenters will use their own laptops. A power outlet should be provided at the presenters desk. USB-C to HDMI video connectors are requested for projection.

Participants

Participants will be required to bring their own laptops. However, since the duration of the proposed tutorial is 2 hours, it is recommended that power strips be provided at the participants tables.

Wi-Fi

The hands-on components of the tutorial require Wi-Fi access for both the presenters and the participants to complete the exercises.