

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

Ewa Deelman, USC Information Sciences Institute, deelman@isi.edu Miron Livny, University of Wisconsin, Madison, miron@cs.wisc.edu



Pegasus WMS

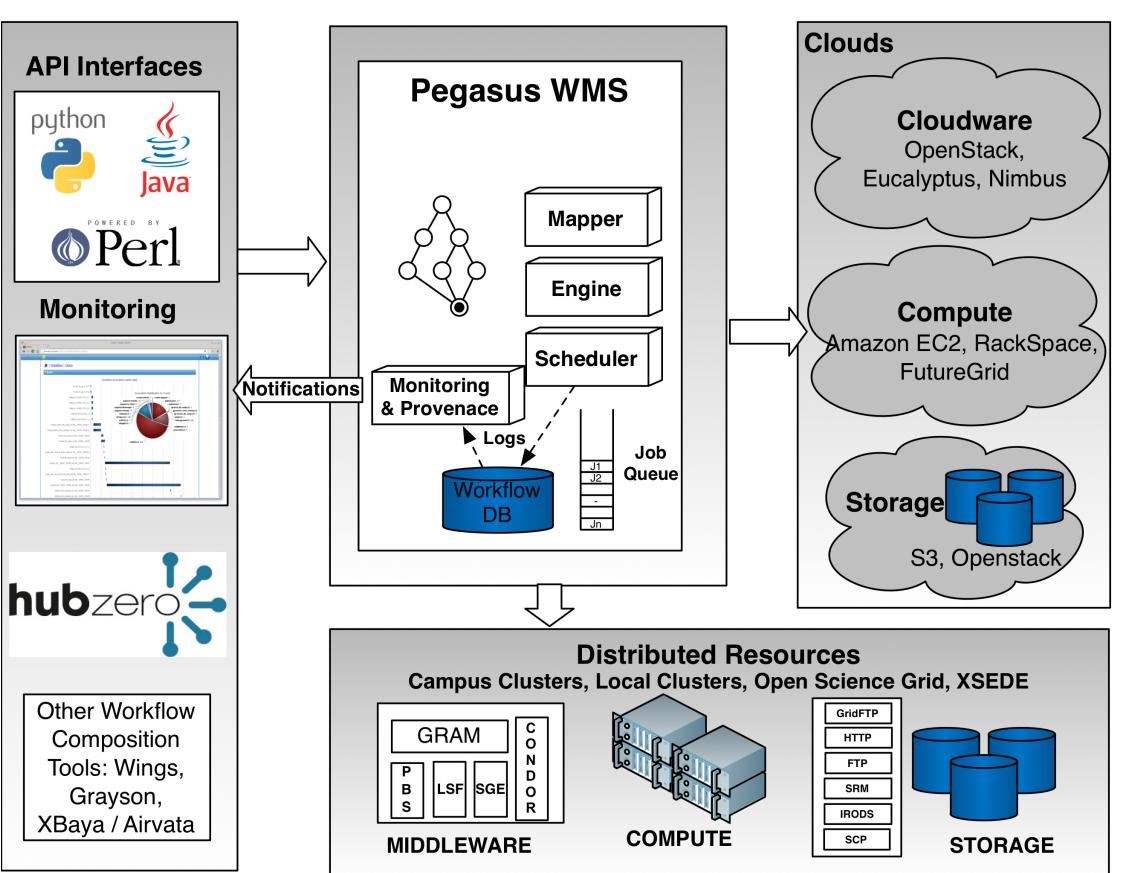
- Pegasus is a system for mapping and executing abstract application workflows over a range of execution environments.
- The same abstract workflow can, at different times, be mapped different execution environments such as XSEDE, OSG, commercial and academic clouds, campus grids, and clusters.
- Pegasus 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.
- Pegasus Workflow Management System (WMS) consists of three main components: the Pegasus Mapper, HTCondor DAGMan, and the HTCondor Schedd.

Canonical Workflow Example Capabilities Higlighted Data Management Data Reuse Job Clustering **Cross Site Runs LEGEND Unmapped Job** С Compute Job mapped to a site A Compute Job mapped to a site B Stage-in Job Stage-Out Job f.d Registration Job Create Directory Job Cleanup Job **Abstract Workflow** Final Executable Workflow Clustered Job

Optimized data transfers and support for many protocols

- Automatic data cleanup to reduce data footprint
- Retries computations in case of failures
- Workflow-level checkpointing through data reuse and DAGMan
- Command line and web-based monitoring and debugging tools to support large workflows
- workflow and task-level notifications hubzero
- Stores provenance of data used and produced, and executable invocations
- Pegasus-MPI-Cluster enables finegrained task graphs to be executed efficiently on HPC resources

System Architecture and Features



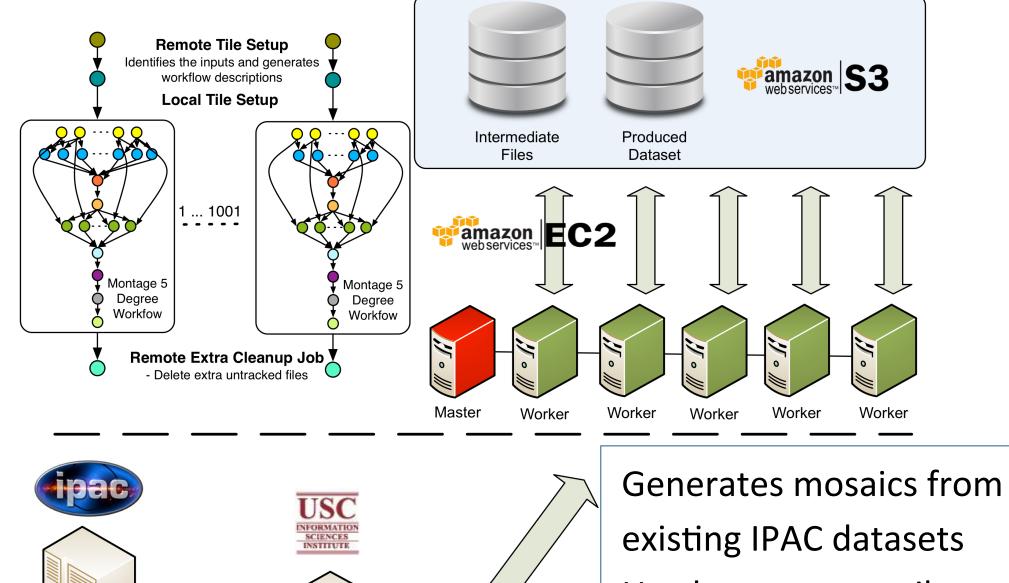
Software Availability Download Options

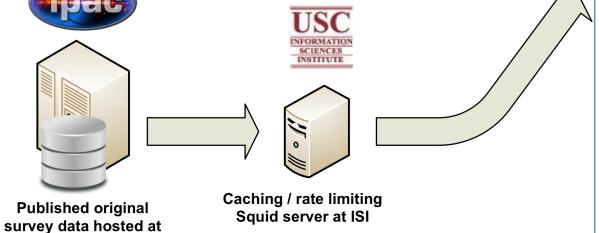
- YUM repository with RPM packages
- APT repository with DEB packages
- Binary packages for Linux and Mac

Documentation / Training Materials

- User Guide
- Quickstart Guide
- Tutorial with Virtual
 Machine
- Software Carpentry Module

Montage Galactic Plane





Generates mosaics from existing IPAC datasets
Used to generate tiles
360 x 40 around the galactic equator

- Tiles are 5° x 5° with 1° overlap with neighbors
- One workflow for each of 17 bands (wavelengths)
- Each workflow uses 3.5TB of input imagery (1.6 million files)
- Each workflow consumes 30K CPU hours and produces
 1,001 tiles in FITS format -published publically in Amazon S3

Applications using Pegasus

Astronomy and Physics:

- Galactic Plane workflow generates mosaics for astronomy surveys
- LIGO workflows help detect gravitational waves
- Periodogram workflows help detect extra solar planets

Seismology:

- CyberShake workflows for seismic hazard analysis of LA basin
- Broadband workflows for accurate predictions of ground motions

Bioinformatics:

- Quality control workflows for data submissions to NRGR repository and PAGE consortium.
- Imputation workflows on PAGE data
- Workflows for Genome and Transcriptome free analysis of RSEQ
- Brain span workflows help study gene expression in the brain
- RNA Sequencing workflows for generating Cancer Genome Atlas
- SIPHT workflows to predict sRNA encoding genes in bacteria
- Proteomics workflows for mass spectrometry based proteomics

Others:

http://pegasus.isi.edu/applications



