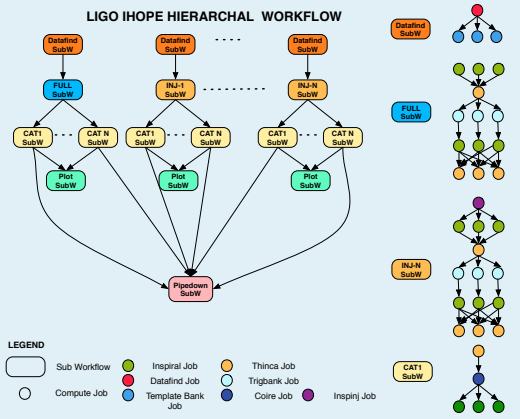
SSI: Distributed Workflow Management Research and Software in Support of Science

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

Pegasus WMS: http://pegasus.isi.edu

- Pegasus is a system for mapping and executing abstract application workflows over a range of execution environments.
- The output is an executable workflow that can be executed over a variety of resources (Clouds, XSEDE, OSG, Campus Grids, Clusters, Workstation)
- Pegasus can run workflows comprising of millions of tasks.
- Pegasus WMS consists of three main components: the Pegasus mapper, Condor DAGMan, and the Condor schedd.
- The mapping of tasks to the execution resources is done by the mapper based on information derived from static and/or dynamic sources. Pegasus adds and manages data transfer between the tasks as required.
- DAGMan takes this executable workflow and manages the dependencies between the tasks and releases them to the Condor schedd for execution.

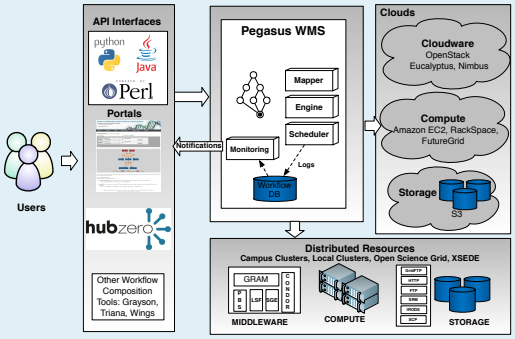


Large Scale Hierarchal Workflows

- Nodes in a workflow can be tasks or another workflow (DAX).
  - Scales up-to order of millions of tasks
- Each sub workflow is mapped when it is ready for execution.

Composing Workflows

- Workflows are expressed in DAX (Directed Acyclic graph in XML)
- DAXes can be generated using Java, Perl or Python API's
- Support for higher level workflow composition tools like Wings, Triana
- Integrated with HUBZero
- Scientists can use application-specific portals such as CGSMD
  - http://portal.nimhgenetics.org



Software Availability

Download Options

- YUM repository with RPM packages
- APT repository with DEB packages
- Binary packages for various linux and Mac platforms.

Training Materials

- Quickstart Guide
- Virtual Machine based Tutorial.

Prepackaged Application VM's

- RNASeq Pegasus VM available at http://genomics.isi.edu/rnaseq

Applications Using Pegasus

Astronomy and Physics

- Galactic Plane for generating mosiacs from the Spitzer Telescope
- LIGO workflows for detecting gravitational waves.
- Periodogram Workflows for detecting extra solar planets .

Earthquake Sciences

- Cybershake workflows for seismic hazard analysis for LA Basin.
- Broadband worklfows for accurate predictions of ground motions.

Bioinformatics

- Brain span workflows to find where in the brain a gene's expressed
- Workflows to compute RNA Seq 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

Pegasus Features

- Clustering of small tasks into large clusters for performance.
- Optimized data transfers and ability to use different protocols.
- Data reuse in case intermediate data products are available
  - workflow-level checkpointing
- Automatic data cleanup
  - reduces data footprint
- Support for Workflow and Task level notifications
- Workflow Progress can be tracked through a database.
- Stores provenance of data used, produced and which software was used with what parameters
- Retries computations in case of failures.
- Monitoring and Debugging tools to debug large scale workflows.
- Integrates with Resource Provisioners like GlideinWMS.
- Support for Shell Code Generator

Pegasus WMS is funded by the National Science Foundation OCI SDCI and SI2 program grants 0722019 and 1148515.