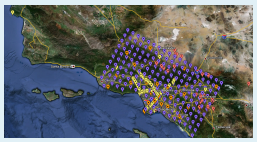
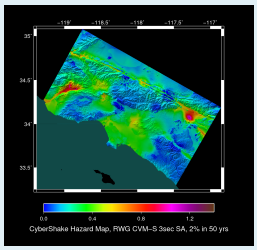
SSI: Distributed Workflow Management Research and Software in Support of Science

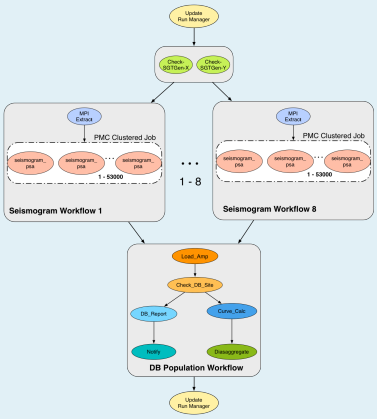
Ewa Deelman, USC Information Sciences Institute, deelman@isi.edu
Miron Livny, University of Wisconsin, Madison, miron@cs.wisc.edu

Pegasus WMS

- Pegasus is a system for mapping and executing abstract application workflows over a range of execution environments
- The output is an executable workflow that can be distributed over a variety of resources (clouds, XSEDE, OSG, campus clusters, grids, workstations)
- Pegasus can run workflows comprising of millions of tasks
- Pegasus WMS consists of three main components: the Pegasus mapper, DAGMan workflow engine, and Condor scheduler
- Tasks are mapped to the execution resources by the Pegasus mapper based on static and/or dynamic information sources Pegasus also automatically plans necessary data transfers and performs static optimizations such as task clustering.
- DAGMan manages the task execution order and provides workflow-level checkpointing and retries
- Condor manages job execution on distributed resources





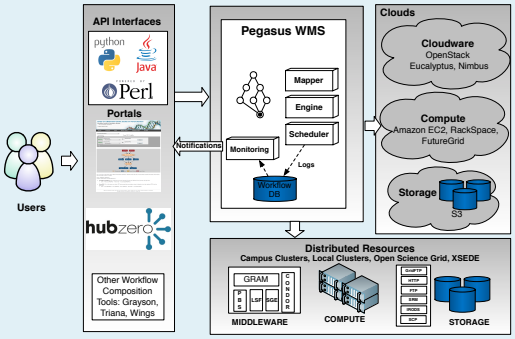


CyberShake Study 13.4

- 2013 Southern California seismic hazard study done by SCEC
- Ran across several large-scale compute resources
  - NICS Kraken, NCSA Blue Waters, TACC Stampede
- 1144 workflows, 470M tasks, 32K jobs, 63.2 TB data, 12 M hours
- Used Pegasus-MPI-Cluster to run HTC tasks on HPC machines

Composing Workflows

- Workflows are expressed in DAX (Directed Acyclic graph in XML)
- DAXes can be generated using Java, Perl or Python APIs
- Higher level workflow composition tools like Grayson, Wings, and Triana can also be used
- Integrated with HUBZero
- Scientists can use application-specific gateways such as CGSMD
  - http://portal.nimhgenetics.org



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

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:

- 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

Pegasus Features

- Clustering of small tasks into batches for performance
- Optimized data transfers and support for many protocols
- Data reuse in case intermediate data products are available
- Automatic data cleanup to reduce data footprint
- Retries computations in case of failures
- Workflow-level checkpointing through data reuse and DAGMan
- Monitoring and debugging tools to support large workflows
- Workflow progress can be tracked through a database
- Support for workflow- and task-level notifications
- Stores provenance of data used and produced, and which software was used with what parameters
- Integrates with resource provisioners like glideinWMS
- Shell code generator compiles workflows into shell scripts
- Pegasus-MPI-Cluster enables fine-grained task graphs to be executed efficiently on HPC resources

http://pegasus.isi.edu

Pegasus WMS is funded by the National Science Foundation SI2 program grant #1148515.