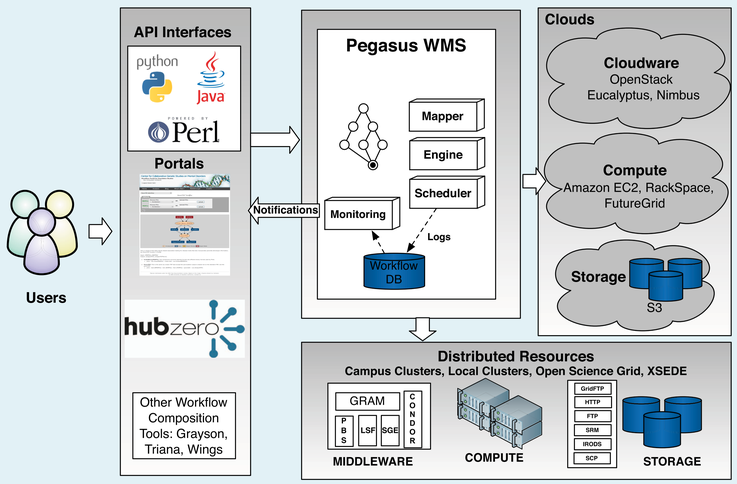
Pegasus WMS: Enabling Large Scale Workflows on National CyberInfrastructure

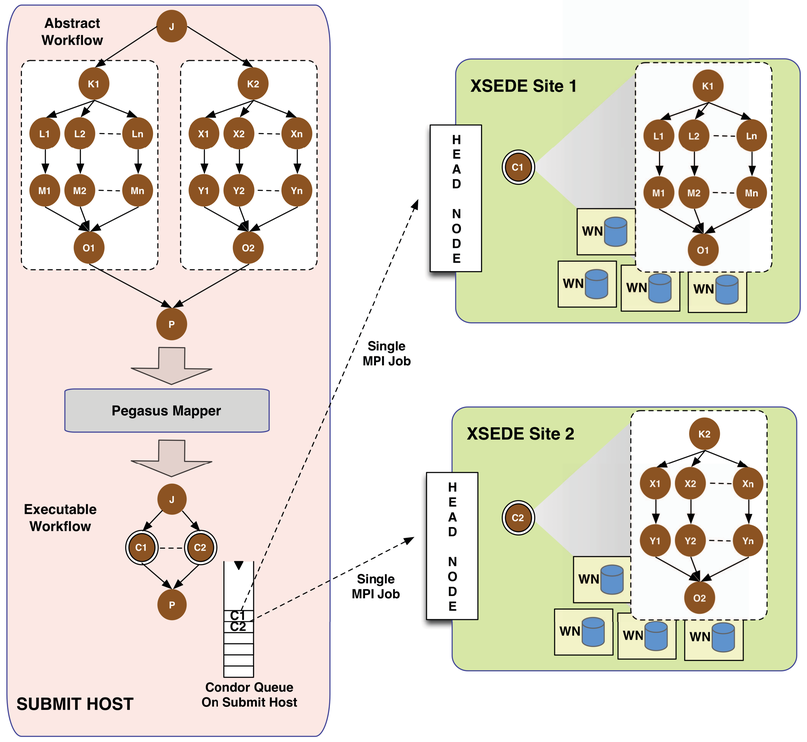
Karan Vahi[1], Ewa Deelman[1], Gideon Juve[1], Mats Rynge[1], Rajiv Mayani[1], Scott Callaghan[2] and Philip Maechling[2]
[1]University of Southern California's Information Sciences Institute, [2]University of Southern California – Southern California Earthquake Center

Overview

- 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 (OSG, XSEDE, commercial and academic clouds, campus grids, clusters, workstation)
- Pegasus can run workflows comprising of millions of tasks.
- Pegasus Workflow Management System (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.
- Pegasus automatically retries failed tasks in case of failures.



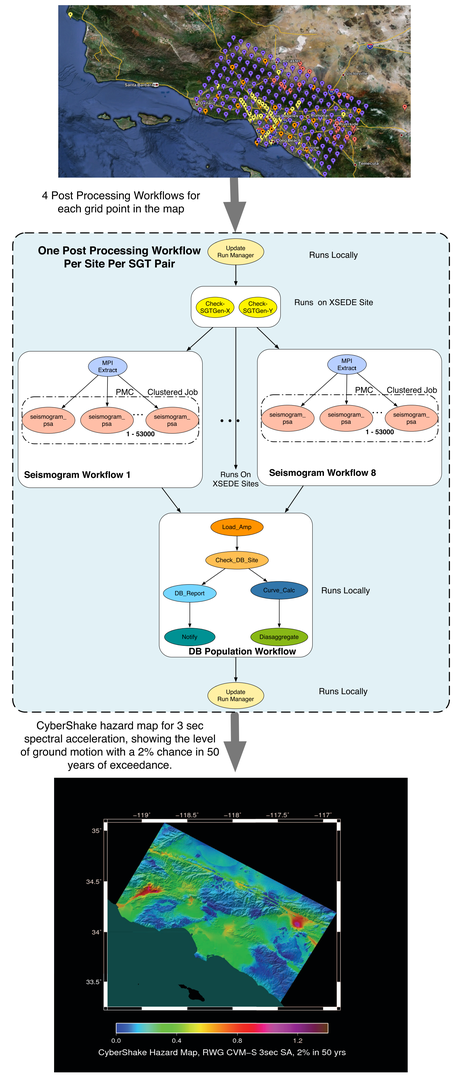
Pegasus MPI Cluster



Distributing large, fine-grained workflows across cluster resources at different sites.

- The large workflow is partitioned into independent sub graphs, which are submitted as self-contained Pegasus MPI Cluster (PMC) jobs to the remote sites.
- The PMC job is expressed as a **DAG** and uses the master-worker paradigm to farm out individual tasks to worker nodes.
- Has in built retry and recovery features. Writes a transaction log to enable recovery in the case of failure.
- Easier to setup than Condor Glideins as no special networking required. Relies on standard MPI constructs

SCEC CyberShake Workflows



Problem Description

- Builders ask seismologists: "What will the peak ground motion be at my new building in the next 50 years?"
- Seismologists answer this question using Probabilistic Seismic Hazard Analysis (PSHA)
- For each site in the input map, generate a hazard curve

Per site post processing workflow

- 410,000 tasks in the workflow
- Input Strain Green Tensor 40 GB
- Outputs about 11GB per site
- CPU Time used : approx 800 hours

Runs on XSEDE in 2013 – CyberShake Study 13.4

- Hazard Map Covering 286 sites with (4 SGT combinations per site)
- Executed 1144 post processing workflows on Stampede.
- **Input Data:** 1144 sets of SGTs x 40 GB/set = 44.7 TB
- **Stored Output Data**: 1144 sites x 11.6 GB/site = 13.0 TB
- **Workflow Logs:** 1144 sites x 4.9 GB/site

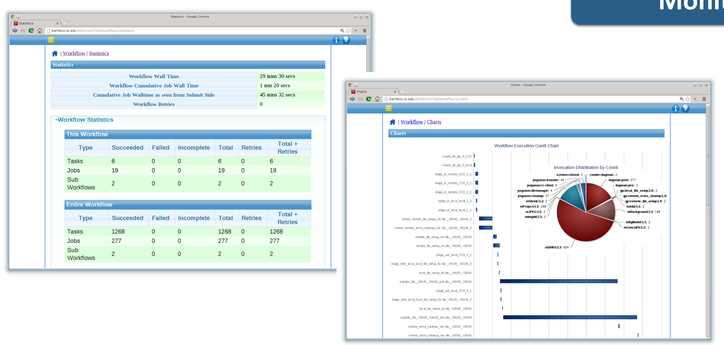
Pegasus Features

- The abstract workflow format (DAX) allows users to represent computations in a portable and infrastructure independent manner. Ideal for sharing!
- Clustering of small tasks into large clusters for performance reducing job scheduling overheads.
- Optimized data transfers and ability to use different protocols.
- Data reuse in case intermediate data products are available
  - workflow-level checkpointing
- Automatic data cleanup which reduces workflow data footprint
- Support for Workflow and Task level notifications (email, instant messenger, user defined script callout)
- Support for Shell Code Generator for local testing / debugging

Data Staging Configurations

- **Shared Filesystem** (Head Node and the worker nodes of execution sites share a filesystem)
- **Non Shared Filesystem with Staging Site** (Head Node and Worker Nodes don't share a filesystem). Data is staged from an external staging site
- **CondorIO** (Head Node and Worker Nodes don't share a filesystem). Data is staged from the submit host using Condor File Transfers

Monitoring and Debugging Capabilities



- Workflow progress can be tracked through a database.
- Database gets populated with workflow and job runtime provenance, including which software was used and with what parameters.
- Command line monitoring and debugging tools to debug large scale workflows.
- A Flask based web dashboard now allows users to monitor their running workflows and drill down to the jobs in a workflow and check their status and output.

Acknowledgments:
- Pegasus WMS is funded by the National Science Foundation OCI SDCI program grant #1148515 .
- Condor : Miron Livny, Kent Wenger, University of Wisconsin Madison

http://pegasus.isi.edu