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 (Clouds, XSEDE, OSG, Campus Grids, Clusters, Workstation).
- Pegasus can run workflows comprising of millions of tasks.
- Pegasus Workflow Management System (WMS) consists of four main components: the Pegasus mapper, Condor DAGMan, Condor schedd, and Pegasus Monitor.
- 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.

Data Staging Configurations

- **Shared Filesystem** (Head Node and the worker nodes of execution sites share a filesystem). Popular on XSEDE.
- **Non Shared Filesystem with Staging Site** (Head Node and Worker Nodes don’t share a filesystem). Data is staged by Pegasus Lite at runtime from an external staging site. Allows users to run across sites. Popular on OSG with SRM as data staging server.
- **Condor** (Head Node and Worker Nodes don’t share a filesystem). Data is staged from the submit host using Condor File Transfers. Popular on OSG and Cloud Environments.

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.

Pegasus MPI Cluster

- **Abstract Workflow**
- **Executable Workflow**
- **Submit Host**
- **DAGMan**
- **Systems**
- **Staging**
- **Head Node**
- **Worker Nodes**
- **Global Data Transfer**

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

Earthquake Science

- 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
- Each per site post processing workflow has
  - 820,000 tasks in the workflow
  - Input Strain Green Tensor 40 GB
  - Outputs about 10 GB per site
  - CPU Time used: 38 days, 23 hrs

Proposed Runs on XSEDE

- 3 Hazard maps each covering 200 sites
- To be run mainly on Kraken using MPI (PMC)
- Inputs SGT: approx 15.6 TB (40 * 400 GB)
- Outputs: 500 million files (820000/site x 600 sites) approx 5.8 TB (600 * 10 GB)
- Number of Output Files: = about 500 million

Acknowledgments:

- Pegasus WMS is funded by the National Science Foundation OCI SDI program grant #0722019 and OCI 502-515 program grant #1348515
- Condor: Miron Livny, Kent Wenger, University of Wisconsin Madison

http://pegasus.isi.edu