Pegasus: Planning for Execution in Grids

Pegasus Team

Ewa Deelman
Carl Kesselman
Gaurang Mehta
Gurmeet Singh
Mei-Hui Su
Karan Vahi
James Blythe
Yolanda Gil

USC/Information Sciences Institute

http://pegasus.isi.edu

VDS Collaborators

Ian Foster
Doug Scheftner
Jens Voeckler
Mike Wilde
Yong Zhao

University of Chicago

http://vds.isi.edu

Pegasus is supported by NSF

Pegasus is a workflow mapping system that maps an abstract workflow description onto Grid resources. The abstract workflows describe the computation in terms of logical files and logical transformations and indicate the dependencies between the workflow components. The abstract workflow is independent of the computational and data resources. Pegasus takes the abstract workflow description and information about the available resources and generates an executable workflow that describes the computations, data transfers and data registration tasks and their dependencies.

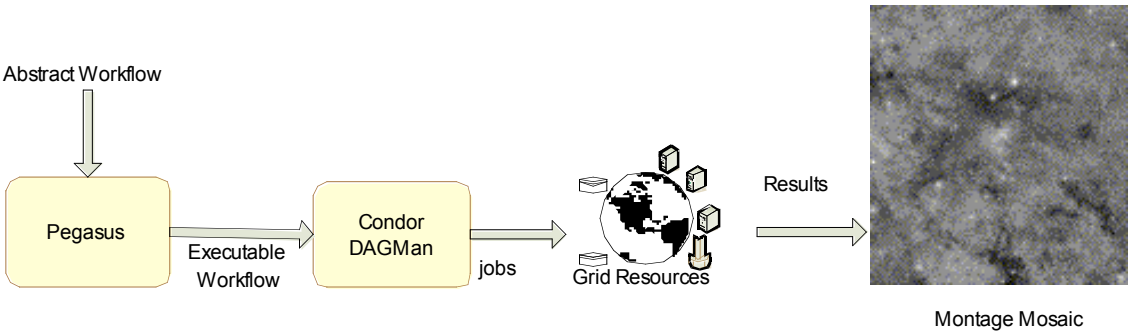
Pegasus consults various Grid information services, such as the Globus Monitoring and Discovery Service (MDS), the Globus Replica Location Service (RLS), the Metadata Catalog Service (MCS), and the Transformation Catalog to determine the available computational resources and data.

Pegasus reduces the abstract workflow based on the available data. For example, if intermediate workflow products are found to be registered in the RLS, Pegasus does not perform the transformations necessary to produce these products.

Given all the information, Pegasus generates an executable workflow, which identifies the resources where the computation will take place, identifies the data movement for staging data in and out of the computation, and registers the newly derived data products in the RLS and MCS.

The executable workflow is given to Condor's DAGman for execution on the Grid.

Pegasus is released as part of the GriPhyN Virtual Data System (**VDS**) which also includes a language used to describe the abstract workflow.



Pegasus' Workflow Mapping

Applications using Pegasus

The LIGO Scientific Collaboration

The Laser Interferometer Gravitational Wave Observatory (LIGO) and the LIGO Scientific Collaboration (LSC) are making great strides toward performing scientifically significant data analysis using Grid resources. LIGO is searching for gravitational waves emitted by binary neutron stars and black holes. LIGO collects data from interferometers during long-duration (several weeks) scientific runs and then performs large-scale analysis.

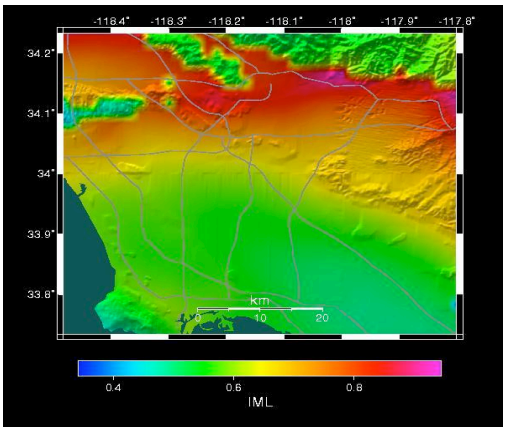
Traditionally, LIGO performed gravitational wave searches on dedicated Beowulf clusters with ~ 300 CPUs per cluster using Condor DAGMan to manage job submission and control. Today, LIGO scientists are developing workflows for a given analysis in an abstract workflow format. Pegasus is then used to generate an executable workflow. This workflow is then submitted to Condor's DAGMan.

www.ligo.caltech.edu

NSF co-op agreement PHY-0107417

Earthquake Center

Southern California Earthquake Center (SCEC), in collaboration with the USC Information Sciences Institute, San Diego Supercomputer Center, the Incorporated Research Institutions for Seismology, and the U.S. Geological Survey, is developing the Southern California Earthquake Center Community Modeling Environment (SCEC/CME).



The SCEC/CME system is an integrated geophysical simulation modeling framework that automates the process of selecting, configuring, and executing models of earthquake systems. By facilitating the investigation, modification, and adoption of these physics-based models, the SCEC/CME can improve the scientists' system-level understanding of earthquake phenomena and can substantially improve the utilization of seismic hazard analysis.

Within SCEC, Pegasus is currently being used in the **CyberShake** Project whose goal is to calculate Probabilistic Seismic Hazard curves for several sites in the Los Angeles area. The hazard curves in this study are generated using 3D ground motion simulations rather than empirically derived attenuation relationships. Currently SCEC is running hundreds of analyses some of which run over a period of several days. For this analysis Pegasus is scheduling the workflows onto the TeraGrid resources.

http://www.scec.org

NSF (ITR) -

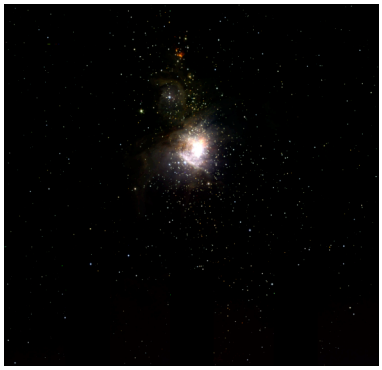
EAR – 0122464

NSF Co-op Agreement EAR-0106924

USGS Co-op Agreement 02HQAG0008

Montage (NVO and NASA)

Montage is a Grid-capable astronomical mosaicking application. Montage is used to re-project, background match and, finally, mosaic many image plates into a single image. Montage has been used to mosaic image plates from synoptic sky surveys, such as 2MASS in the infrared wavelengths. Montage is being developed under the ESTO CT project by a team that includes Caltech IPAC, Caltech CACR, and JPL. Currently Pegasus is being deployed at Caltech IPAC in a service that will provide a wide range of mosaics to the astronomy community.



The picture on the right shows the Sword of Orion (M42, Trapezium, Great Nebula). This mosaic was obtained by running a Montage workflow through Pegasus and executing the workflow on the Teragrid resources.

NASA's Earth Science Tech Office, Computational Technologies Project, under Co-op Agreement NCC5-626 between NASA and CIT.

montage.ipac.caltech.edu

Thanks to the TeraGrid and OSG for the use of the grid resources.

www.teragrid.org

www.opensciengrid.org

Southern California

Other Applications

Astronomy: Galaxy Morphology (National Virtual Observatory)

- Investigates the dynamic state of galaxy clusters.

Bioinformatics: BLAST (PACI Data Quest)

- Set of sequence comparison algorithms that are used to search sequence databases for optimal local alignments to a query

Medicine: Tomography (NIH-funded project)

- Derivation of 3D structure from a series of 2D electron microscopic projection images
- Reconstruction and detailed structural analysis of complex structures like synapses and large structures like dendritic spines