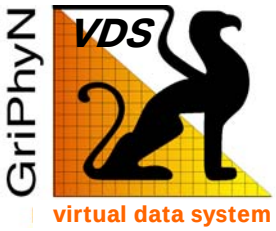


Delivering Workflow Solutions using the Virtual Data System

VDS Team:

Argonne NATIONAL LABORATORY

THE UNIVERSITY OF CHICAGO

Ben Clifford
Ian Foster
Mike Wilde
Yong Zhao

ISI Information Sciences Institute

Ewa Deelman
Carl Kesselman
Gaurang Mehta
Mei-Hui Su
Karan Vahi
Jens Voeckler

VDS is funded by the National Science Foundation.

http://vds.uchicago.edu

http://pegasus.isi.edu

VDS, The Virtual Data System, provides a set of tools for expressing, executing, and tracking the results of workflows.

Workflows consist of graphs of application invocations, and can be expressed in a location-independent, high-level "Virtual Data Language" - VDL - which frees the user from specifying details of the location of files and programs in a distributed environment.

VDL workflows can be executed in a variety of environments ranging from the desktop to Grids such as the Open Science Grid and TeraGrid. VDL definitions are stored in a Virtual Data Catalog - VDC - that provides for the tracking of the provenance of all files derived by the workflow. Both VDL and the workflows that it generates can be expressed in XML documents. The workflow is expressed in abstract terms and is used by the VDS planner, Pegasus, to generate executable forms of the workflow.

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.

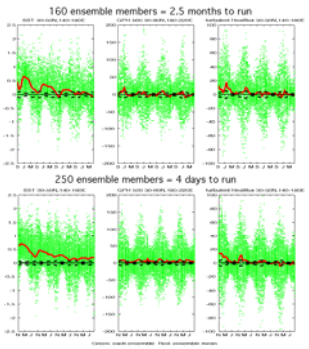
Given all the information, Pegasus generates an executable workflow, which identifies the resources where the computation will take place, identifies the data movement for staging data in and out of the computation, and registers the newly derived data products in the RLS and MCS.

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

GriPhyN VDS virtual data system

VDS Applications

Climate Modeling



Ensemble simulations are a promising technique for identifying the signal of atmospheric response to extra-tropical sea surface temperature variability with high statistical significance. The large atmospheric weather noise at daily time scales means that the ensemble size needs to be at the order of 1000 to obtain statistically significant results.

Ensemble simulations are performed using the Fast Ocean Atmosphere Model (FOAM) and the Virtual Data System and run on resources provided by the TeraGrid. FOAM jobs utilize 8 processors and run on the order of 3 hours.

A significant run of FOAM performed on SDSC's TeraGrid resources processed an ensemble of 250 members in 2 days, representing roughly 7,000 CPU-hours with an average parallelism of 140 concurrent jobs. Originally, this type of computation took 10 weeks of manual effort to complete on a shared supercomputer.

Genome Analysis and Database Update system

The Computational Biology Group at Argonne National Laboratory uses VDS within GADU - the Genome Analysis and Database Update system, to perform high-throughout analysis and annotation of the genomics information that it regularly fuses from multiple public information sources, providing an integrated facility that supports research programs within DOE as well as public visitors to its web portal. VDS runs huge workflows across the Open Science Grid and TeraGrid, applying tools such as BLAST, BLOCKS and PFAM to enrich the warehouse.

Science Education

In Science Education, VDS is being used by the QuarkNet and I2U2 projects to run workflows for student analysis of data from cosmic rays, high energy physics events, and environmental sensors from the LIGO gravitational wave observatory. Students at over a hundred high schools around the U.S. are using the system in an inquiry-based approach to scientific discovery.

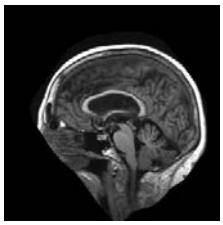
Social Informatics

pegasus

The Social Informatics Data Grid project is testing VDS for use in the analysis of multi-modal time series data from physiological and neuroscience experiments, and providing a portal through which scientists can select datasets, perform large-scale analyses that leverage Grid resources, and track the provenance of their results.

GriPhyN VDS virtual data system

Medical Applications



The Brain Research Imaging Center at the University of Chicago is testing VDS for use in managing the datasets and workflows of its neuroscientists in an integrated fashion that promotes scientific validation and collaboration.

With the Department of Radiology and the Graduate Programs in Medical Physics at the University of Chicago, VDS is being used to leverage Grid computing to improve the prediction accuracy of artificial intelligence and computer vision algorithms in identifying cancerous lesions in radiographic images.

Astronomy and High-energy Physics

Montage is a Grid-capable astronomical mosaicking application. Montage is used to re-project input images, rectify their backgrounds to a common level, and, finally, mosaic them into a single image. Montage has been used to mosaic image plates from synoptic sky surveys, such as the Two-Micron All Sky Survey (2MASS). Montage is being developed under the ESTO CT project by a team that includes Caltech Infrared Processing and Analysis Center (IPAC), Caltech CACR and JPL. Currently VDS is being deployed at Caltech IPAC in a service that will deliver image mosaics to the astronomy community.

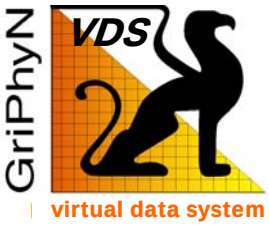
The picture on the right shows the Sword of Orion (M42, Trapezium, Great Nebula). This mosaic was obtained by planning a Montage workflow using Pegasus and executing the workflow on TeraGrid resources.

VDS has also been applied to problems of galaxy cluster finding and large-scale co-addition in the Sloan Digital Sky Survey, and to generate simulated high energy physics events within the ATLAS experiment.

Gravitational Wave Physics

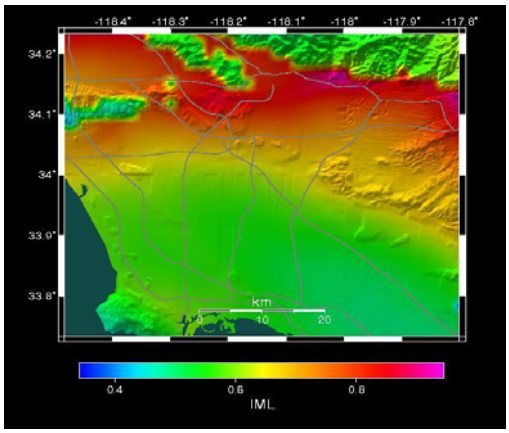
The Laser Interferometer Gravitational Wave Observatory (LIGO) and the LIGO Scientific Collaboration (LSC) are making great strides toward performing scientifically significant data analysis using Grid resources. LIGO is searching for gravitational waves emitted by binary neutron stars and black holes.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

pegasus



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, which is then submitted to Condor DAGMan to make use of the vast resources of the Open Science Grid.

Earthquake Science



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, VDS 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.

Thanks to TeraGrid and OSG for the use of Grid resources:

www.teragrid.org

www.opensciencegrid.org