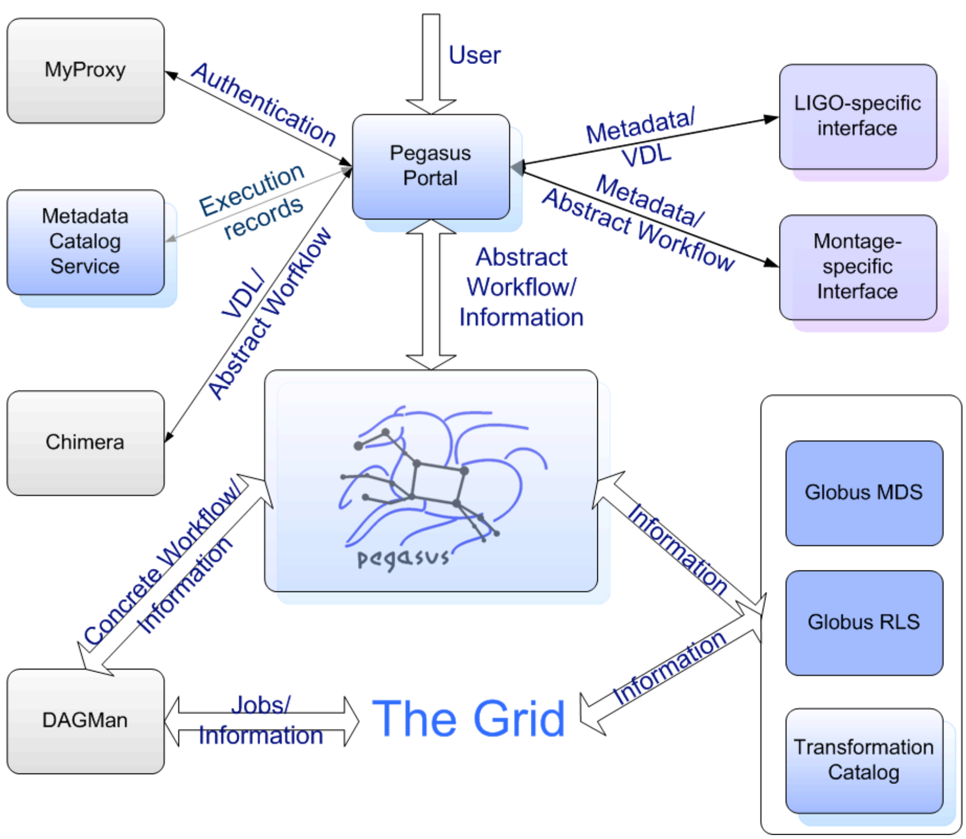
LIGO-T030249-00-Z

Pegasus: Planning for Execution in Grids

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 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, 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 figure below is a simplified view of the system being demonstrated at SC 2003.



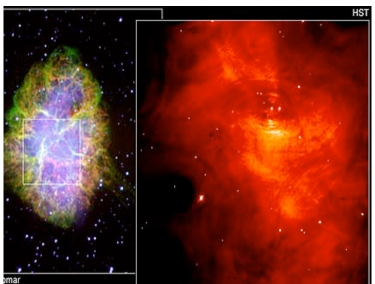
People Involved

Ewa Deelman
Carl Kesselman
Saurabh Khurana
Gaurang Mehta
Sonal Patil
Gurmeet Singh
Mei-Hui Su
Karan Vahi

USC
Information Sciences Institute

http://pegasus.isi.edu

The LIGO Scientific Collaboration



People Involved

Bruce Allen[3]
Stuart Anderson[2]
Kent Blackburn[2]
Steffen Grunewald[1]
Yousuke Itoh[1]
Albert Lazzarini [2]
Scott Koranda[3]
Dan Kozak[2]
Brian Moe[3]
Maria Alessandra Papa[1]
Hari Pulapaka[2]
Xavier Siemens[3]
Peter Shawhan[2]

[1] Albert Einstein Institute, Germany
[2] Caltech, USA
[3] University of Wisconsin Milwaukee, USA

www.ligo.caltech.edu

www.lsc-roup.phys.uwm.edu/lscdatagrid/

ttp://pandora.aei.mpg.de/merlin/

The Laser Interferometer Gravitational Wave Observatory (LIGO) and the LIGO Scientific Collaboration (LSC) are making great strides towards performing scientifically significant data analysis using Grid resources. LIGO is using the occasion of SC2003 to initiate a month-long production run with recent S2 science data to search for potential sources of continuous periodic waves near the Galactic Center.

Continuous gravitational waves, nearly sinusoidal signals in the rest frame of the source, are expected from rotating neutron stars. Due to selection effects of our observations we know of only a small fraction of the existing population of these sources. The chances are that the source nearest to us, is actually unknown. Hence, the interest in performing blind searches, scanning the regions of the sky where we have no a priori information of the presence of a source.

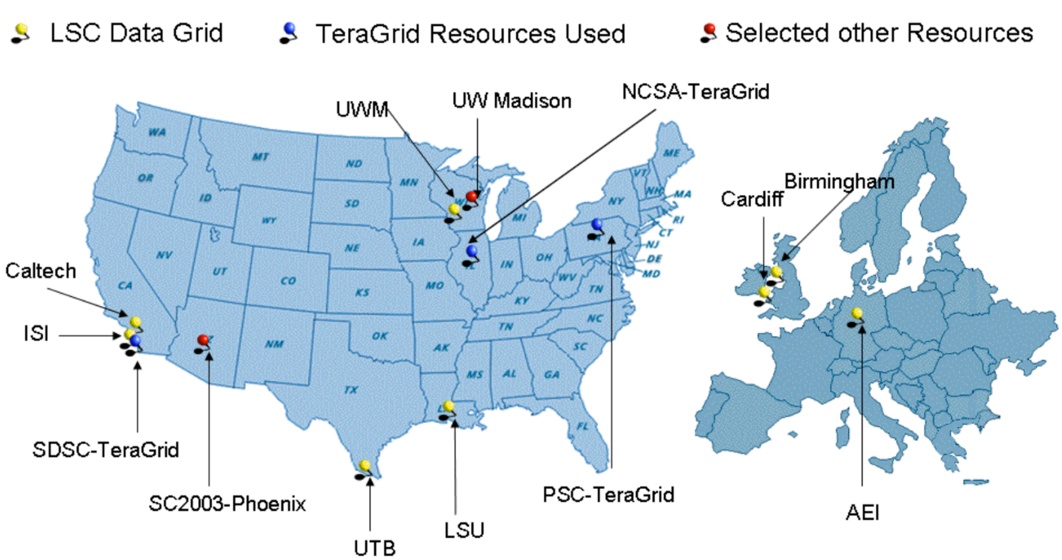
The sensitivity of searches for continuous gravitational waves increases with the length of the observation time. The resolution also increases with the observation time_very rapidly. For an observation of 10 hours a resolution of order mstrad (this is the area of a typical grid patch) is enough to ensure good coverage (less than 10% signal-to-noise-ratio loss) of most (95%) of the sources. But if the observation time increases to several weeks the resolution increases to grid sizes smaller than ~nstrdians. This makes the computational requirements for blind searches for continuous signals over extended observation times prohibitive.

The search codes to perform wide-area, wide-frequency blind searches were designed so that the computation load can be distributed among different and very loosely coupled computers -- the results of the calculations performed on one machine are completely independent of what is happening on any other machine.

These codes have performed short observation time searches using dedicated Beowulf clusters with ~ 300 CPUs per cluster. What we are presenting here is the first attempt to perform a deep area search, therefore employing a factor of ten more computing power than we have used up to now.

The second science run of the LIGO interferometric gravitational wave detectors took place over a time of 8 weeks during March-April 2003. The data from the three LIGO detectors will be analyzed to search for continuous gravitational wave signals coming from the region in the direction of the Galactic Center and the galactic core.

The resources used in this production run are drawn from the LSC testbed (see picture below) spanning condor pools at the Albert Einstein Institute, Birmingham, Caltech, Cardiff University, Louisiana State University, USC/ISI, UT Brownsville, UW Madison, UW Milwaukee as well as the resources of the iVDGL Grid3 testbed and the Teragrid.



Montage

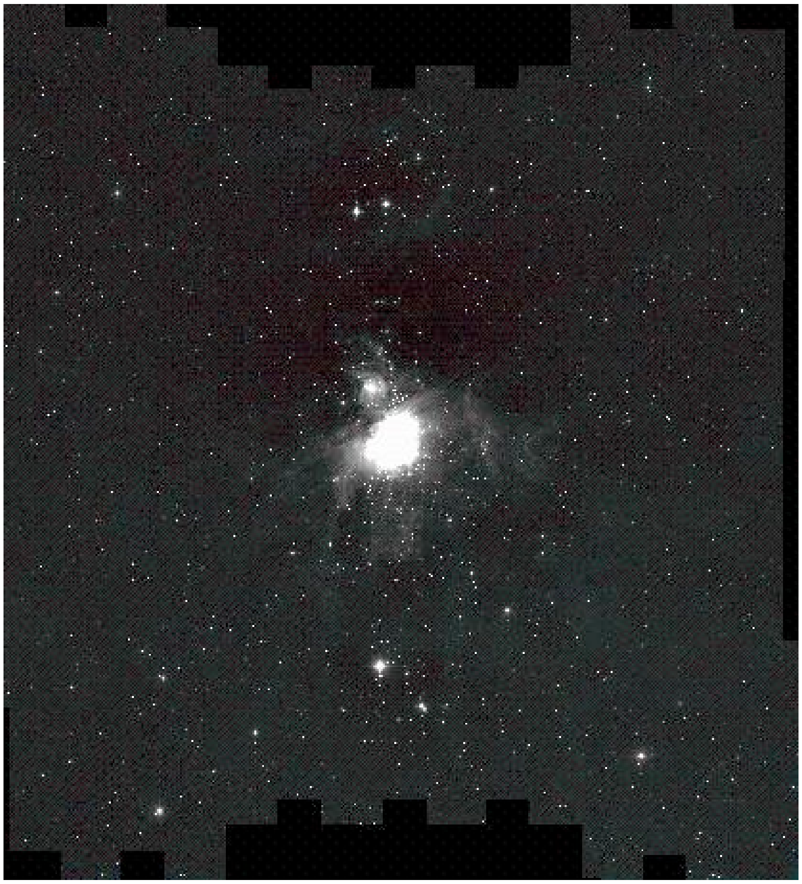
Montage is a grid-capable astronomical mosaicking application. Montage is used to reproject, background match, and finally mosaic many image plates into a single image. Montage has been used to mosaic image plates from synoptic sky surveys, such as 2MASS in the infrared wavelengths. Montage is being developed under the ESTO CT project by a team that includes Caltech IPAC, Caltech CACR, and JPL. This grid-enabled prototype of Montage is being built in collaboration with USC ISI.

The Grid testbed for Montage applications consists of condor pools at USC/ISI, UW Madison, and Teragrid resources at NCSA, PSC, and SDSC.

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

In the constellation of Orion, the hunter's belt is seen as three distinctive stars in a row. Hanging just "below" this belt is a fuzzy region, easily seen with the naked eye on a dark night, which is known as the "Sword," or "Great Nebula" of Orion.

At a distance of 1500 light years, this is the nearest large region of star formation and is one of the most intensely studied regions of the sky. A giant stellar nursery, the Orion nebula is home to thousands of young stars, some of which are known to have disks of dust that may be forming new planets. At its core, a cluster of four massive stars (known as the Trapezium) provide most of the energy that causes this nebula to glow.

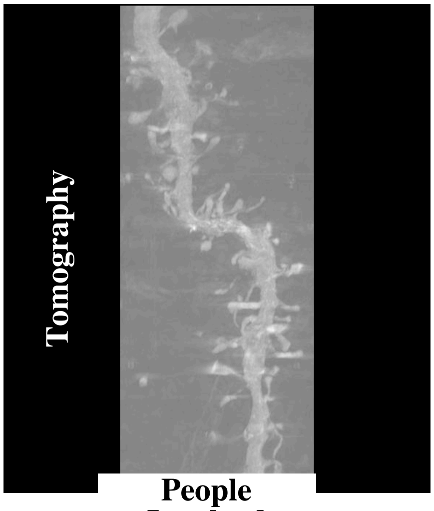


People Involved

Bruce Berriman[1]
John Good[1]
Joseph C. Jacob [2]
Daniel S. Katz[2]
Anastasia Laity[1]

[1] Caltech/IPAC
[2] JPL

http://montage.ipac.caltech.edu/



People Involved

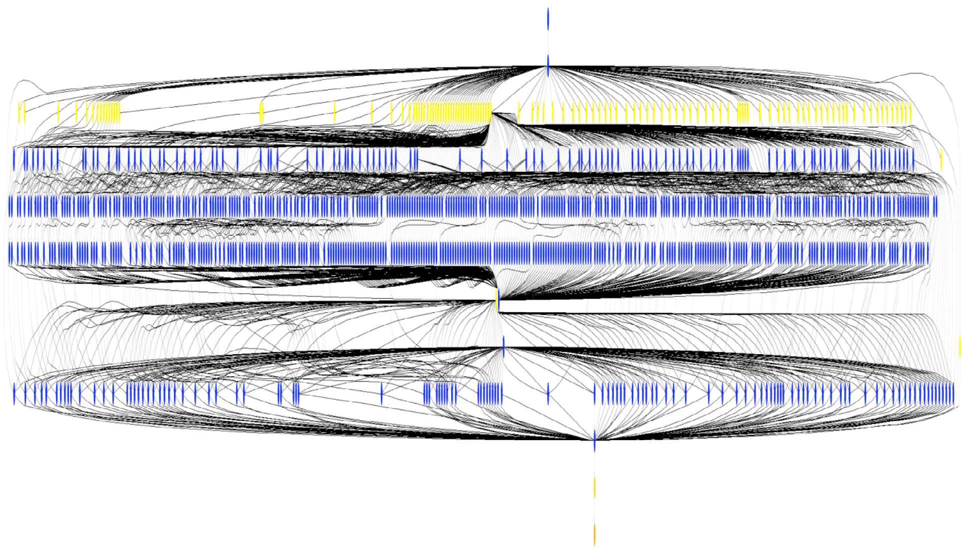
Too many GriPhyN participants to list

James Annis, Gretchen Green, Robert Hanisch, Ray Plante, and many others (Fermi, JHU, NCSA)

Natalia Matlsev, Michael Milligan, Veronika Nefedova, Alexis Rodriguez, Dinanath Sulakhe, Jens Voeckler, Mike Wilde (ANL)

Mark Ellisman, Steve Peltier, Abel Lin, Thomas Molina (SDSC)

Example Workflow_M42 Montage



Other Pegasus Applications

All GriPhyN applications: Atlas, CMS, LIGO, and SDSS.

Astronomy: Galaxy Morphology (National Virtual Observatory)

- Investigates the dynamical state of galaxy clusters
- Explores galaxy evolution inside the context of large-scale structure.
- Uses galaxy morphologies as a probe of the star formation and stellar distribution history of the galaxies inside the clusters.
- Data intensive computations involving hundreds of galaxies in a cluster

Bioinformatics: BLAST (PACI Data Quest)

- Set of sequence comparison algorithms that are used to search sequence databases for optimal local alignments to a query
- Two major runs performed:
    - 60 genomes (4,000 sequences each) processed in 24 hours,
        - 67 CPU-days of processing time delivered ~ 10,000 Grid jobs, 50 GB of data generated
    - 450 genomes (4,000 sequences each) processed in 84 hours
        - 700 CPU-days of processing time delivered, ~10,000 Grid Jobs, ~70GB of data generated

Speedup of 5-20 times were achieved because the compute nodes we used efficiently by keeping the submission of the jobs to the compute cluster constant.

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