Pegasus
A Framework for Workflow Planning on the Grid

Ewa Deelman

USC Information Sciences Institute

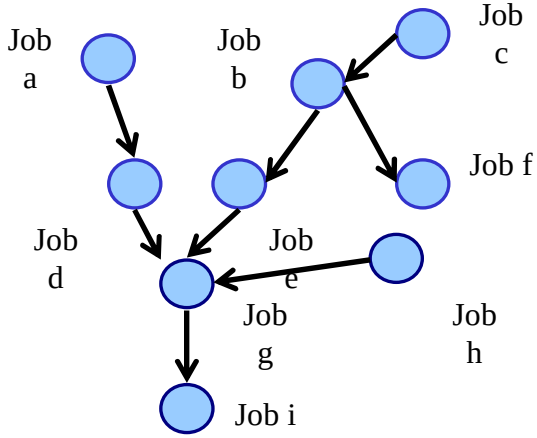
Pegasus Acknowledgments:
Carl Kesselman, Gaurang Mehta, Mei-Hui Su, Gurmeet Singh, Karan Vahi

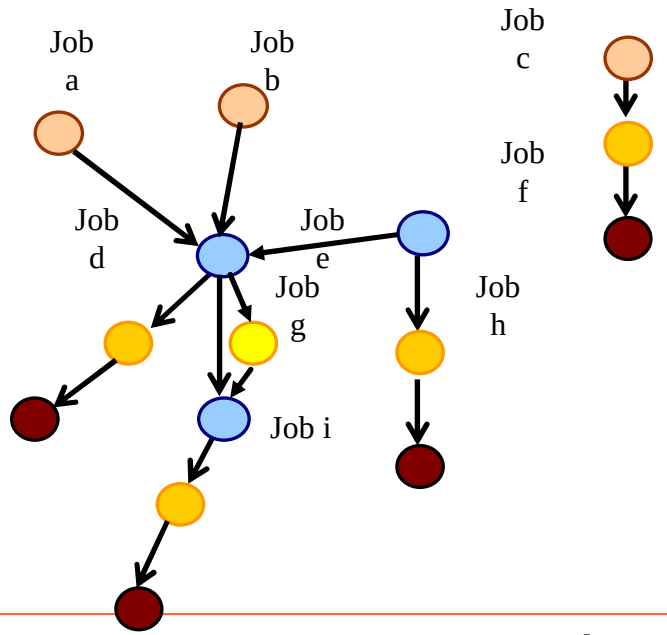
Pegasus

- Flexible framework, maps abstract workflows onto the Grid
- Possess well-defined APIs and clients for:
  - Information gathering
    - > Resource information
    - > Replica query mechanism
    - > Transformation catalog query mechanism
  - Resource selection
    - > Compute site selection
    - > Replica selection
  - Data transfer mechanism
- Can support a variety of workflow executors

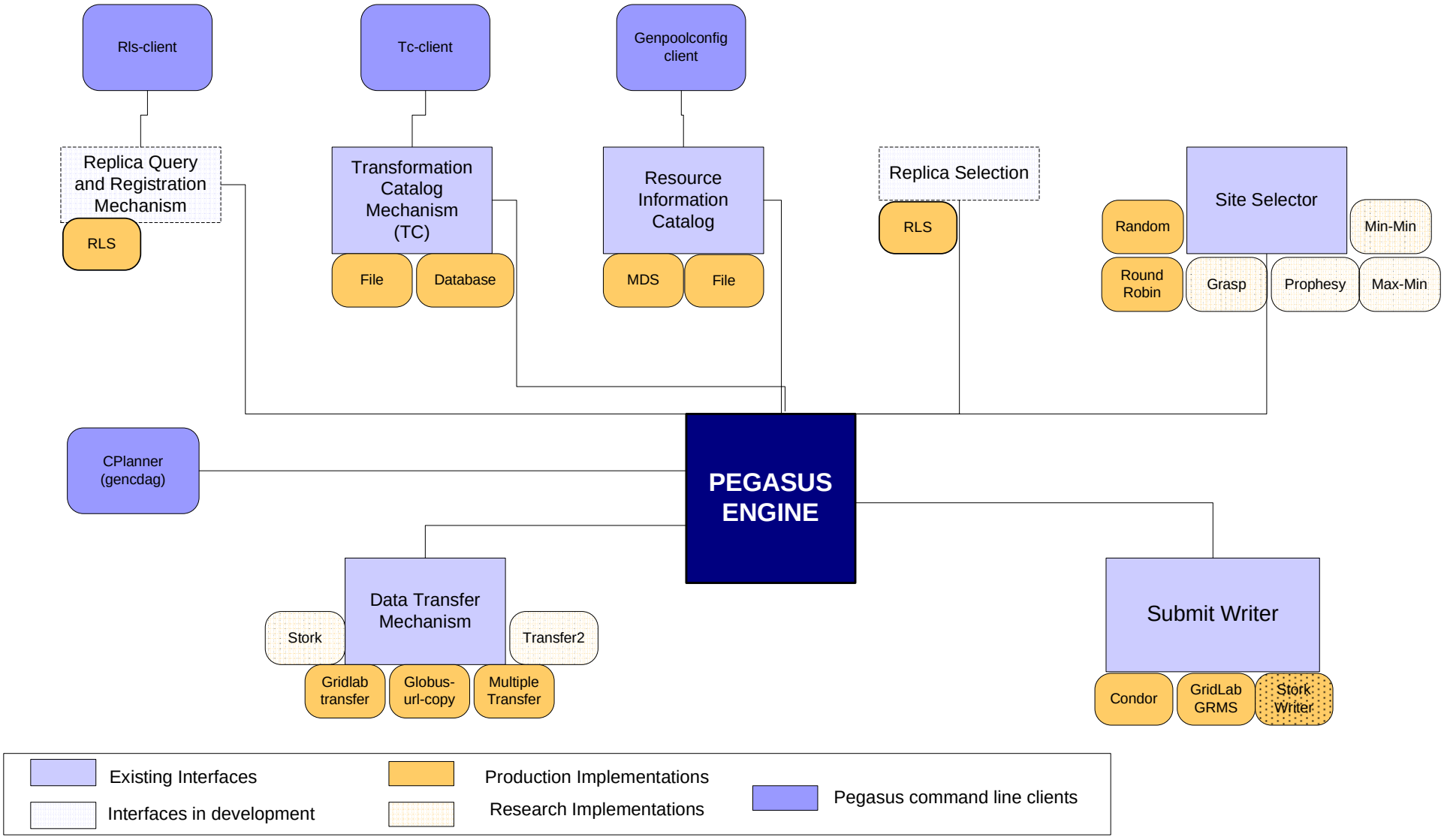
Pegasus

- May reduce the workflow based on available data products
- Augments the workflow with data stage-in and data stage-out
- Augments the workflow with data registration



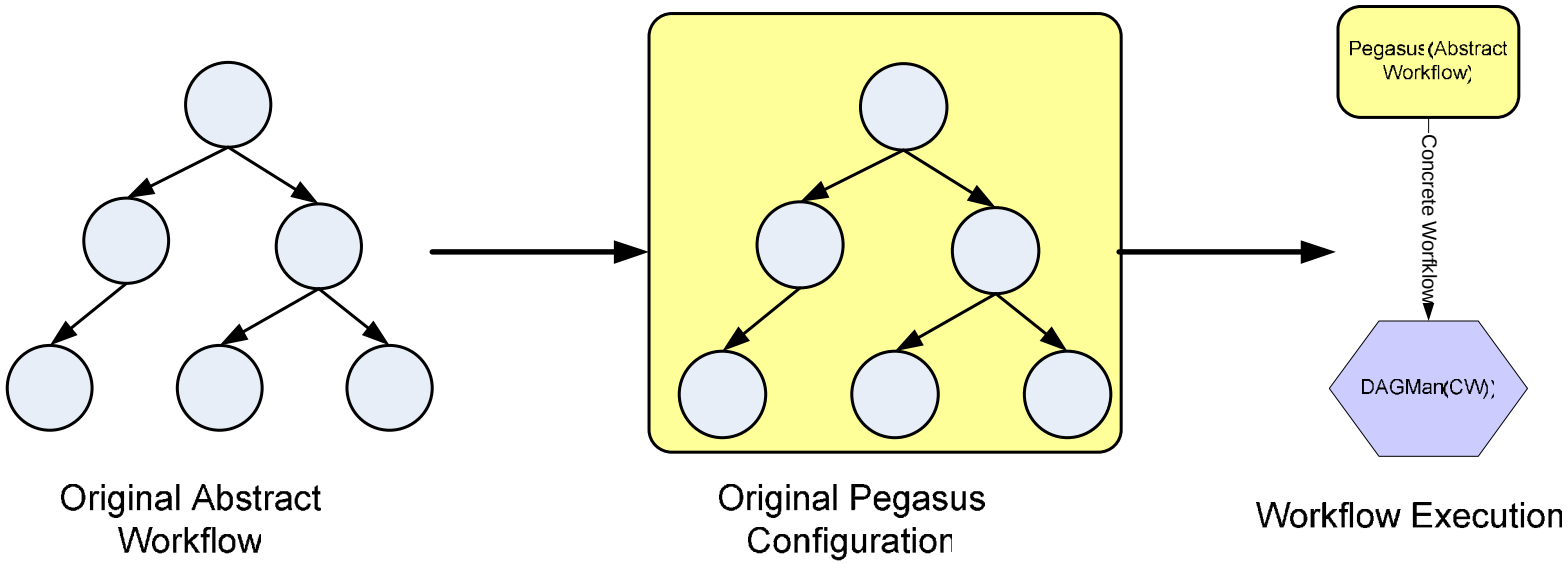


KEY
The original node
Pull transfer node
Registration node
Push transfer node
Inter-pool transfer node



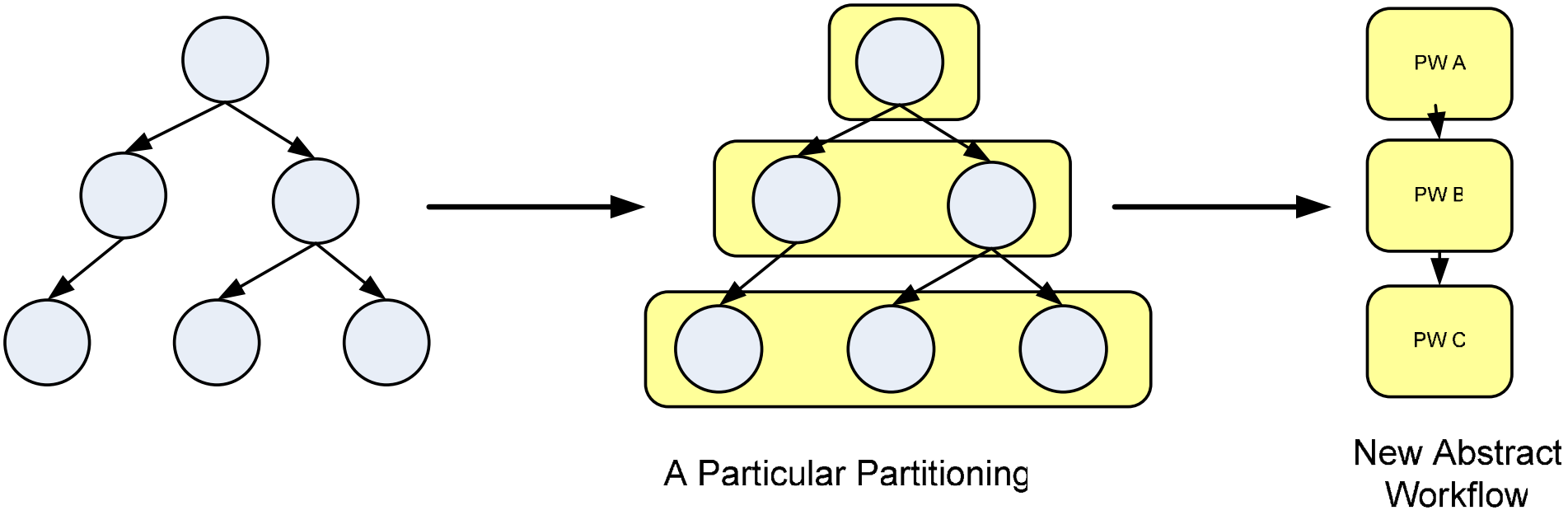
Pegasus Components

Original Pegasus configuration



Simple scheduling: random or round robin using well-defined scheduling interfaces.

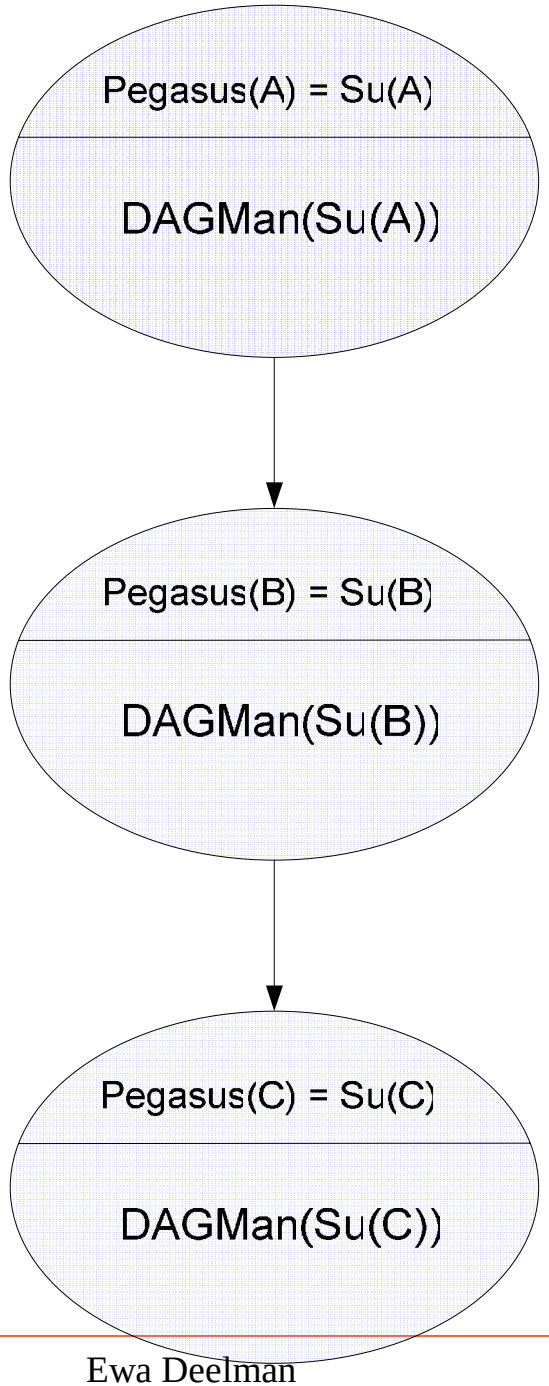
Deferred Planning through Partitioning



A variety of planning algorithms can be implemented

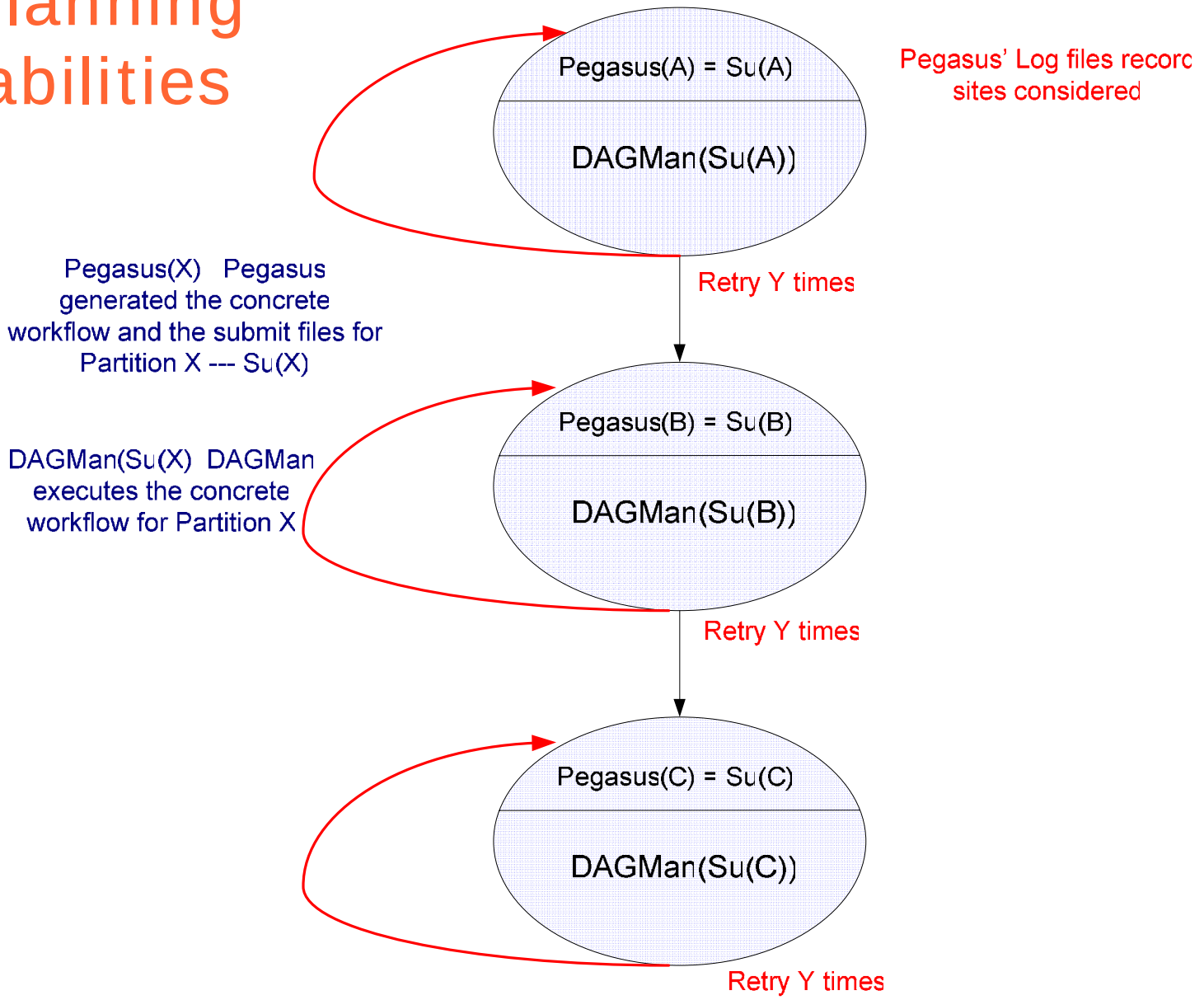
Pegasus(X): Pegasus generated the concrete workflow and the submit files for Partition X -- Su(X)

DAGMan(Su(X): DAGMan executes the concrete workflow for X

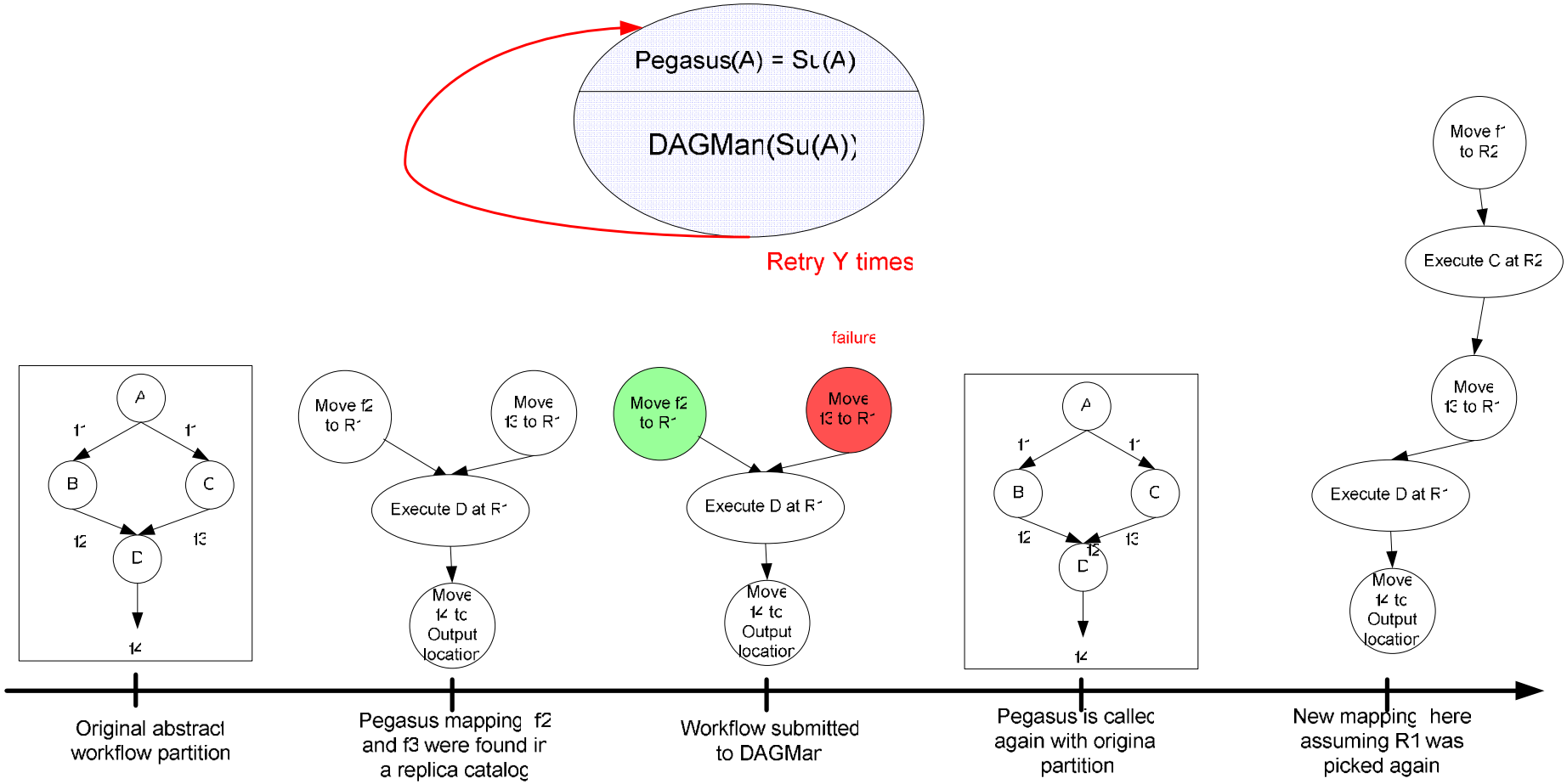


Mega DAG is created by Pegasus and then submitted to DAGMan

Re-planning capabilities



Complex Replanning for Free (almost)



Optimizations

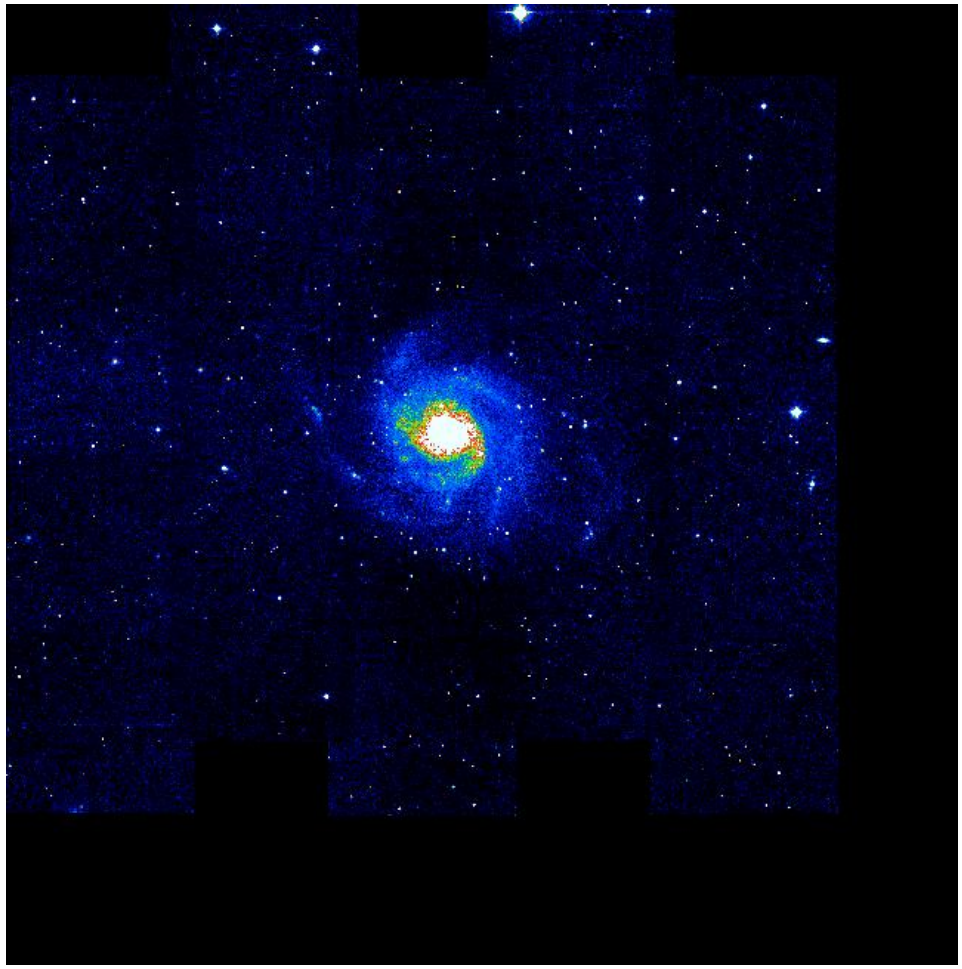
- If the workflow being refined by Pegasus consists of only 1 node
  - Create a condor submit node rather than a dagman node
  - This optimization can leverage Euryale's super-node writing component

Planning & Scheduling Granularity

- Partitioning
  - Allows to set the granularity of planning ahead
- Node aggregation
  - Allows to combine nodes in the workflow and schedule them as one unit (minimizes the scheduling overheads)
  - May reduce the overheads of making scheduling and planning decisions
- Related but separate concepts
  - Small jobs
    - > High-level of node aggregation
    - > Large partitions
  - Very dynamic system
    - > Small partitions

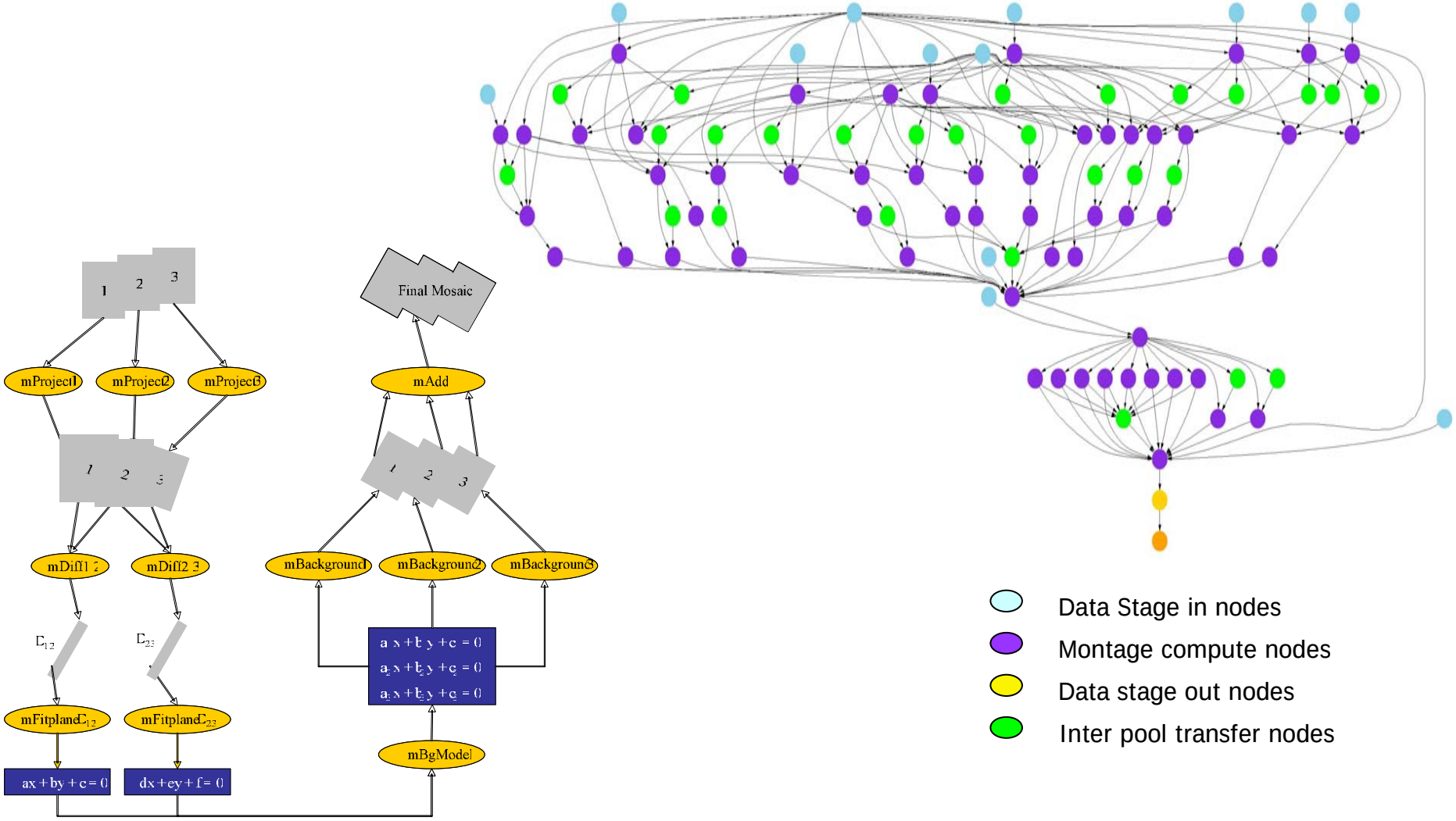
- Montage (NASA and NVO)
  - Deliver science-grade custom mosaics on demand
  - Produce mosaics from a wide range of data sources (possibly in different spectra)
  - User-specified parameters of projection, coordinates, size, rotation and spatial sampling.
- Bruce Berriman, John Good, Anastasia Laity, Caltech/IPAC
- Joseph C. Jacob, Daniel S. Katz, JPL
- Doing large: 6 and 10 degree dags (for the m16 cluster).
- The 6 degree runs had about 13,000 compute jobs and the 10 degree run had about 40,000 compute jobs

Montage



Mosaic created by Pegasus based Montage from a run of the M101 galaxy images on the Teragrid.

Montage Workflow

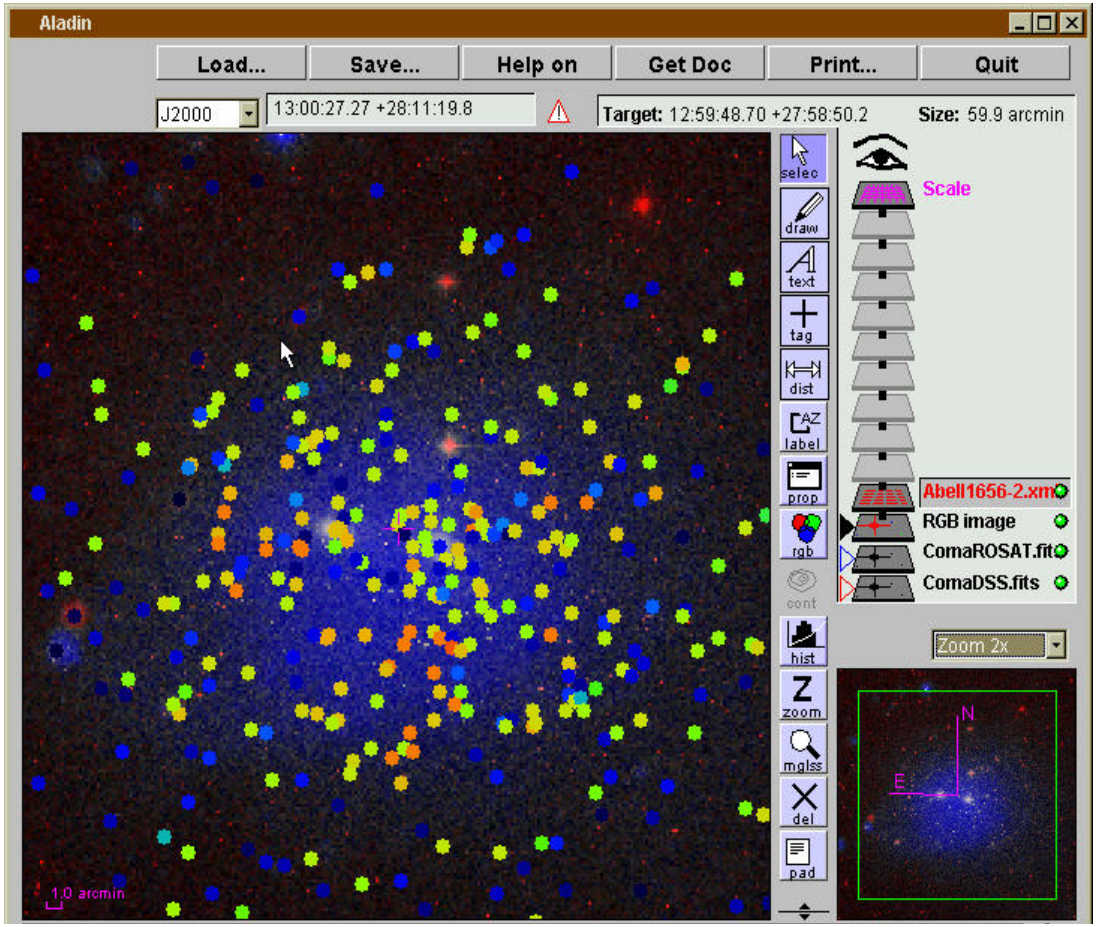


Future work

- Staging in executables on demand
- Expanding the scheduling plug-ins
- Investigating various partitioning approaches
- Investigating reliability across partitions

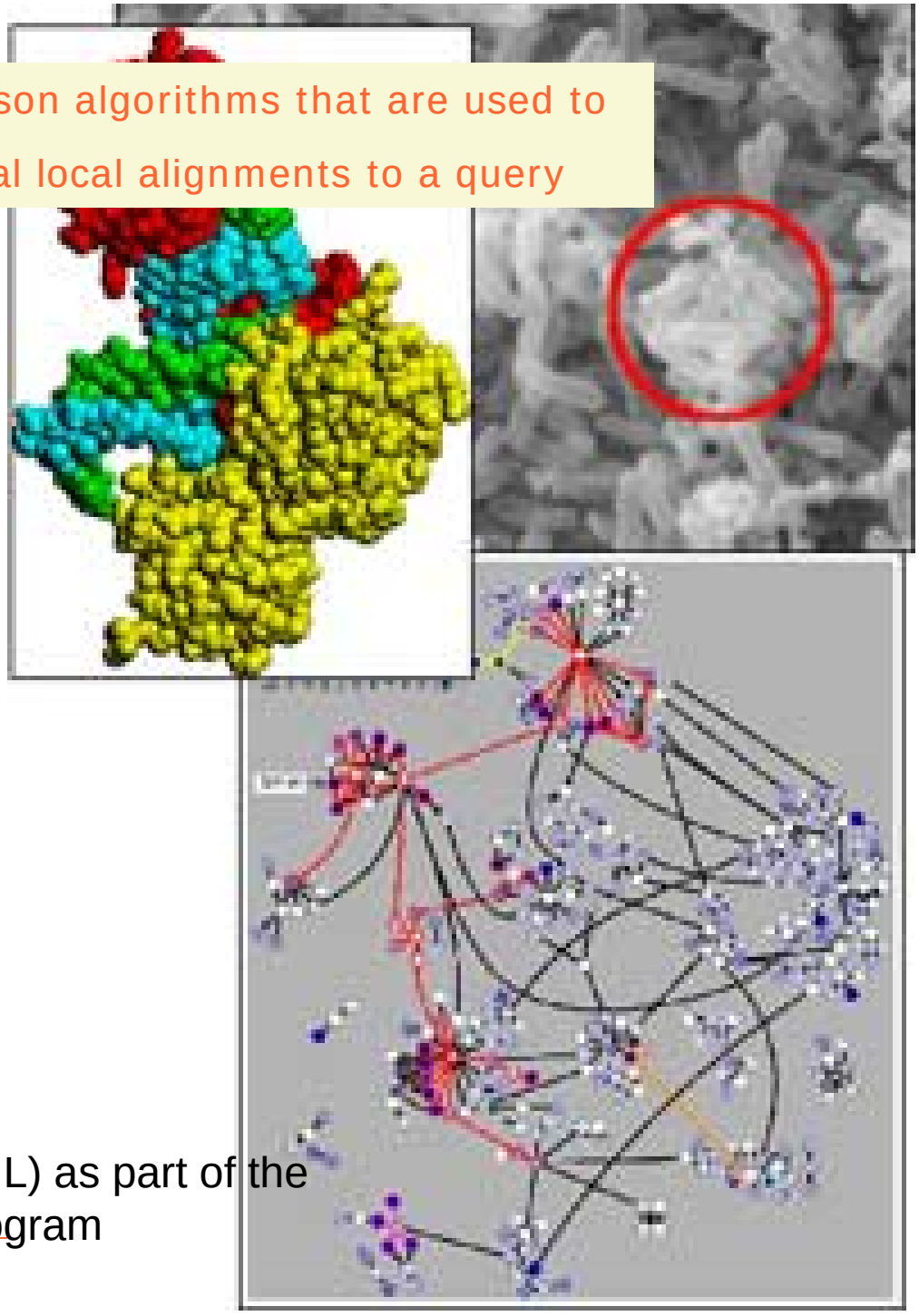
Non-GriPhyN applications using Pegasus

- 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



The x-ray emission is shown in blue, and the optical mission is in red. The colored dots are located at the positions of the galaxies within the cluster; the dot color represents the value of the asymmetry index. Blue dots represent the most asymmetric galaxies and are scattered throughout the image, while orange are the most symmetric, indicative of elliptical galaxies, are concentrated more toward the center.

BLAST: set of sequence comparison algorithms that are used to search sequence databases for optimal local alignments to a query



2 major runs were performed using Chimera and Pegasus:

1) 60 genomes (4,000 sequences each),

In 24 hours processed Genomes selected from DOE-sponsored sequencing projects

- 67 CPU-days of processing time delivered
- ~ 10,000 Grid jobs
- >200,000 BLAST executions
- 50 GB of data generated

2) 450 genomes processed

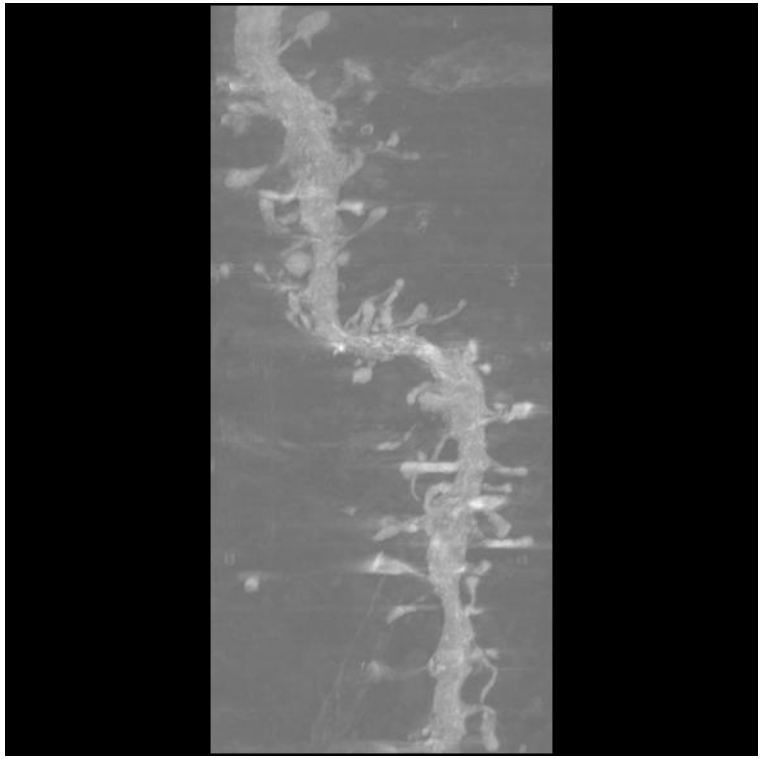
Speedup of 5-20 times were achieved because the compute nodes we used efficiently by keeping the submission of the jobs to the compute cluster constant.

Lead by Veronika Nefedova (ANL) as part of the PACI Data Quest Expedition program

Biology Applications (cont'd)

Tomography (NIH-funded project)

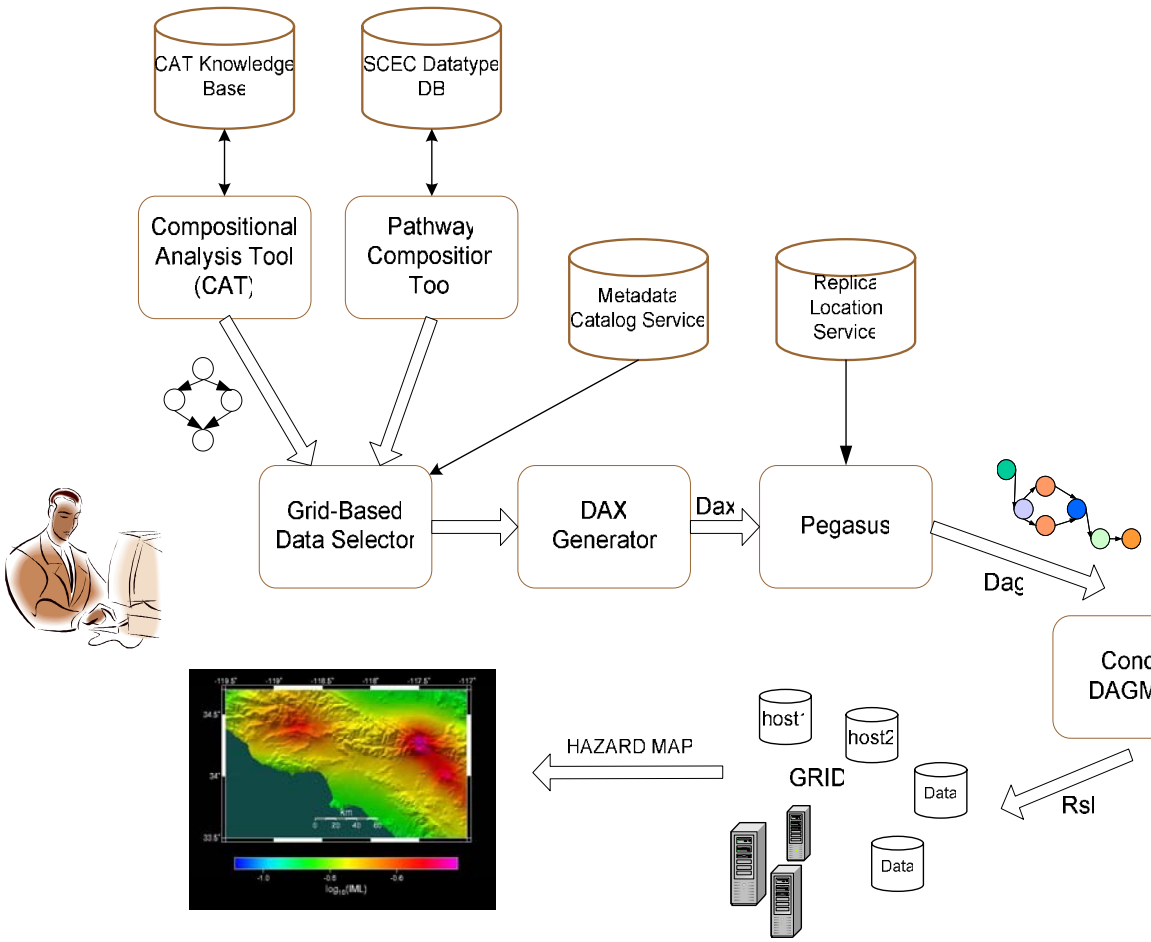
- Derivation of 3D structure from a series of 2D electron microscopic projection images,
- Reconstruction and detailed structural analysis
  - complex structures like synapses
  - large structures like dendritic spines.
- Acquisition and generation of huge amounts of data
- Large amount of state-of-the-art image processing required to segment structures from extraneous background.



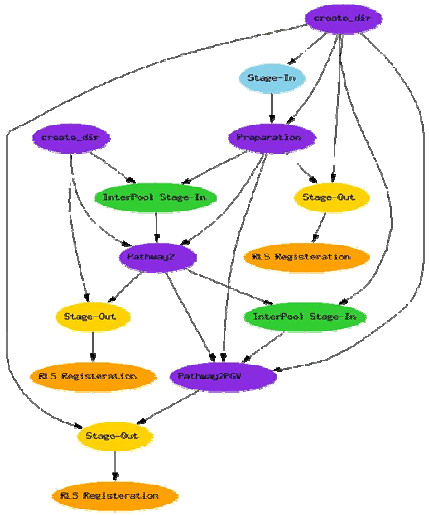
Dendrite structure to be rendered by Tomography

Work performed by Mei Hui-Su with Mark Ellisman, Steve Peltier, Abel Lin, Thomas Molina (SDSC)

Southern California Earthquake Center



The SCEC/IT project, funded by (NSF), is developing a new framework for physics-based simulations for seismic hazard analysis building on several information technology areas, including knowledge representation and reasoning, knowledge acquisition, grid computing, and digital libraries.



People involved: Vipin Gupta, Phil Maechling (USC)