Pegasus Workflows on XSEDE

Mats Rynge

<rynge@isi.edu>

USC Information Sciences Institute

XSEDE ECSS Workflow Community Applications Team

https://www.xsede.org/web/guest/ecss-workflows

The XSEDE Workflow Community Applications Team's charter is to assist researchers to use scientific workflow technologies on XSEDE to solve challenging scientific problems involving parameter sweeps, multiple applications combined in dependency chains, tightly coupled applications, and similar execution patterns that require multiple applications and multiple XSEDE resources. The workflow team accomplishes its mission through the use of third party workflow software in collaboration with the workflow developers, service providers and XSEDE Extended Collaborative Support Services.

Community mailing list: workflows@xsede.org

– To subscribe, email majordomo@xsede.org with "subscribe workflows" in the body of the message

More Information: Marlon Pierce (mpierce@iu.edu)
Mats Rynge (rynge@isi.edu)
Suresh Marru (smarru@iu.edu)

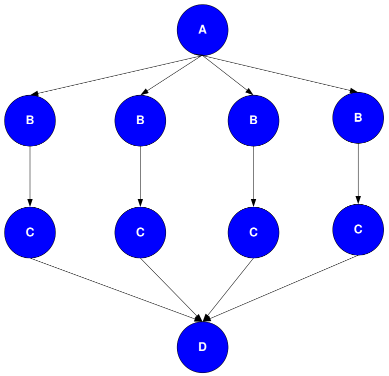
Pegasus Workflow Management System

- NSF funded project and developed since 2001 as a collaboration between USC Information Sciences Institute and the HTCondor Team at UW Madison

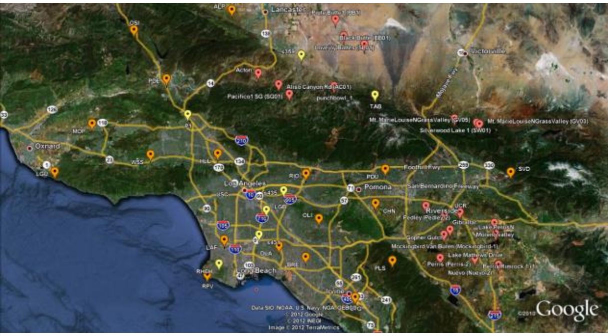
- Builds on top of HTCondor DAGMan.

- Abstract Workflows - Pegasus input workflow description
  - Workflow "high-level language"
  - Only identifies the computation, devoid of resource descriptions, devoid of data locations

- Pegasus is a workflow "compiler" (plan/map)
  - Target is DAGMan DAGs and HTCondor submit files
  - Transforms the workflow for performance and reliability
  - Automatically locates physical locations for both workflow components and data
  - Collects runtime provenance

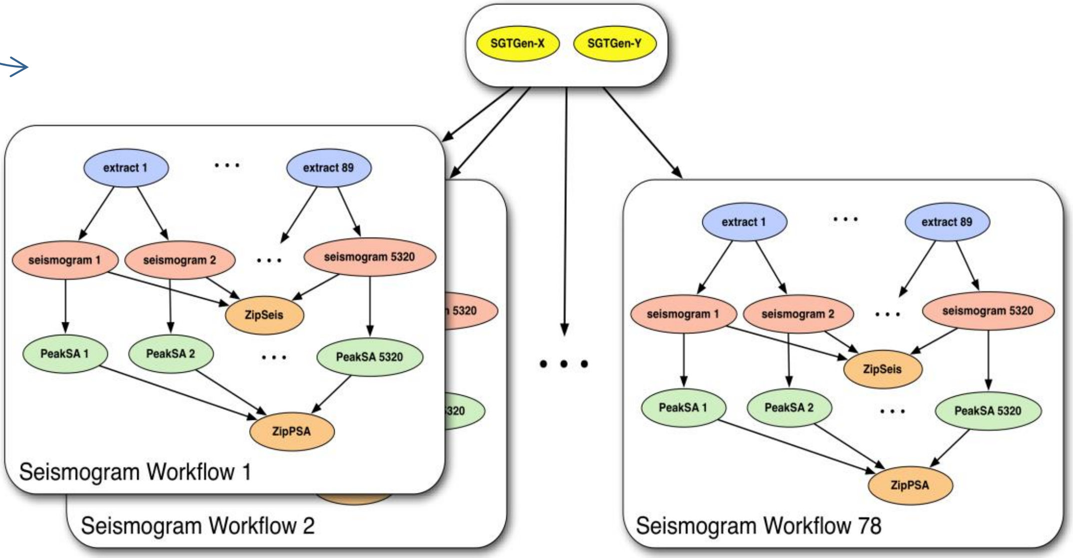


Southern California Earthquake Center's "CyberShake" workflow



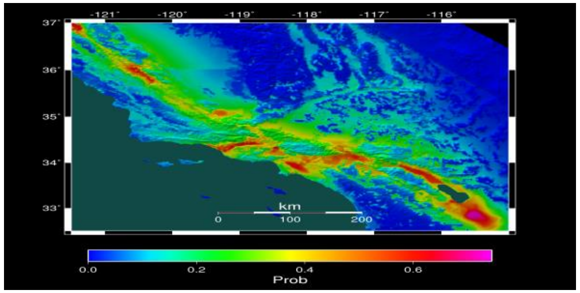
Description

- 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)



- Hierarchal workflow with mix of large MPI jobs (top) and large number of serial tasks (post processing)
- The size of the workflow depends on the number of sites and sampling frequency
- For each of the 600 sites in the input map, generate a hazard curve
- Each site has a sub-workflow with 820,000 tasks

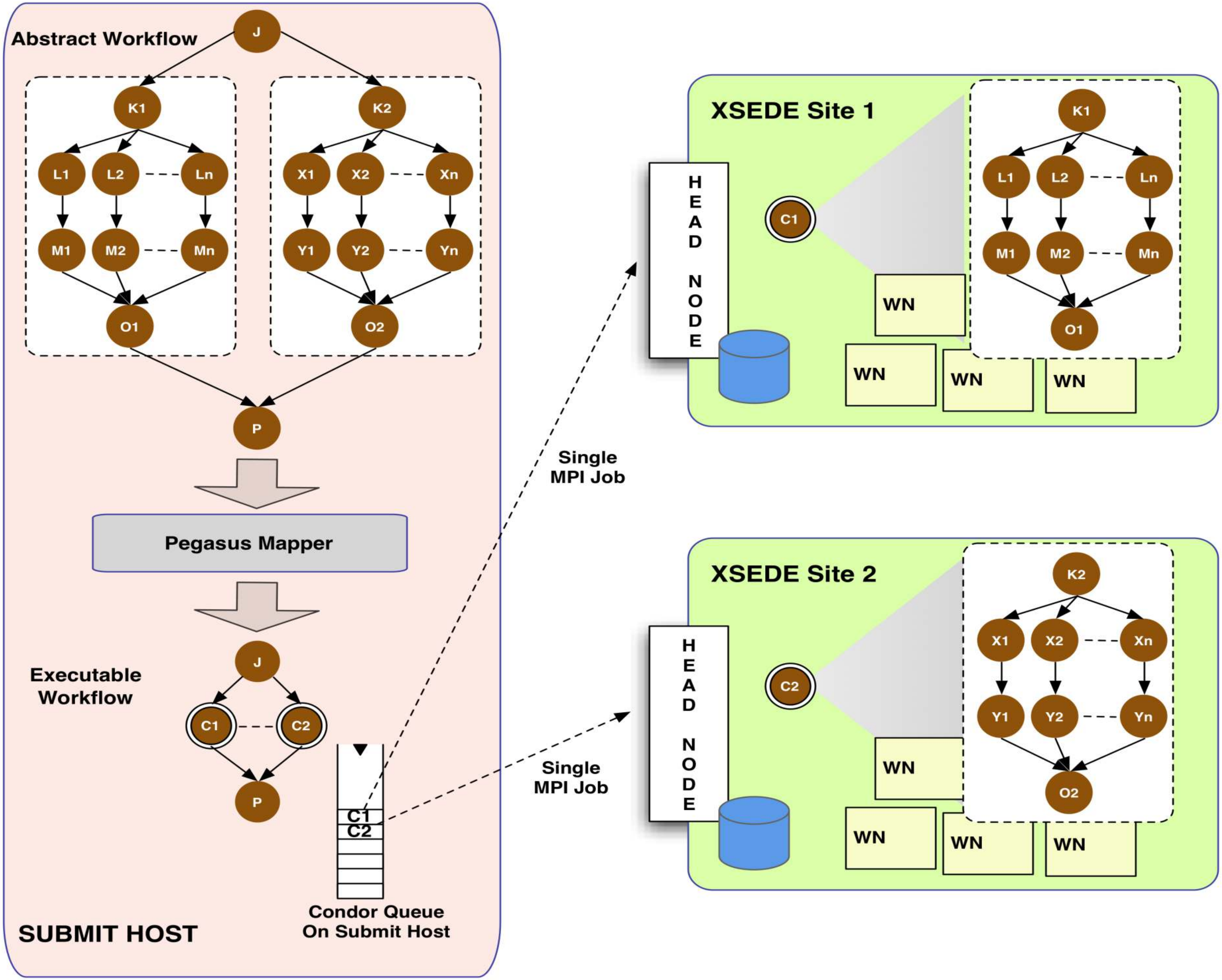
- SGT (Strain Green Tensor - MPI jobs) output: **15.6 TB** (40 * 400 GB)
- Final outputs: **500 million files** (820,000/site x 600 sites) **5.8 TB** (600 * 10 GB)



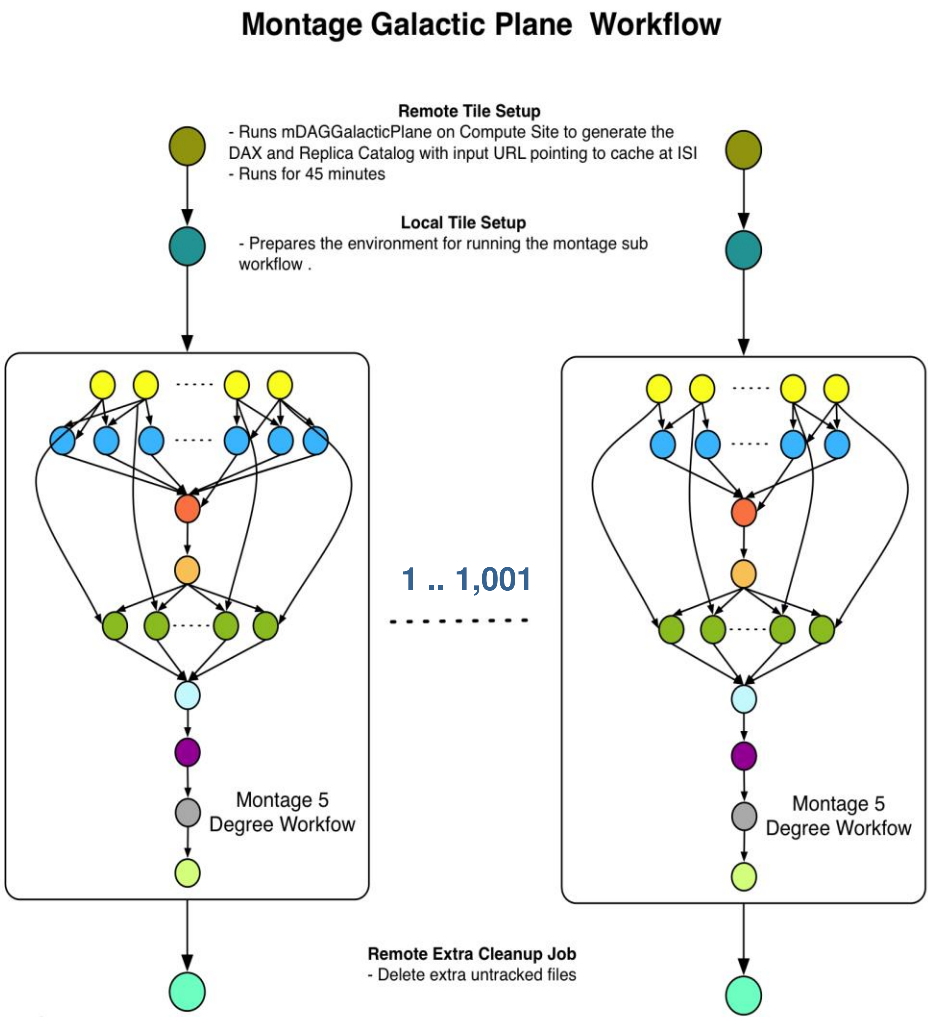
Probability of exceeding 0.1g in 50 yrs

pegasus-mpi-cluster

- Master/worker paradigm
- Master manages the subgraph tasks, handing out work to the workers
- Efficient scheduling / handling of input/outputs
- Subgraph described in a DAG-similar format
- Failure management / rescue DAG


Abstract Workflow
J
K1
L1
L2
Ln
M1
M2
Mn
O1
K2
X1
X2
Xn
Y1
Y2
Yn
O2
P
Pegasus Mapper
Executable Workflow
J
C1
C2
P
C1
C2
Condor Queue On Submit Host
SUBMIT HOST
Single MPI Job
Single MPI Job
XSEDE Site 1
HEAD NODE
C1
WN
WN
WN
WN
K1
L1
L2
Ln
M1
M2
Mn
O1
XSEDE Site 2
HEAD NODE
C2
WN
WN
WN
WN
K2
X1
X2
Xn
Y1
Y2
Yn
O2

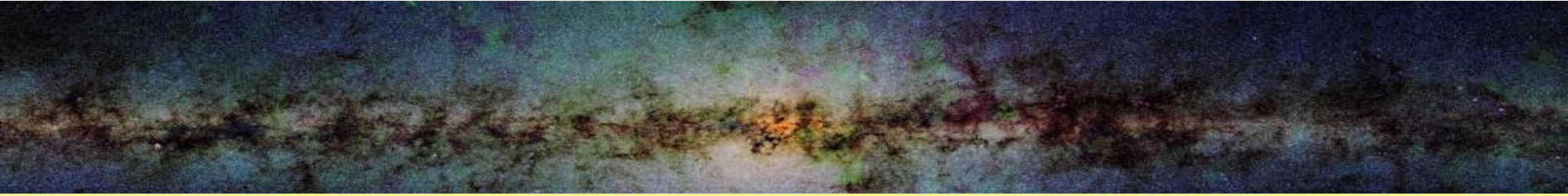
GALACTIC PLANE WORKFLOW


Montage Galactic Plane Workflow
Remote Tile Setup
- Runs mDAGGalacticPlane on Compute Site to generate the DAX and Replica Catalog with input URL pointing to cache at ISI
- Runs for 45 minutes
Local Tile Setup
- Prepares the environment for running the montage sub workflow .
1 .. 1,001
Montage 5 Degree Workflow
Montage 5 Degree Workfow
Remote Extra Cleanup Job
- Delete extra untracked files

- Galactic Plane for generating mosaics from the NASA Telescope Missions like Spitzer etc.
- Used to generate tiles 360 x 40 degrees around the galactic equator
- A tile 5 x 5 with 1 overlap with neighbors
- Output datasets to be potentially used in NASA Sky and Google Sky
- Each per band workflow
  - **1.6 million** input files
  - **10.5 million** tasks
  - Consumes **34,000 CPU hours**
  - Generates **1,001 tiles** in FITS format

× 16

- **Ran on XSEDE, moved to Amazon EC2**
- Run workflows corresponding to each of the 16 bands (wavelengths)
- Total Number of Data Files – **18 million**
- Potential Size of Data Output – **45 TB**



Acknowledgements: Bruce Berriman, John Good at IPAC - http://montage.ipac.caltech.edu/gallery.html

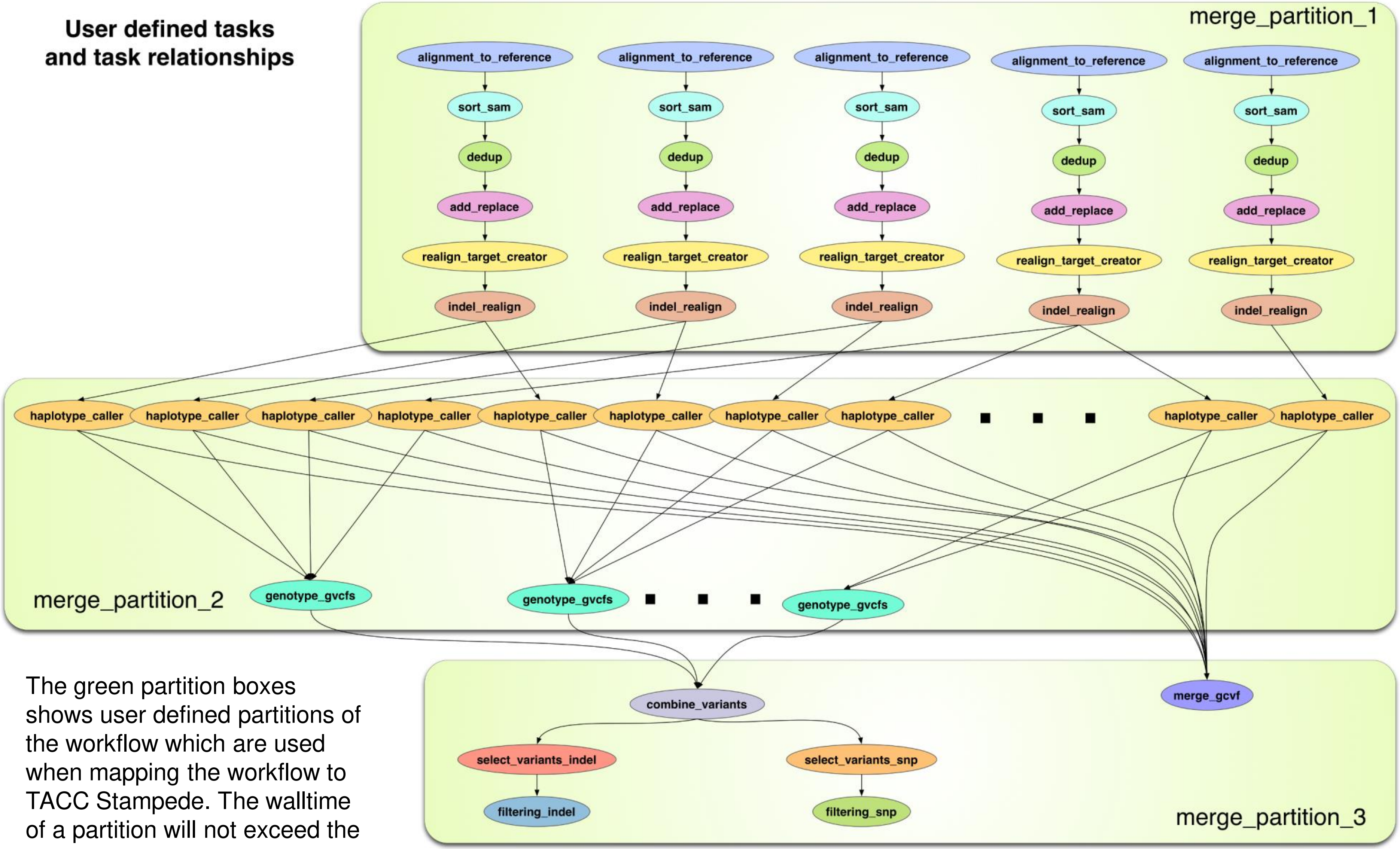
SoyKB – Soybean Knowledge Base

- Bioinformatics analysis of 1000+ resequenced soybean germplasm lines selected for major traits including oil, protein, soybean cyst nematode resistance (SCN), abiotic stress resistance (drought, heat and salt) and root system architecture.

- Sequencing pipeline using standard tools: Picard, BWA, GATK

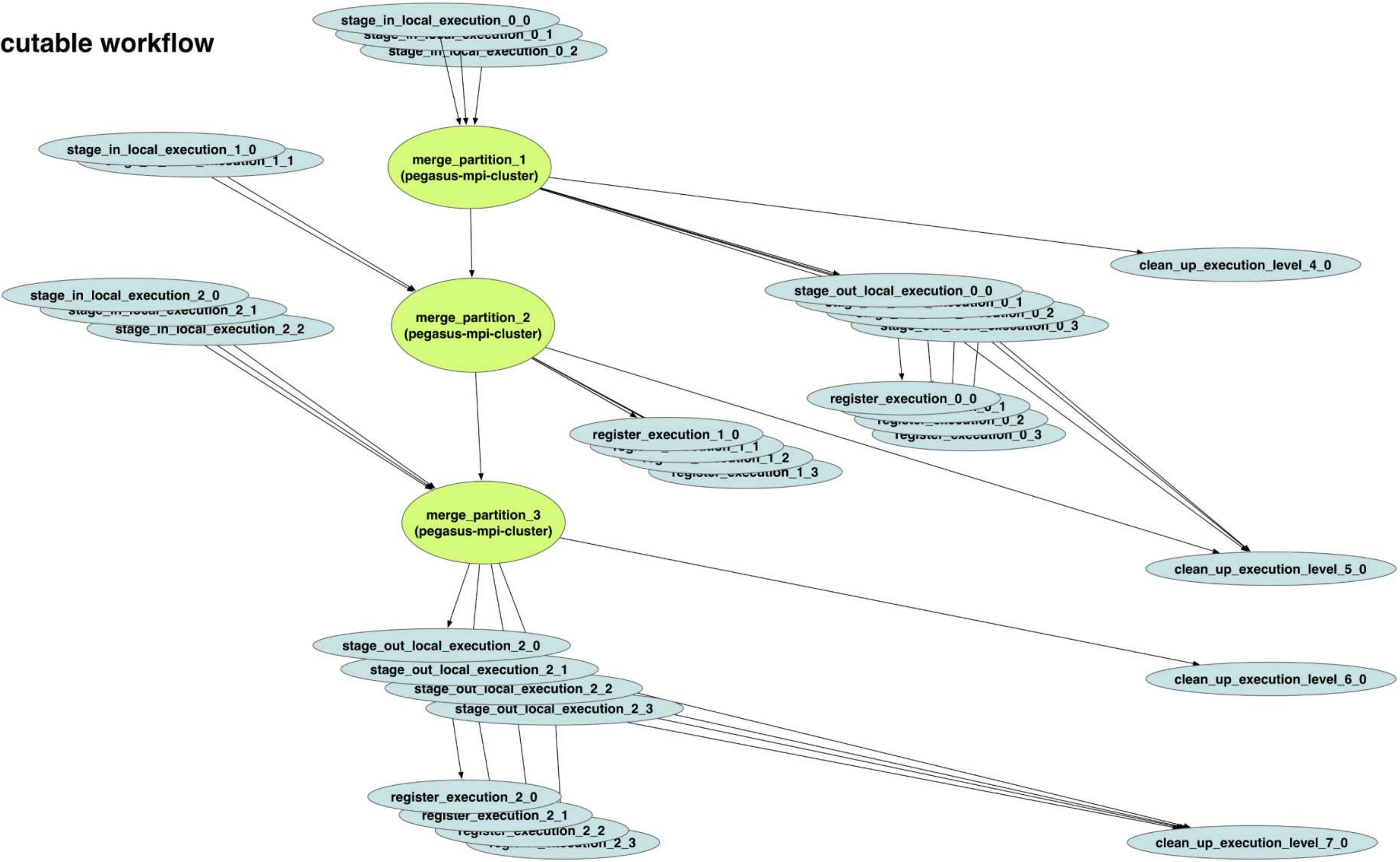
- Environment
  - iPlant Data Store for inputs and outputs
  - iPlant Atmosphere for VMs for workflow management
  - TACC Stampede as execution environment

User defined tasks and task relationships



The green partition boxes shows user defined partitions of the workflow which are used when mapping the workflow to TACC Stampede. The walltime of a partition will not exceed the 48 hour wall clock limit, given a certain number of compute nodes.

Executable workflow


stage_in_local_execution_0_0
stage_in_local_execution_0_1
stage_in_local_execution_0_2
stage_in_local_execution_1_0
stage_in_local_execution_1_1
merge_partition_1
(pegasus-mpi-cluster)
clean_up_execution_level_4_0
stage_in_local_execution_2_0
stage_in_local_execution_2_1
stage_in_local_execution_2_2
merge_partition_2
(pegasus-mpi-cluster)
stage_out_local_execution_0_0
stage_out_local_execution_0_1
stage_out_local_execution_0_2
stage_out_local_execution_0_3
register_execution_0_0
register_execution_0_1
register_execution_0_2
register_execution_0_3
register_execution_1_0
register_execution_1_1
register_execution_1_2
register_execution_1_3
merge_partition_3
(pegasus-mpi-cluster)
clean_up_execution_level_5_0
stage_out_local_execution_2_0
stage_out_local_execution_2_1
stage_out_local_execution_2_2
stage_out_local_execution_2_3
clean_up_execution_level_6_0
register_execution_2_0
register_execution_2_1
register_execution_2_2
register_execution_2_3
clean_up_execution_level_7_0

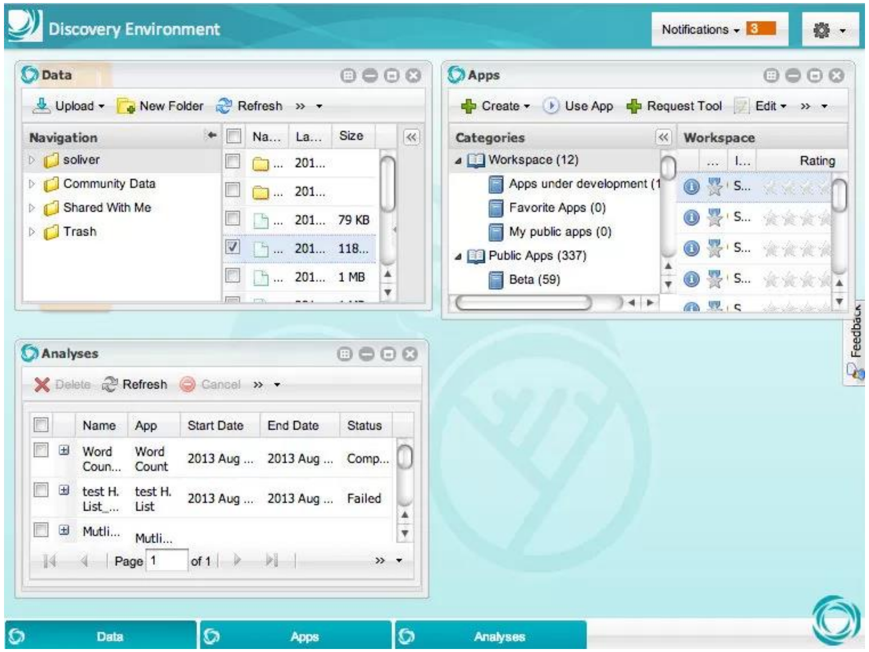
The green jobs are MPI container jobs of the partitions shown in the earlier slide. These are run on the remote supercomputer. The blue auxiliary tasks are run on the workflow submit host, and handle things like data staging and cleanup.

SoyKB - Highlights

- Workflow can be mapped to a distributed or a supercomputer execution environment
  - Distributed environment is used for testing the pipeline
  - The supercomputer environment is used for production runs, and tasks are packed up in MPI container jobs (pegasus-mpi-cluster)

- Input data is fetched from iPlant Data Store, or if already replicated, from the TACC iPlant Data Store node for close to computation access

- Outputs are automatically put back into the Data Store for easy access and further analysis in the iPlant Discovery Environment



☐	Na...	La...	Size
☐	...	201...	
☐	...	201...	
☐	...	201...	79 KB
☑	...	201...	118...
☐	...	201...	1 MB

☐		Name	App	Start Date	End Date	Status
☐	⊞	Word Coun...	Word Count	2013 Aug ...	2013 Aug ...	Comp...
☐	⊞	test H. List_...	test H. List	2013 Aug ...	2013 Aug ...	Failed
☐	⊞	Mutli...	Mutli...			

Page 1 of 1

Relevant Links

http://pegasus.isi.edu

Tutorial and documentation:
http://pegasus.isi.edu/wms/docs/latest/

Mats Rynge   rynge@isi.edu