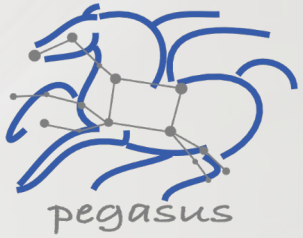


Living in a Heterogenous World: How scientific workflows bridge diverse cyberinfrastructure

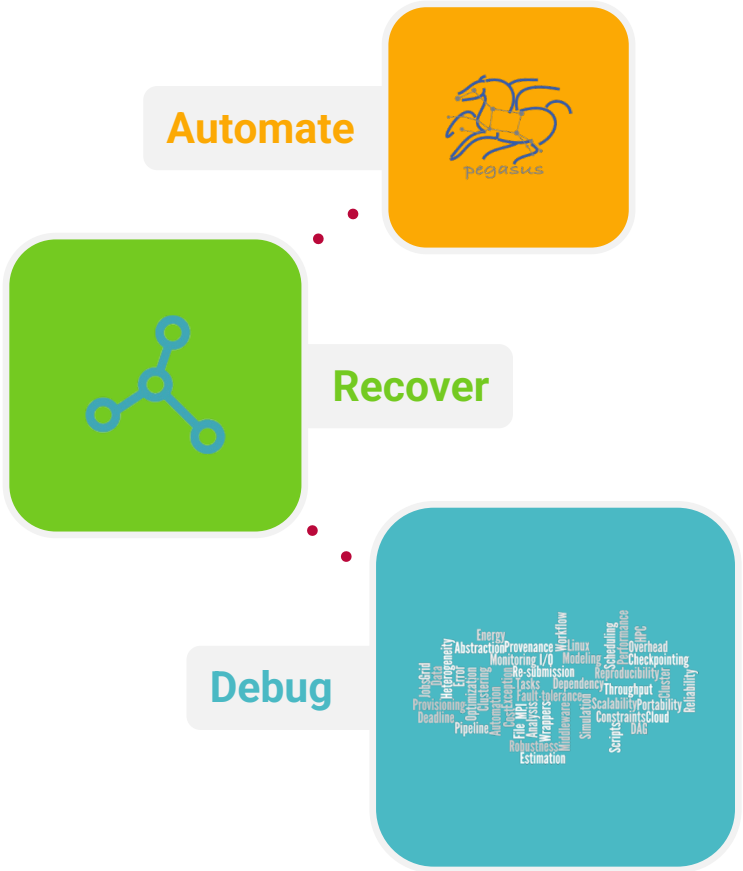
Ewa Deelman

University of Southern California, School of Engineering
Information Sciences Institute
deelman@isi.edu

This work is funded by NSF awards #2018074 and #1664162.

USC Viterbi School of Engineering Information Sciences Institute

Pegasus Workflow Management System

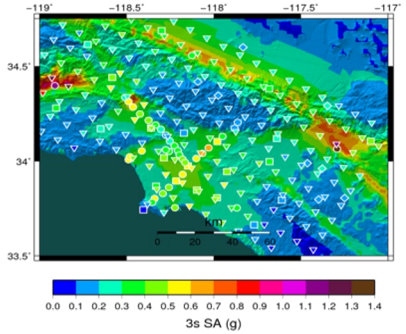


- **Automates Complex**, Multi-stage Processing Pipelines
- Enables Parallel, **Distributed Computations**
- **Automatically Executes** Data Transfers
- Reusable, Aids **Reproducibility**
- Records How Data was Produced **(Provenance)**
- Handles **Failures** with to Provide Reliability
- Keeps Track of Data and **Files**

HTCondor High Throughput Computing

NSF funded project since 2001,
with close collaboration with HTCondor team

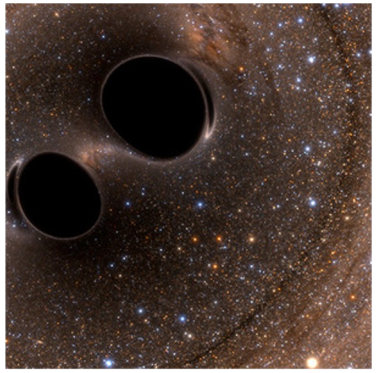
Southern California Earthquake Center's CyberShake



Mix of MPI and single-core jobs, mix of CPU, GPU codes.
Large data sets (10s of TBs), ~300 workflows with 420,000 tasks each
Supported since 2005: changing CI, x-platform execution

First Physics-Based "Shake map" of Southern California

Laser Interferometer Gravitational-Wave Observatory (LIGO)

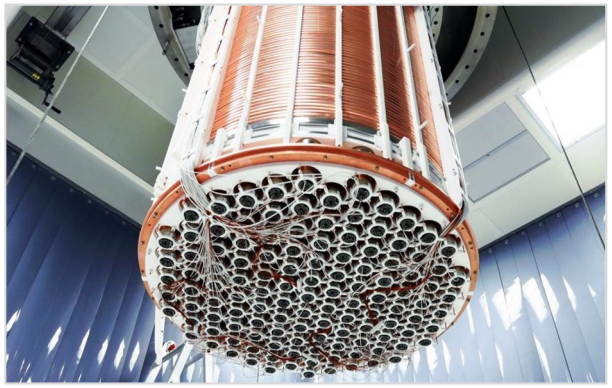


High-throughput computing workload, access to HPC resources, ~ 21K Pegasus workflows, ~ 107M tasks

Supported since 2001, distributed data, opportunistic computing resources

First direct detection of a gravitational wave (colliding black holes)

XENONnT - Dark Matter Search



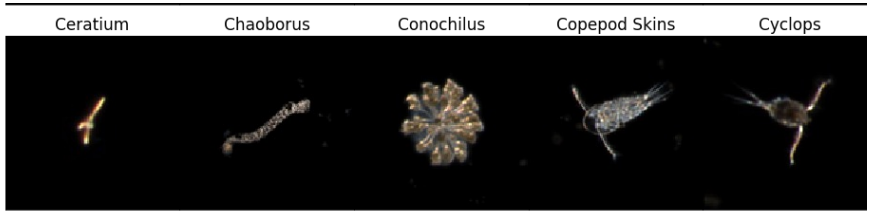
Custom data management
Rucio for data management
MongoDB instance to track science runs and data products.

Monte Carlo simulations and the main processing pipeline.

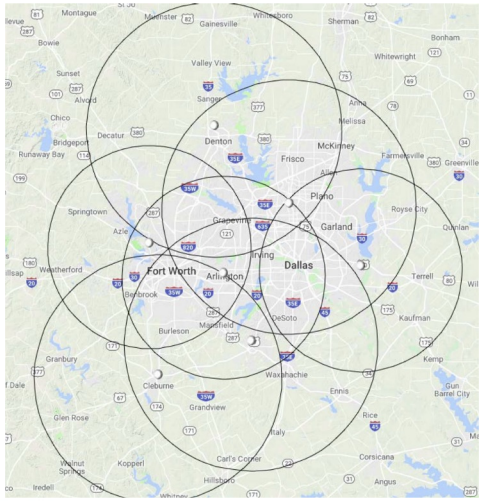
Edge Instruments and Sensors

Sensor data are increasingly used across science* domains

- Zooplankton classification
- Weather forecasting
- Study of the marine life



https://www.frontiersin.org/articles/10.3389/fmicb.2021.746297/full



CASA: Collaborative and Adaptive Sensing of the Atmosphere

- Has deployed a network of short-range Doppler radars
- Compute and data repositories at the edge, close to the radars
- Use on demand cloud resources to scale up their computations

http://www.casa.umass.edu/

OOI: Ocean Observatories Initiative

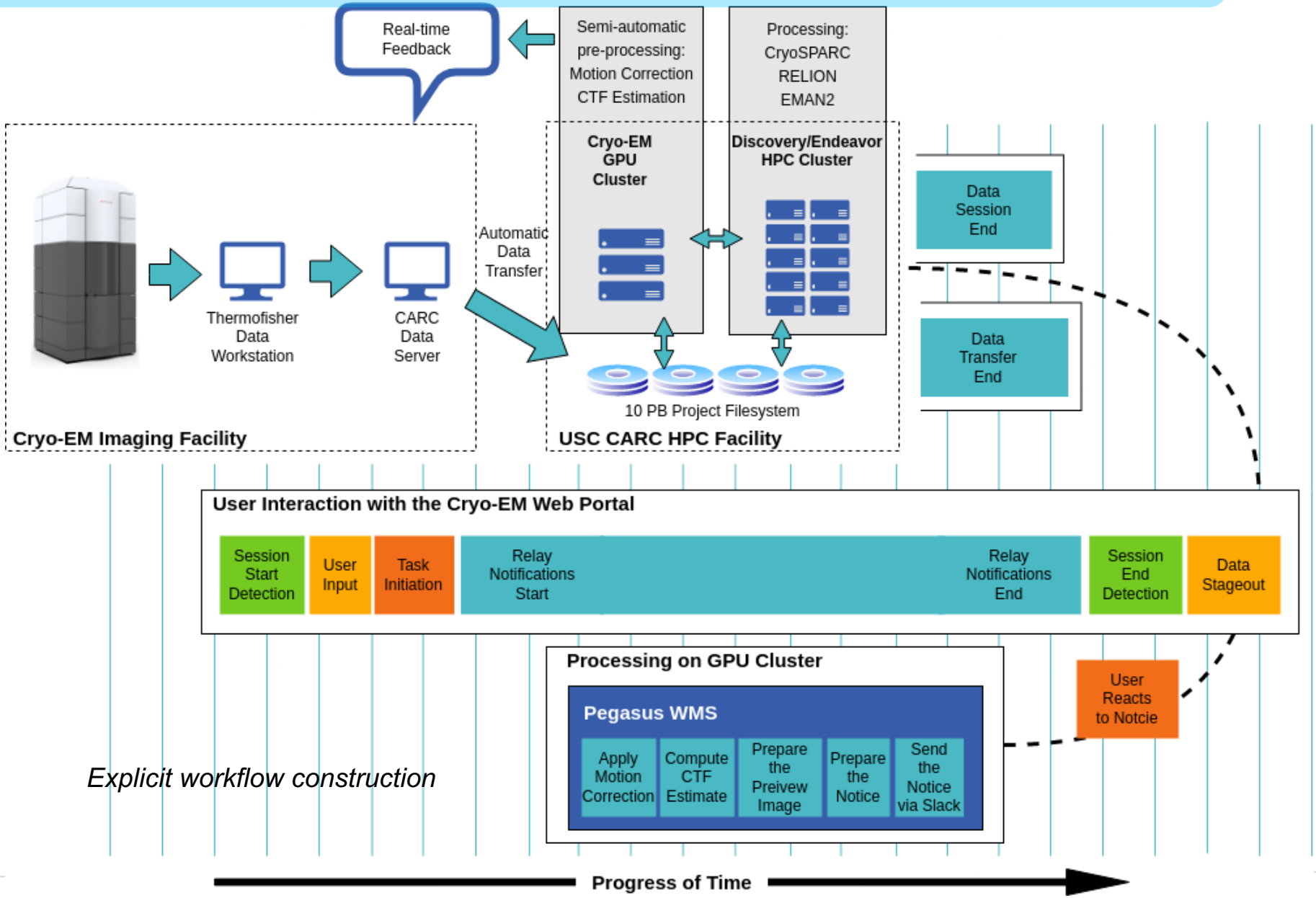
- Has deployed a variety of sensors in the Atlantic and the Pacific oceans to study the oceans and the marine life
- Hydrophone sensors have been deployed in three locations in the state of Washington
- The Orcasound community initiative is using them to study Orca whales in the Pacific Northwest region.

https://www.orcasound.net/

These applications need workflow technologies to orchestrate computations edge-to-cloud

* Work in progress

Processing instrument data in real time



Hidden workflows

Explicit workflow construction

Pegasus Workflow Management System

Workflow Challenges Across Domains

Describe complex workflows in a simple way

Access distributed, heterogeneous data and resources (heterogeneous interfaces)

Deal with resources/software that change over time

Ease of use. Ability to debug and monitor large workflows

Our Focus

- Separation between workflow description and workflow execution
- Workflow planning and scheduling (scalability, performance)
- Task execution (monitoring, fault tolerance, debugging, web dashboard)
- Workflow optimization, restructuring for performance and fault tolerance.

1. Resource-independent Specification

Input Workflow Specification YAML formatted

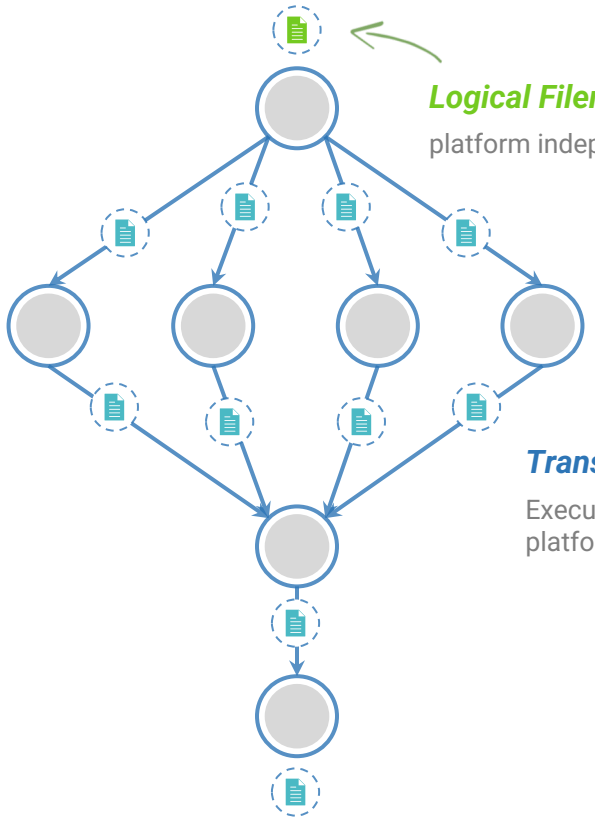
directed-acyclic graphs

Output Workflow

Portable Description

Users do not worry about low level execution details

ABSTRACT WORKFLOW

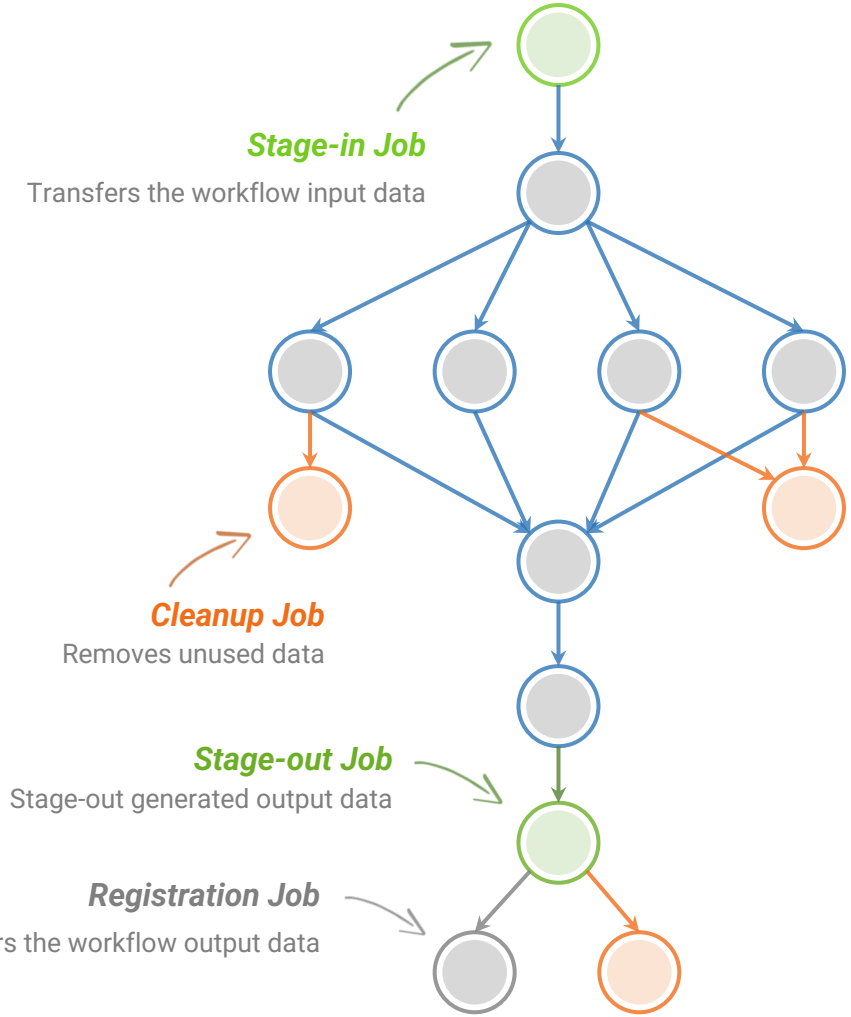


Logical Filename (LFN)

platform independent (abstraction)

Transformation

Executables (or programs)
platform independent



Stage-in Job

Transfers the workflow input data

Cleanup Job

Removes unused data

Stage-out Job

Stage-out generated output data

Registration Job

Registers the workflow output data

EXECUTABLE WORKFLOW

Pegasus

2. Submit locally, run globally

- **Pegasus WMS ==** **Pegasus planner (mapper) + DAGMan workflow engine + HTCondor scheduler/broker**

  Pegasus maps workflows to target infrastructure (1 or more resources)

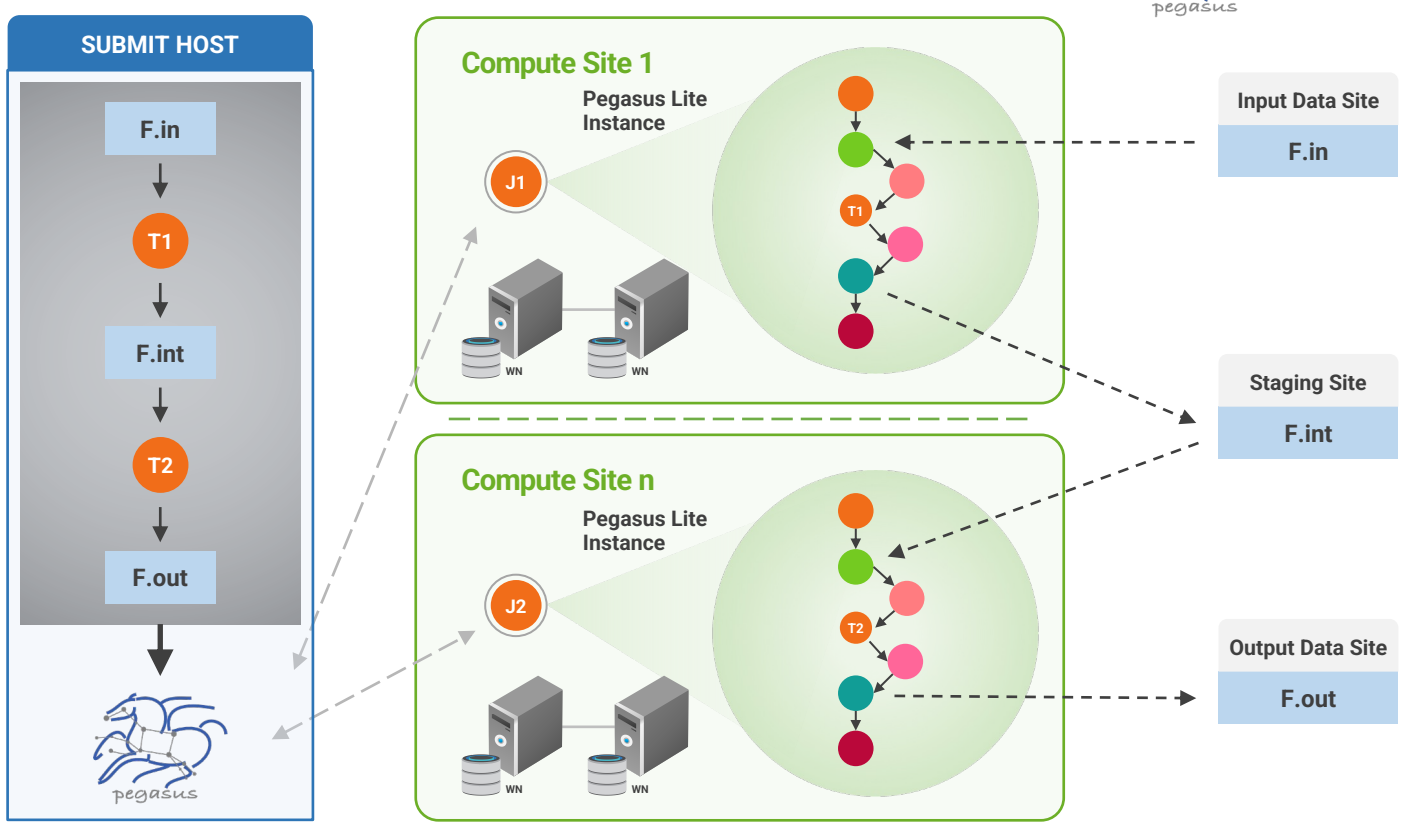
  DAGMan manages dependencies and reliability

  HTCondor is used as a broker to interface with different schedulers

- **Planning converts an abstract workflow into a concrete, executable workflow**

  Planner is like a compiler
  Optimized performance
  Provides fault tolerance

- **Can leverage distributed and heterogeneous CI**



3. Flexible Data Staging Configurations

HTCondor I/O (HTCondor pools, OSG, …)

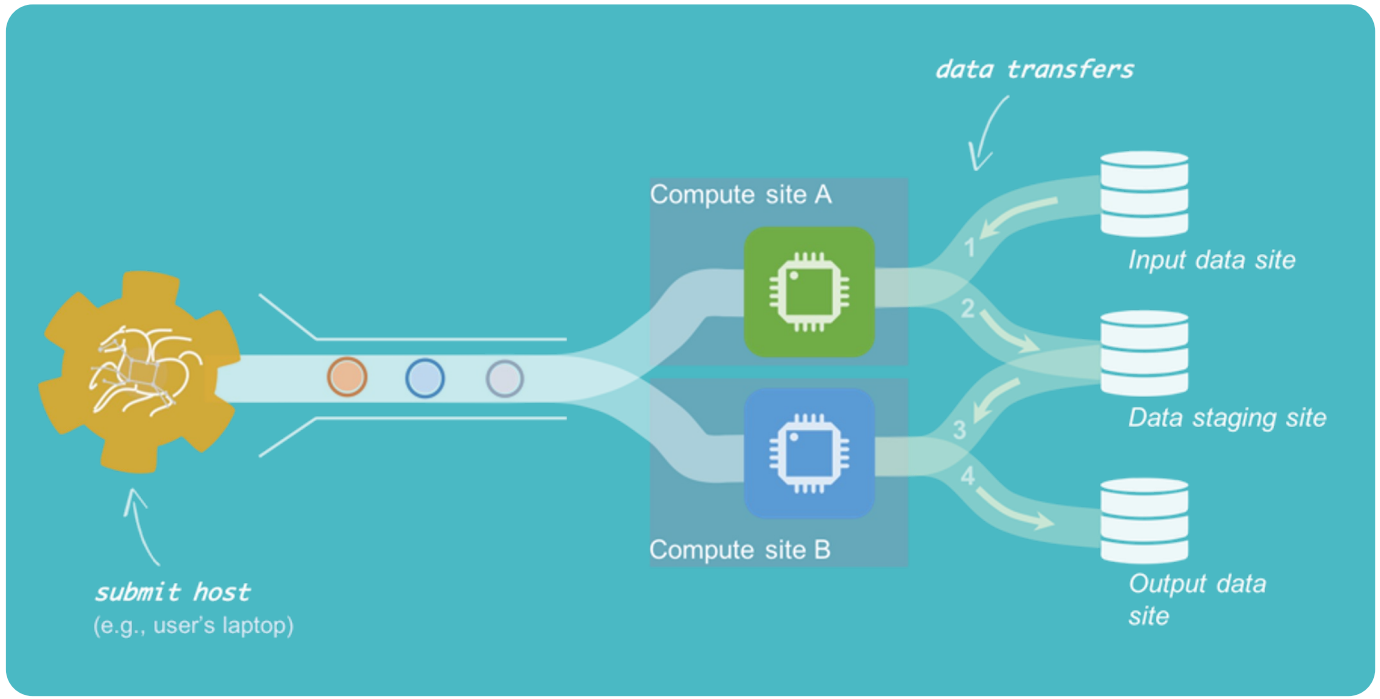
Worker nodes do not share a file system
Data is pulled from / pushed to the submit host via HTCondor file transfers
Staging site is the submit host

Non-shared File System (clouds, OSG, …)

Worker nodes do not share a file system
Data is pulled / pushed from a staging site, possibly not co-located with the computation

Shared File System (HPC sites, XSEDE, Campus clusters, …)

I/O is directly against the shared file system



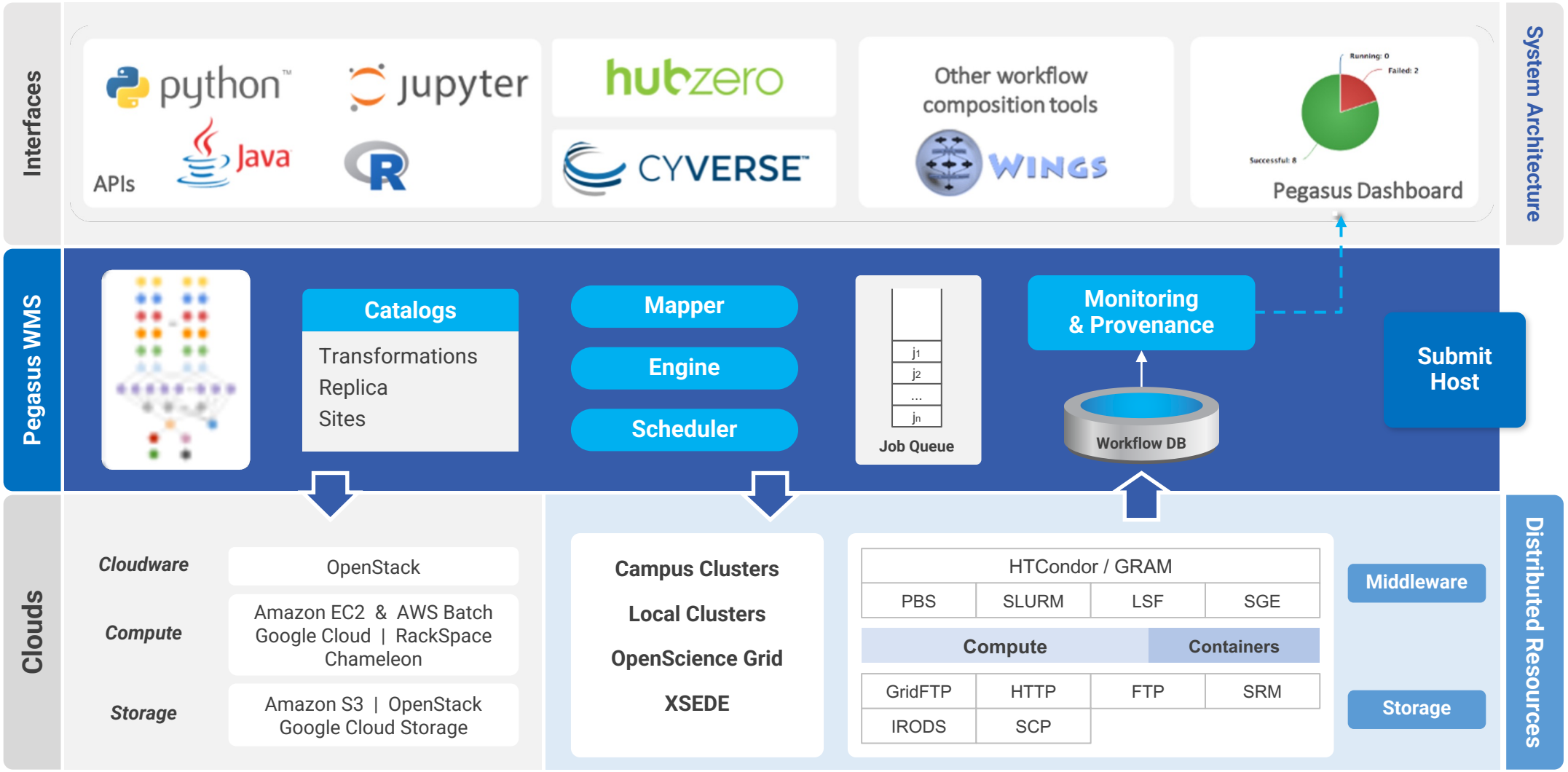
4. Flexible Data movement Pegasus-transfer

Pegasus' internal data transfer tool with support for a number of different protocols

- **Directory creation, file removal**
  If protocol can support it, also used for cleanup
- **Two stage transfers**
  e.g., GridFTP to S3 = GridFTP to local file, local file to S3
- **Parallel transfers**
- **Automatic retries**
- **Credential management**
  Uses the appropriate credential for each site and each protocol (even 3rd party transfers)

```
HTTP
SCP
GridFTP
Globus
Online
iRods
Amazon S3
Google
Storage
SRM
FDT
Stashcp
Rucio
cp
ln -s
```

5. “Up and down” integrations with diverse CI, common languages, and Portal/GUI interfaces



Event Horizon Telescope

Bringing Black Holes into Focus

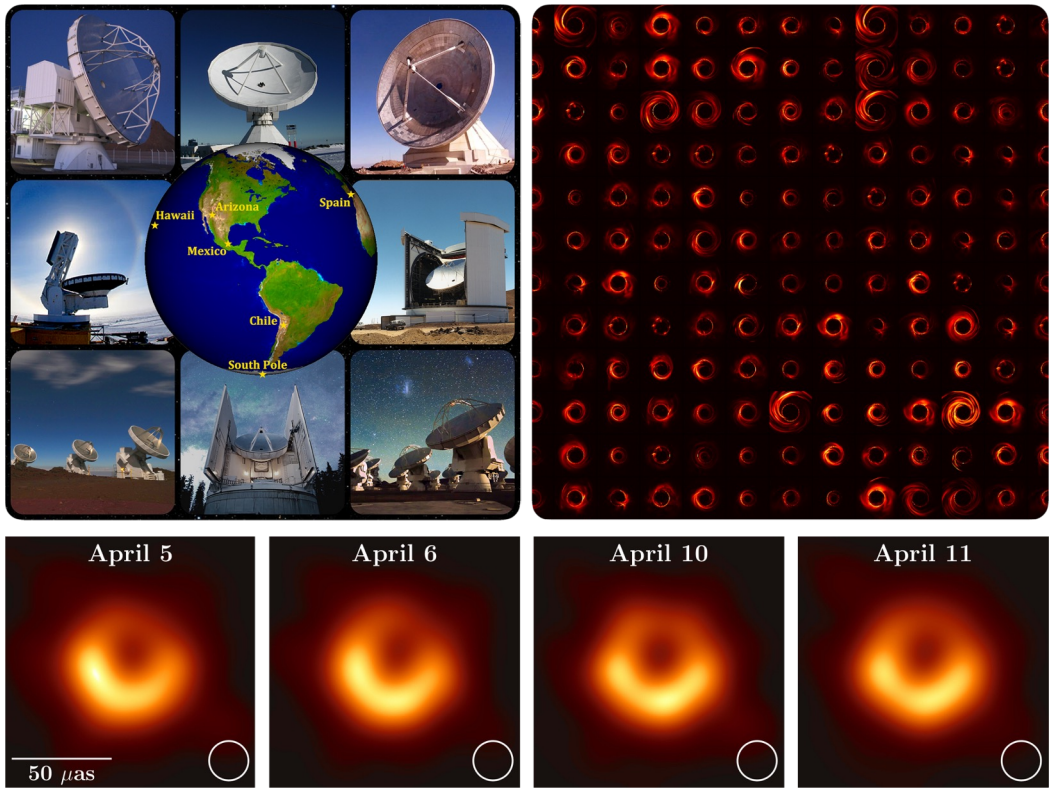
480,000 jobs - 2,600,000 core hours

#15 in all OSG projects in last 6 months

#2 in all OSG astronomy projects in the last 6 months

8 telescopes: 5 PB of data

60 simulations: 35 TB data

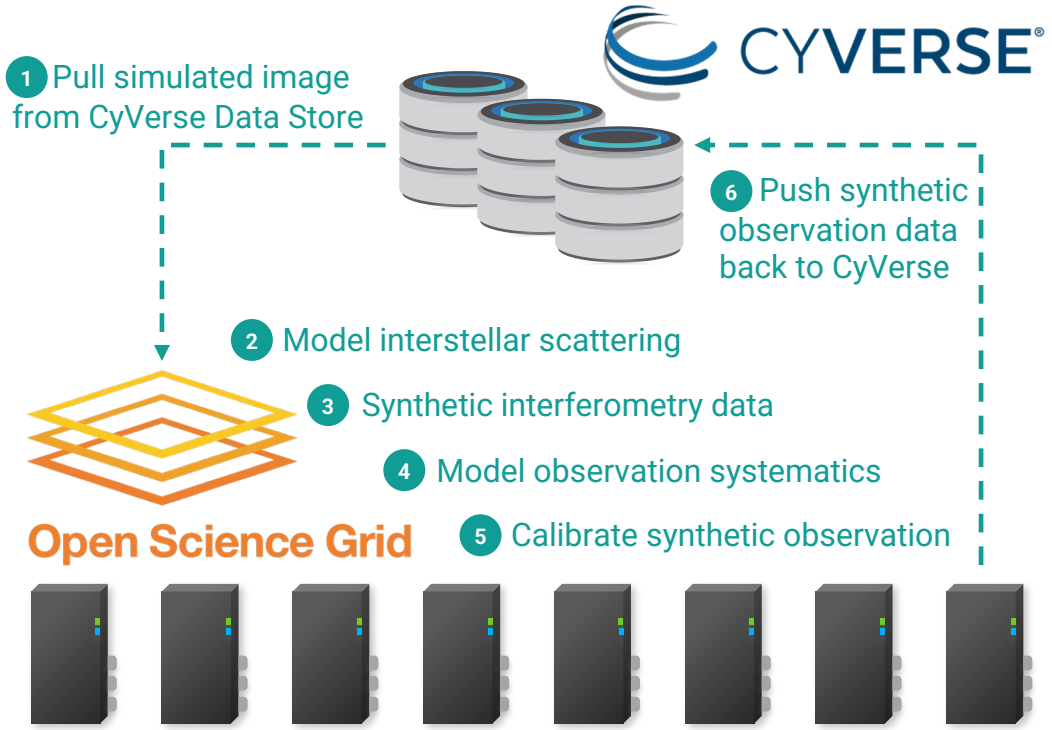


First images of black hole at the center of the M87 galaxy

Improve constraints on Einstein's theory of general relativity by 500x

Pegasus-SYMBA Pipeline

Physically accurate synthetic observation data from simulations are keys to develop calibration and imaging algorithms, as well as comparing the observation with theory and interpreting the results.



USC Viterbi School of Engineering Information Sciences Institute

https://pegasus.isi.edu

XENONnT - Dark Matter Search

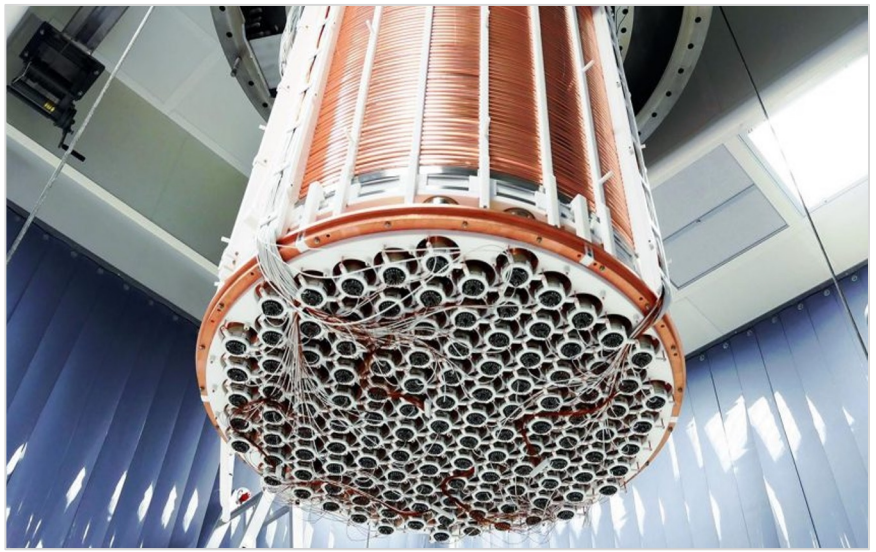
Two Workflows

Monte Carlo simulations and the main processing pipeline.

Workflows execute across Open Science Grid (OSG) & **European Grid Infrastructure (EGI)**

Rucio for data management

MongoDB instance to track science runs and data products.

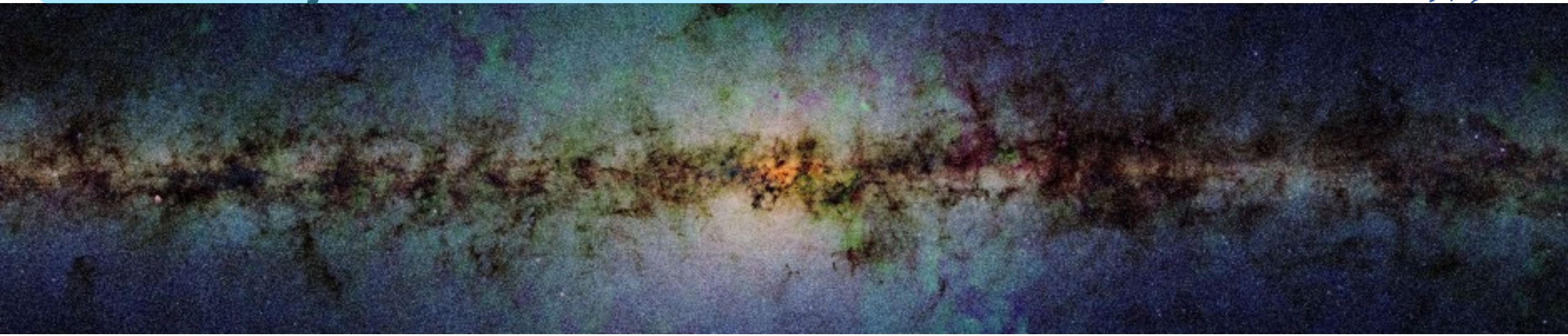


```
------------------------------------------------------------------------------
Type           Succeeded Failed  Incomplete  Total     Retries   Total+Retries
Tasks          4000      0       0           4000      267       4267
Jobs           4484      0       0           4484      267       4751
Sub-Workflows  0         0       0           0         0         0
------------------------------------------------------------------------------

Workflow wall time                                       : 5 hrs, 2 mins
Cumulative job wall time                                 : 136 days, 9 hrs
Cumulative job wall time as seen from submit side        : 141 days, 16 hrs
Cumulative job badput wall time                          : 1 day, 2 hrs
Cumulative job badput wall time as seen from submit side : 4 days, 20 hrs
```

The challenge is to manage the workflows accessing different data management systems.

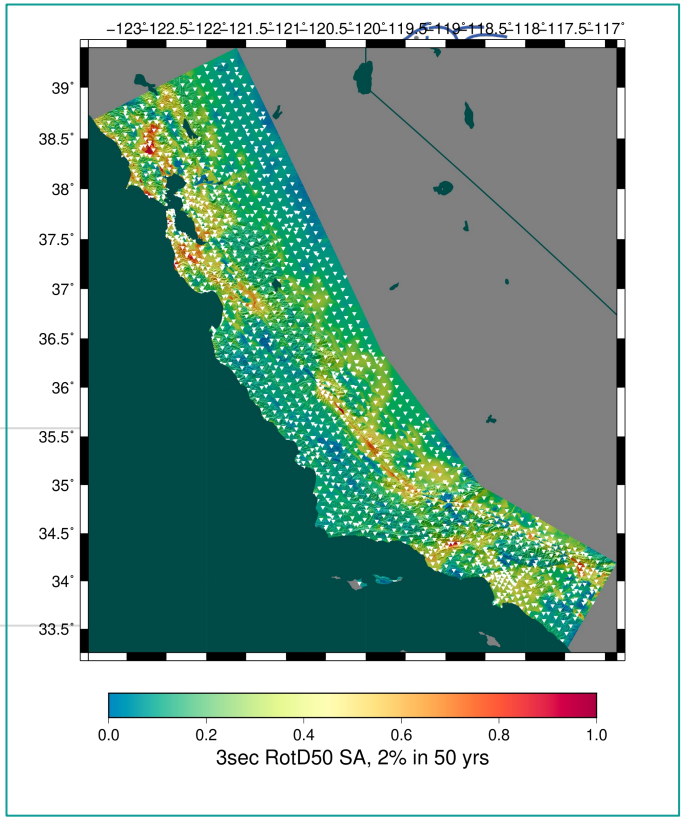
Community Archives: Galactic Plane Atlas



- 18 million input images (~2.5 TB)
- 900 output images (2.5 GB each, 2.4 TB total)
- Measuring the global star formation rate in the galaxy
- Studying the energetics of the interaction of molecular clouds with the interstellar medium
- Determining whether coagulation or fragmentation governs the formation of massive stars
- Assessing the supernova rate in the Galaxy

} × 17

Ensemble Manager



Allow users to submit a collection of workflows (ensembles)

- Automatically **spawn** and **manage** collections of workflows

Trigger submission of workflows

Cron workflow trigger

File pattern workflow trigger

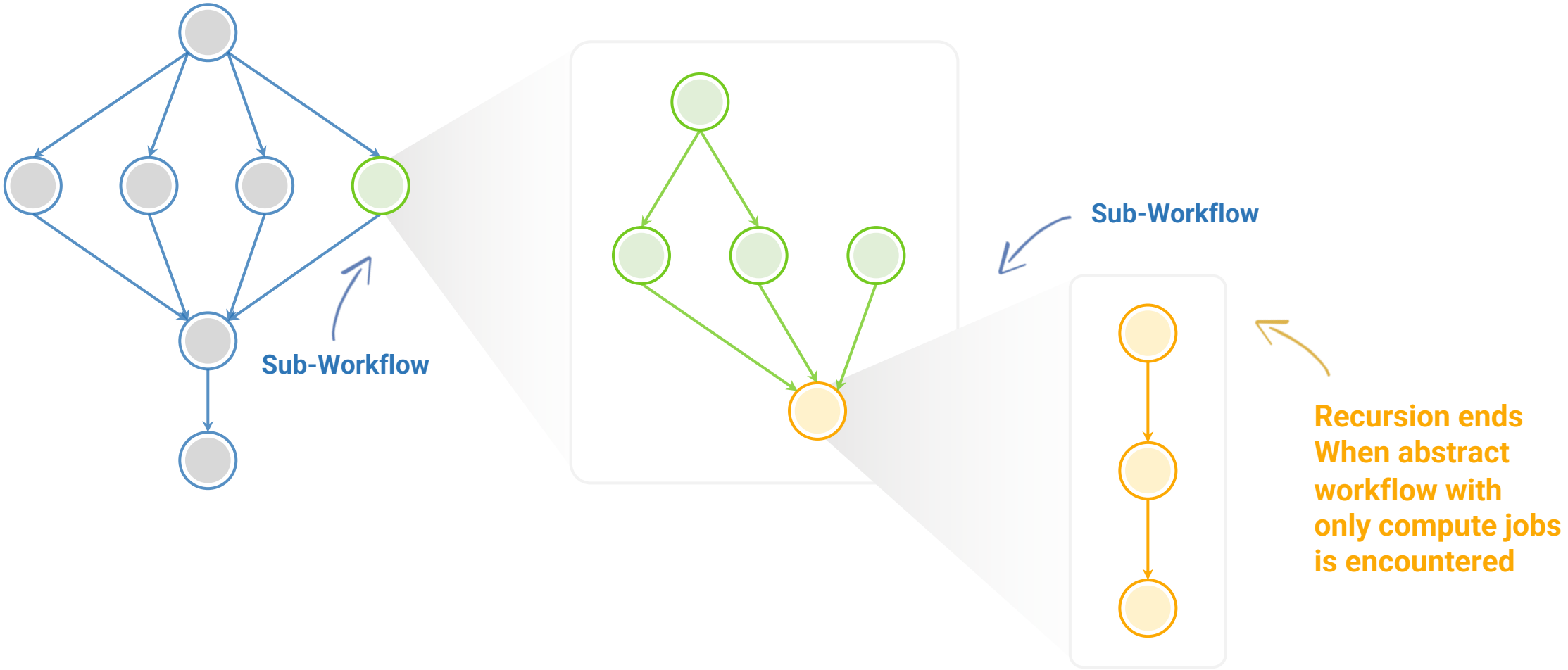
Properties

- Workflows within an ensemble may have **different priorities**
  - > *Priorities can also be changed at runtime*
- Ensembles may limit the number of **concurrent** planned and running workflows

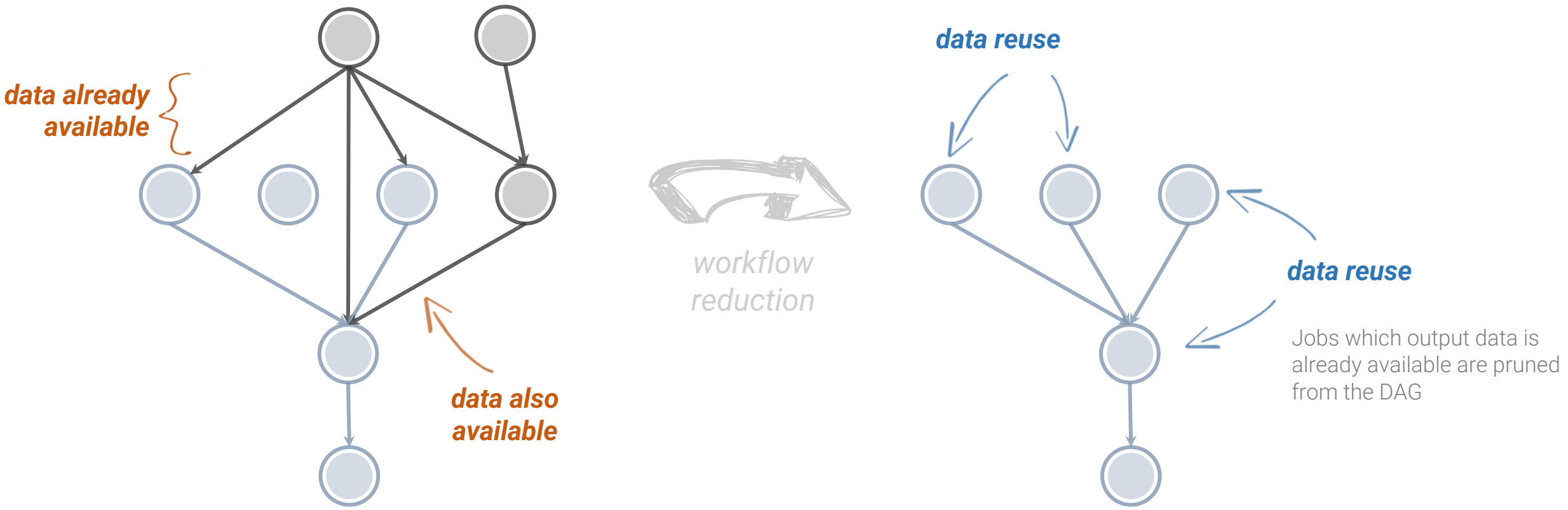
Additional Actions

- Ensembles can be **paused, resumed, removed, re-planned**, and **re-executed**
- A **debugging** mechanism is also provided to investigate failures in workflow runs
- Actions can be performed both to ensembles and single workflows within ensembles

Handling of large-scale workflows



Data Reuse prune jobs if output data already exists



Performance optimization, Fault recovery strategy

Types of Workflow Applications: **Supporting community-based analysis**

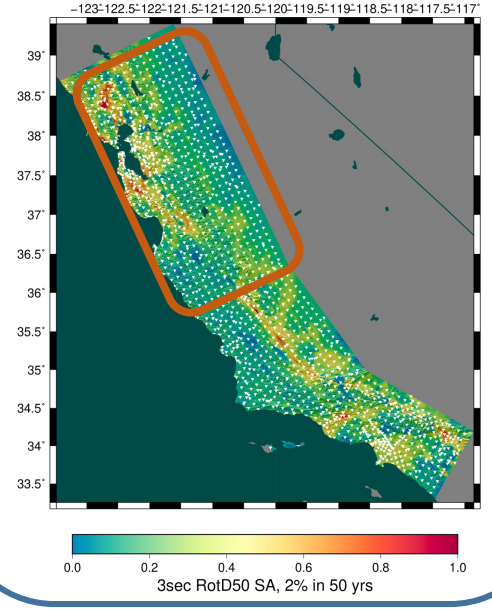
SCEC's CyberShake:
What will the peak earthquake motion be over the next 50 years?

Southern California Earthquake Center

- Codes are collaboratively developed
- Codes are "strung" together to model complex systems
- Ability to correctly connect components
- Automating the flow of data (instead of emails)
- Automatic fault recovery and support for scalability

Useful information for:

- Building engineers
- Disaster planners
- Insurance agencies



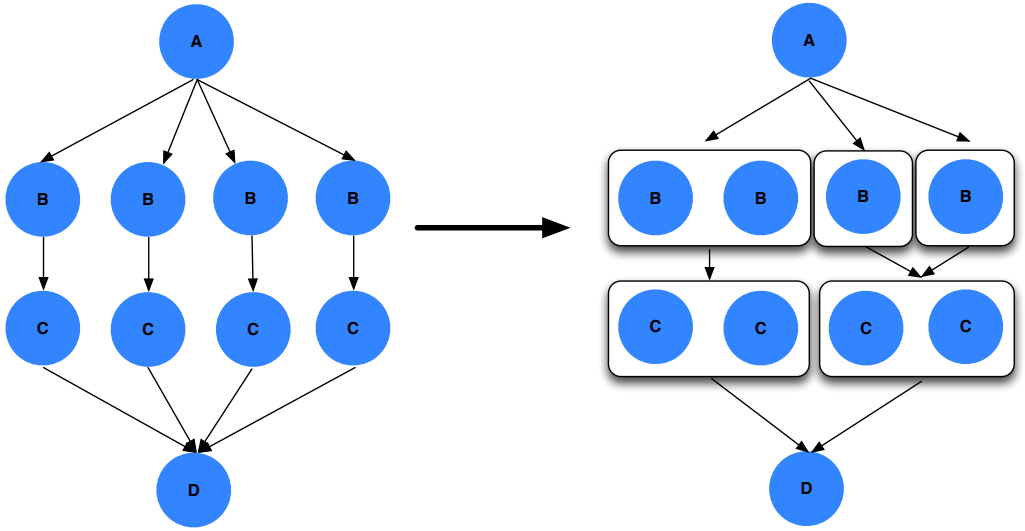
NCSA *Blue Waters*
OLCF *Titan*

- 120 million core-hours
- 39,285 jobs
- 1.2 PB of data managed
- 157 TB of data automatically transferred
- 14.4 TB of output data archived

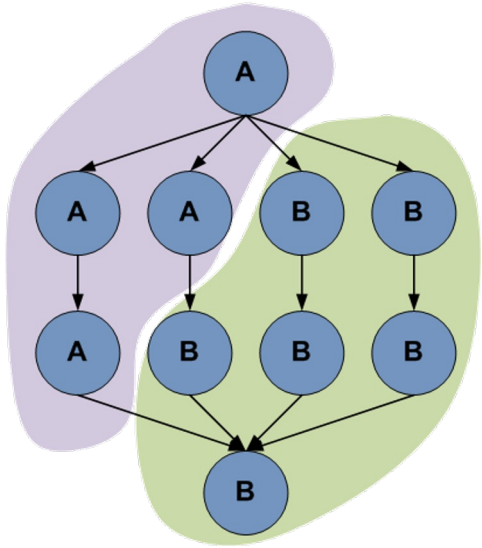
Slide credit: USC's Southern California Earthquake Center

Increasing computational granularity:

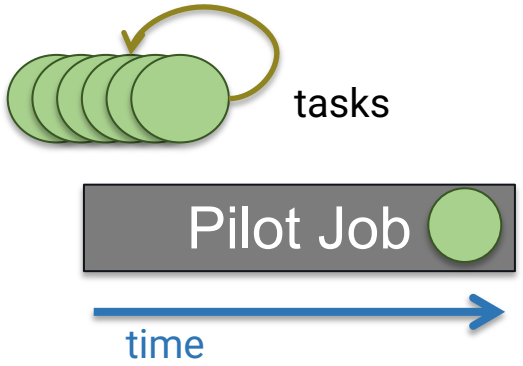
Task clustering



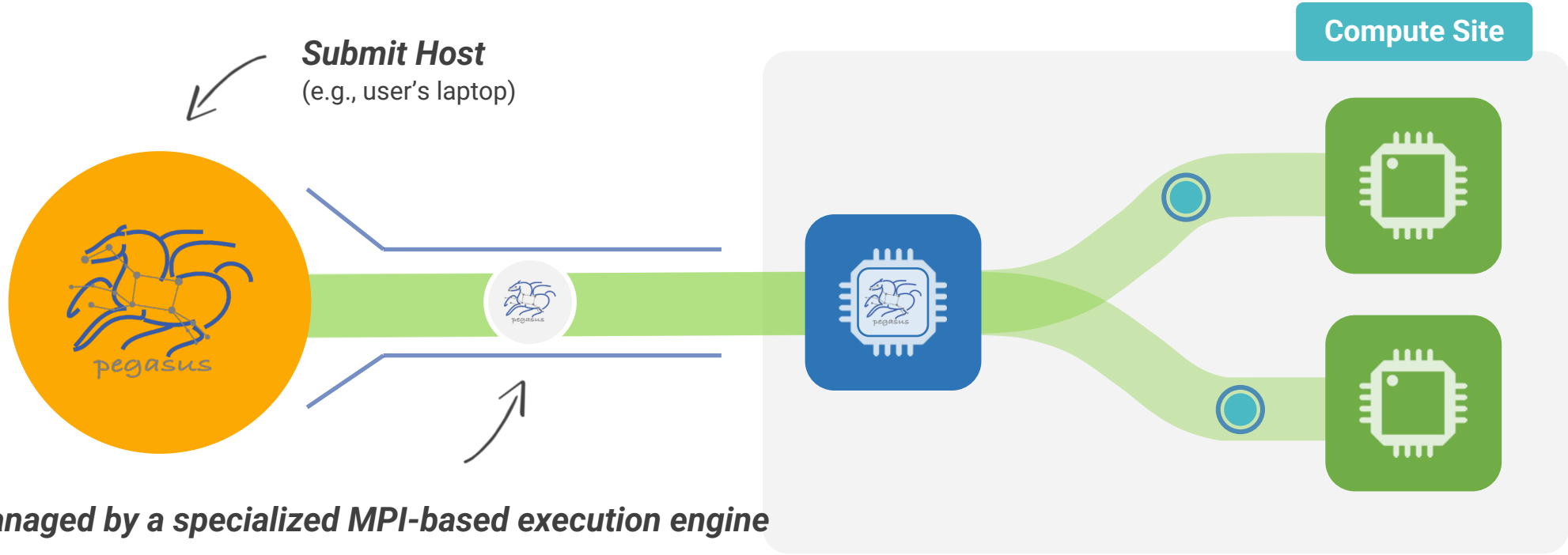
Partition the workflow into sub-workflows and send them for execution to the target system



Use “pilot” jobs to dynamically provision a number of resources at a time



Handling heterogeneous workloads: Running HTC jobs on HPC systems…



Workflow managed by a specialized MPI-based execution engine

Allows sub-graphs of a Pegasus workflow to be submitted as monolithic jobs to remote resources

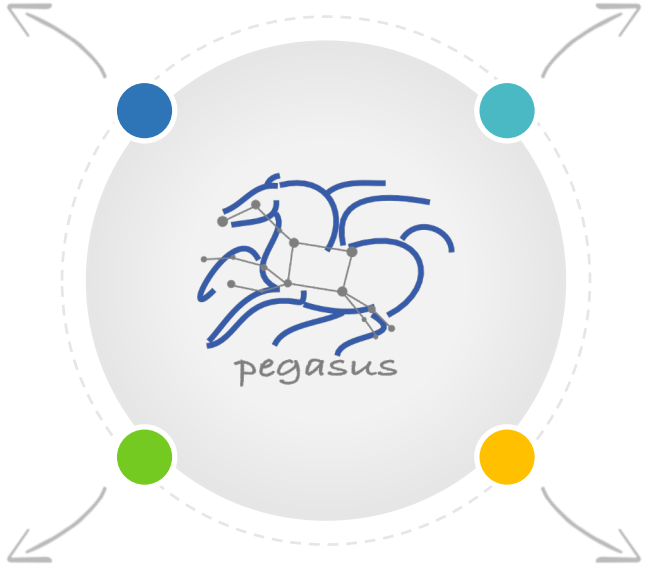
Better fault tolerance

Postscript

detects non-zero exit code output parsing for success or failure message exceeded timeout do not produced expected output files

Job Retry

helps with transient failures set number of retries per job and run



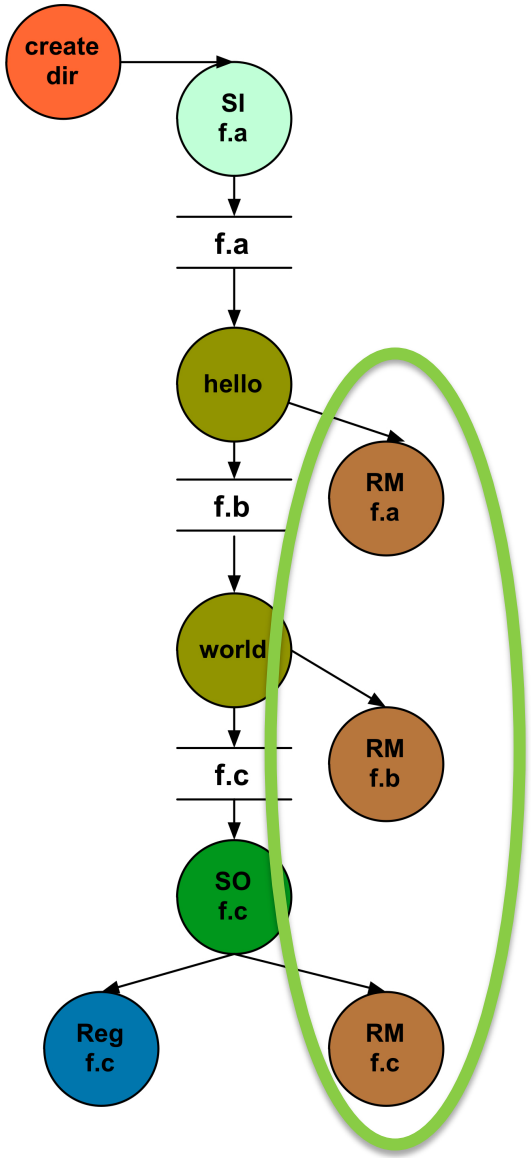
Checkpoint Files

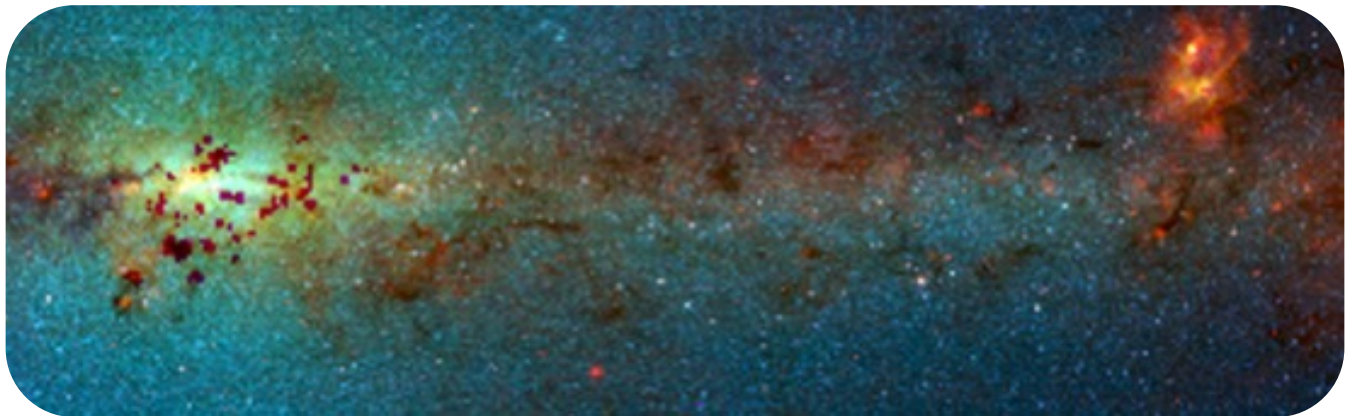
job generates checkpoint files staging of checkpoint files is automatic on restarts

Rescue DAGs

workflow can be restarted from checkpoint file recover from failures with minimal loss

Montage Astronomy Workflow

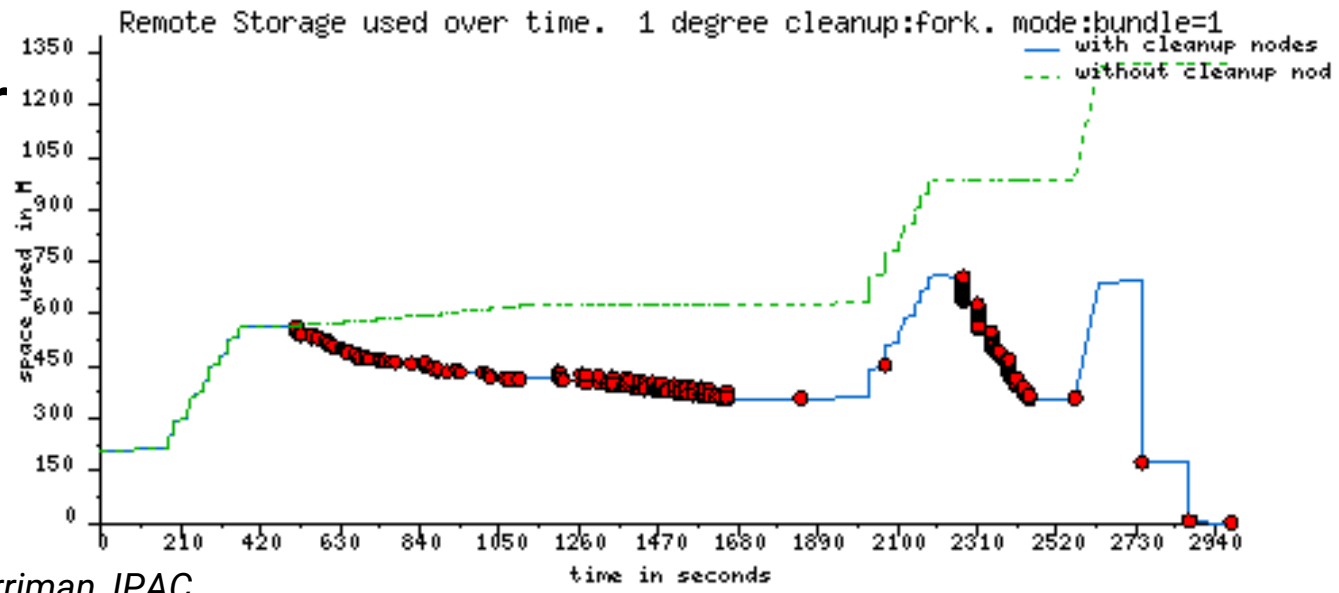




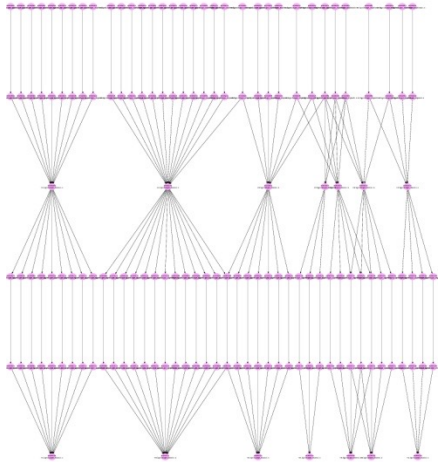
2Mass Mosaic, IPAC, Caltech

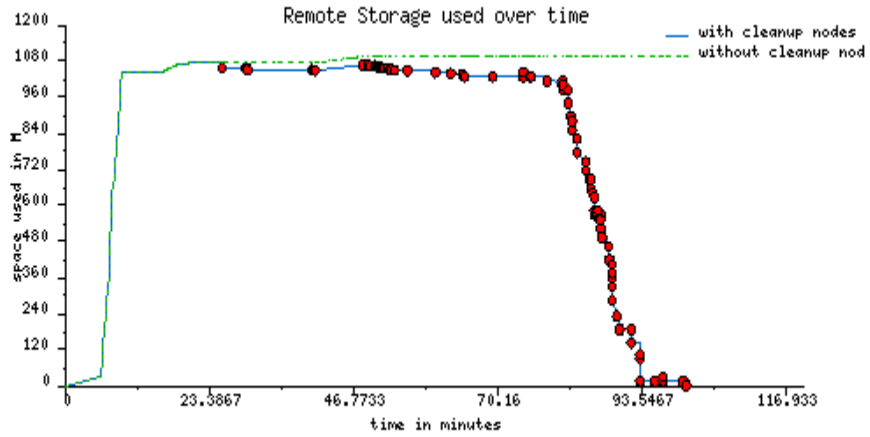
Automatically add tasks to "clean up" data no longer needed

1.25GB versus 700 MB



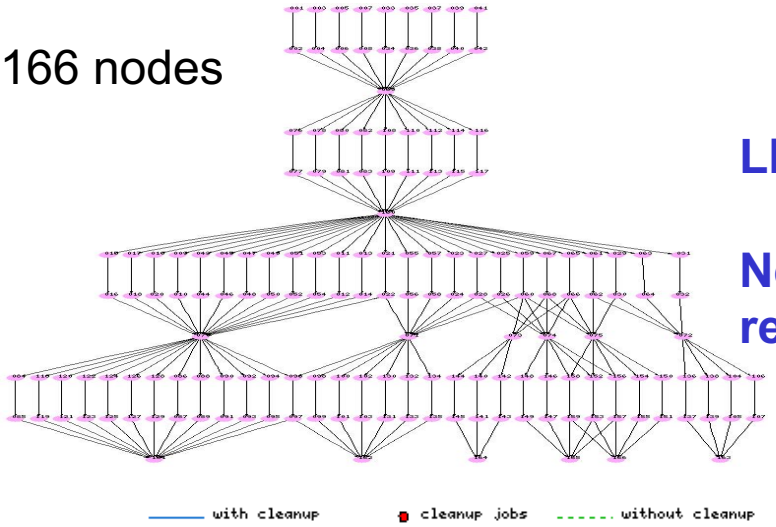
Collaboration with Rizos Sakellariou, U Manchester and Bruce Berriman, IPAC

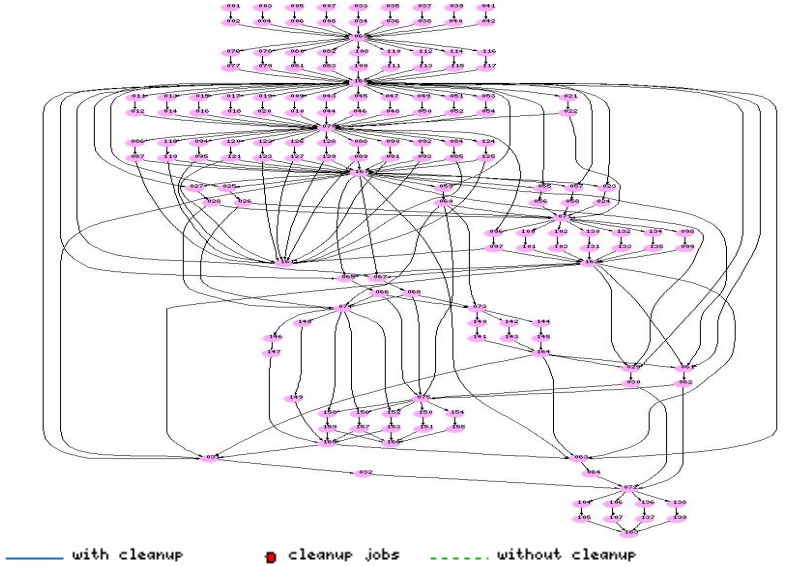




Full workflow:
185,000 nodes 466,000 edges
10 TB of input data
1 TB of output data.

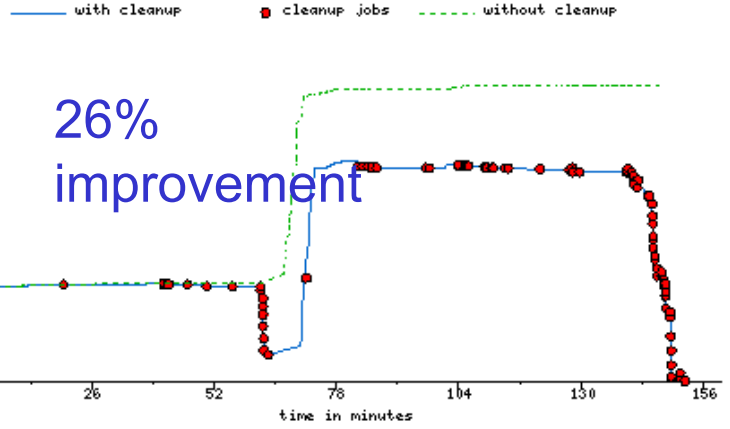
166 nodes



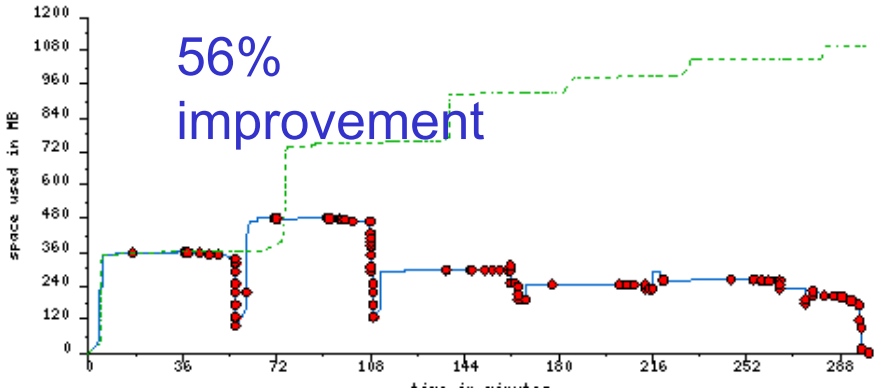


LIGO Workflows

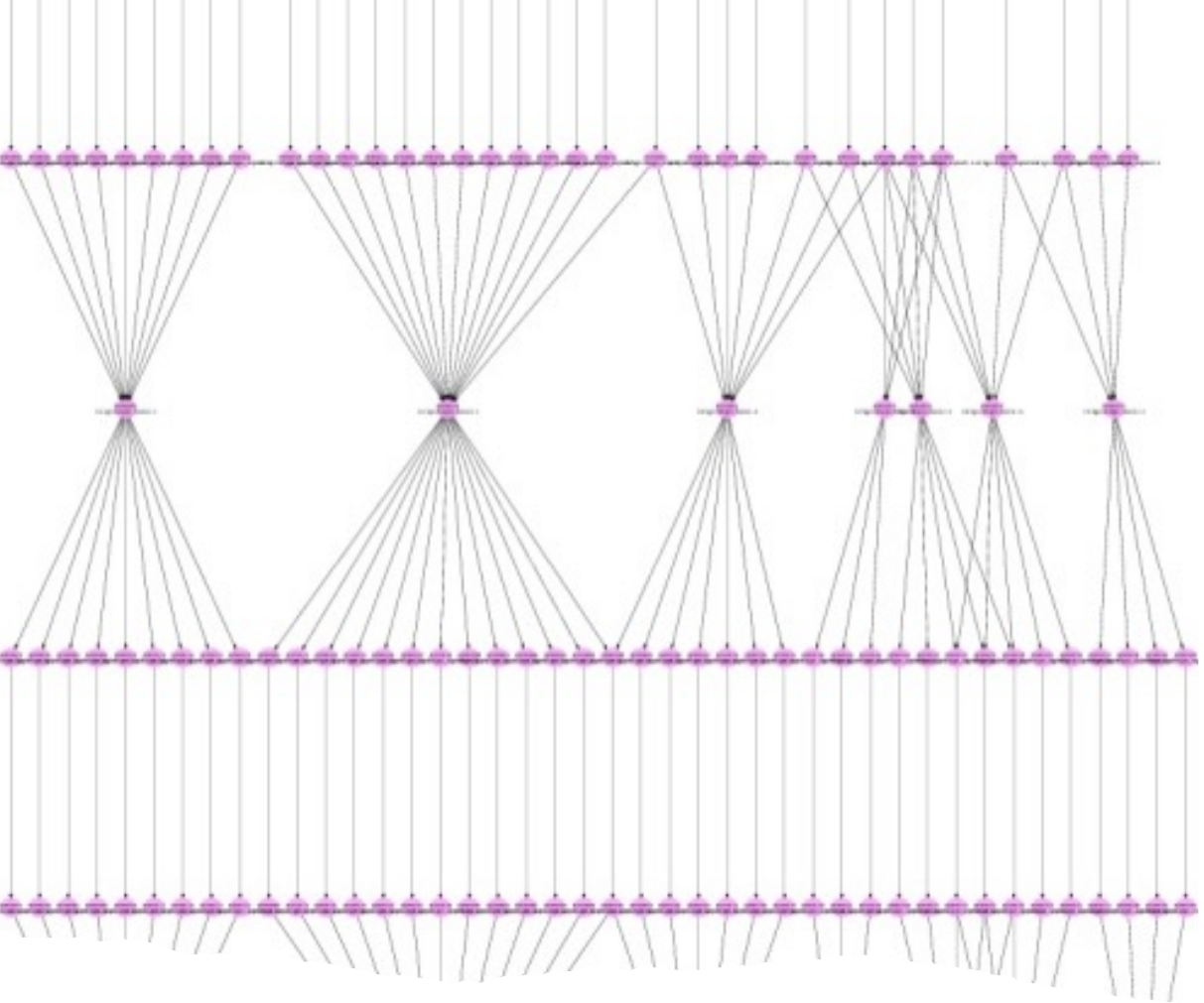
Need additional restructuring



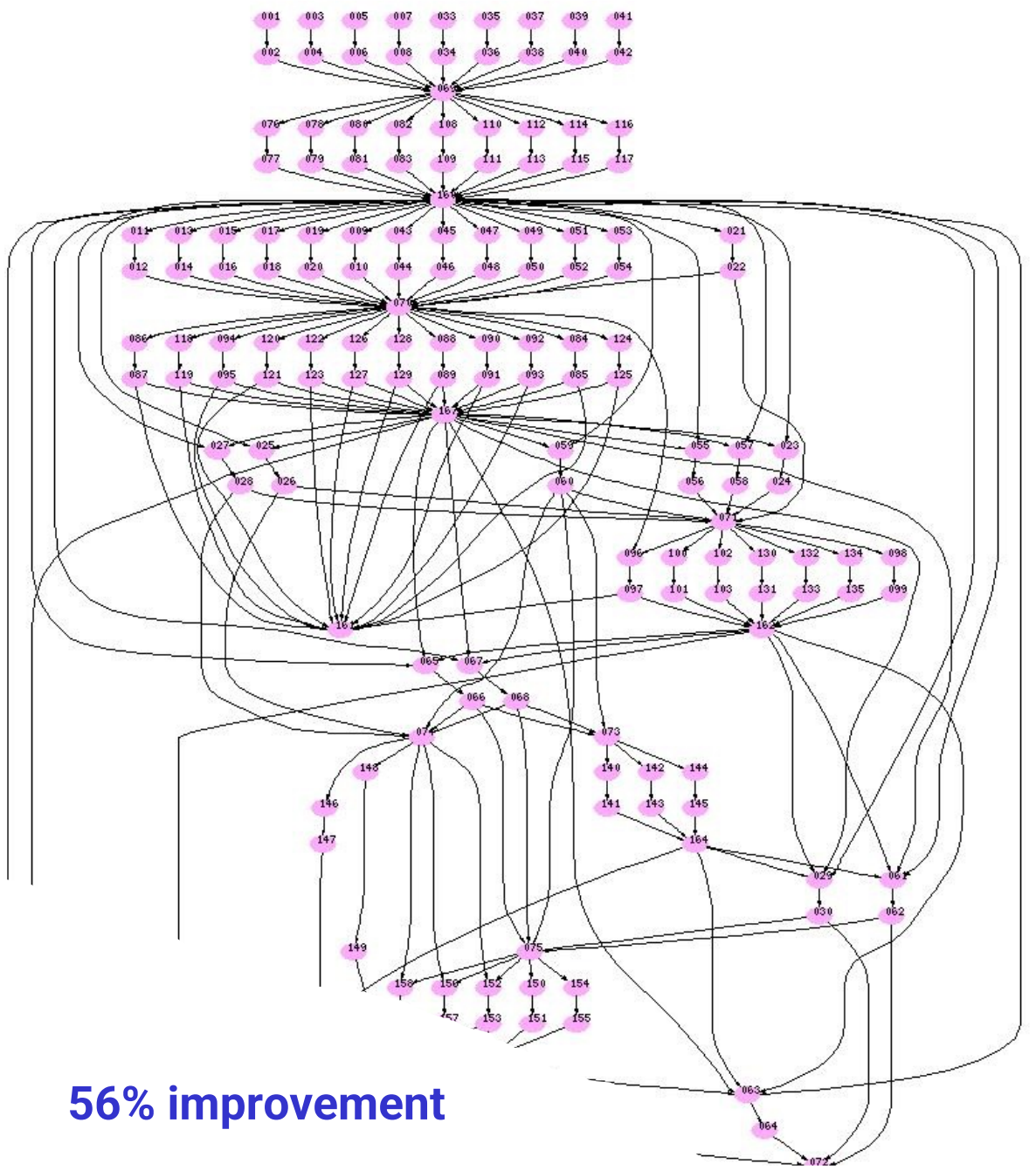
26% improvement



56% improvement

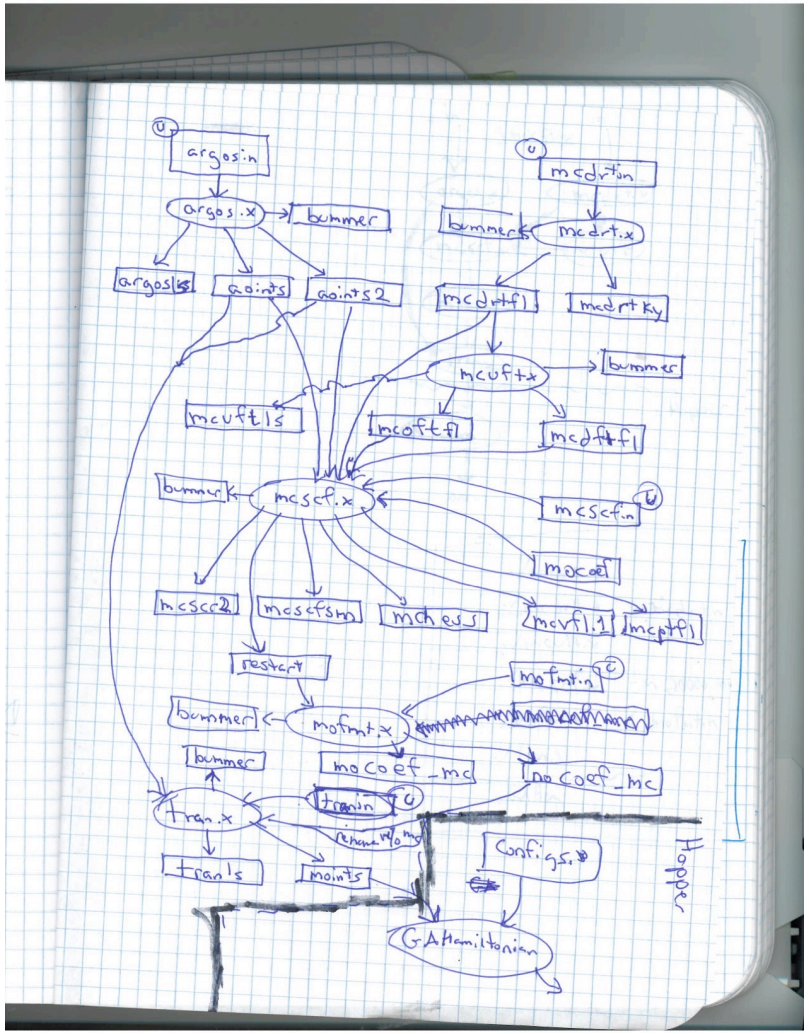


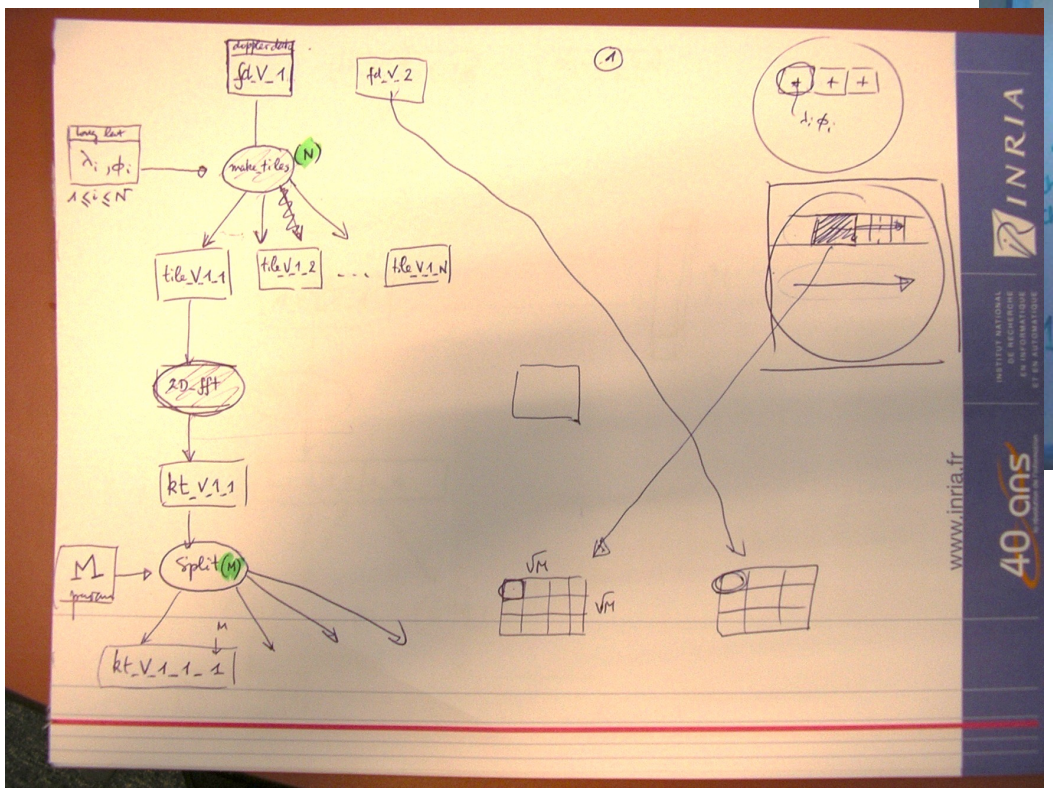
Is this a good thing?

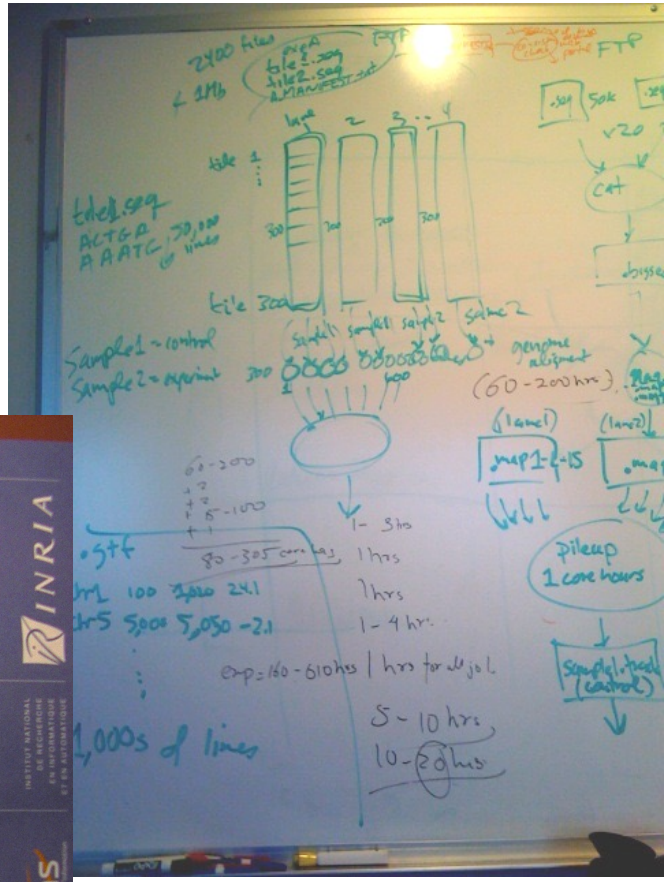


56% improvement

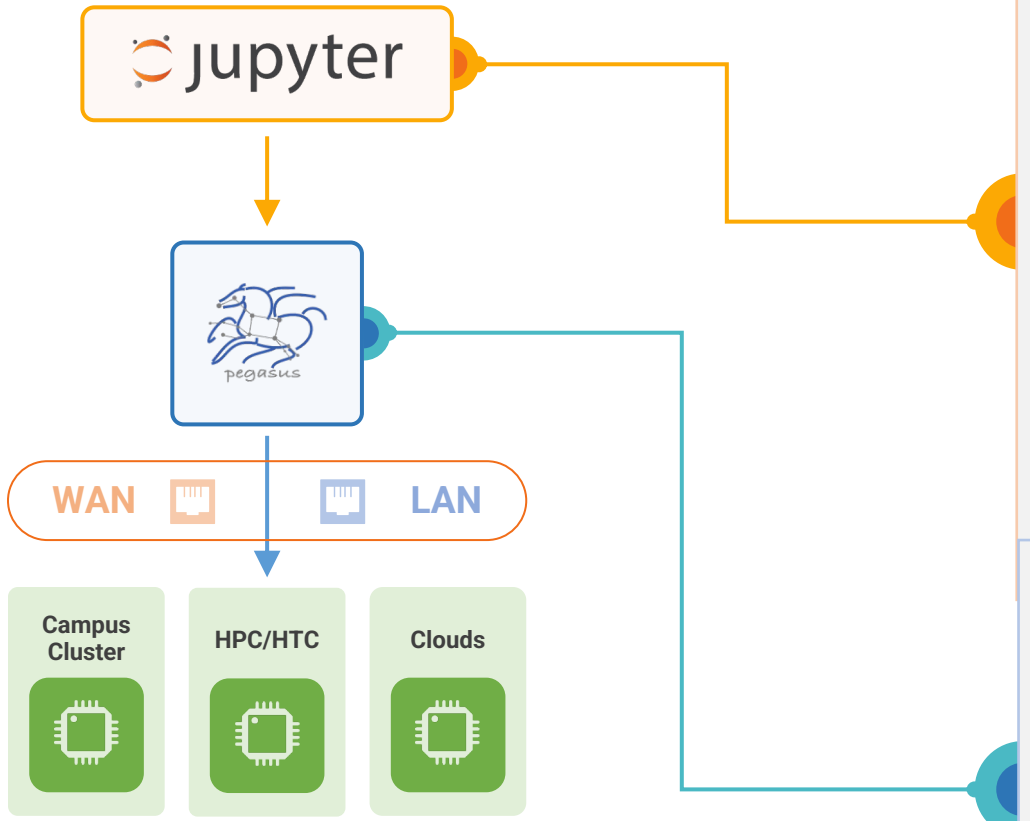
We cannot forget about the users

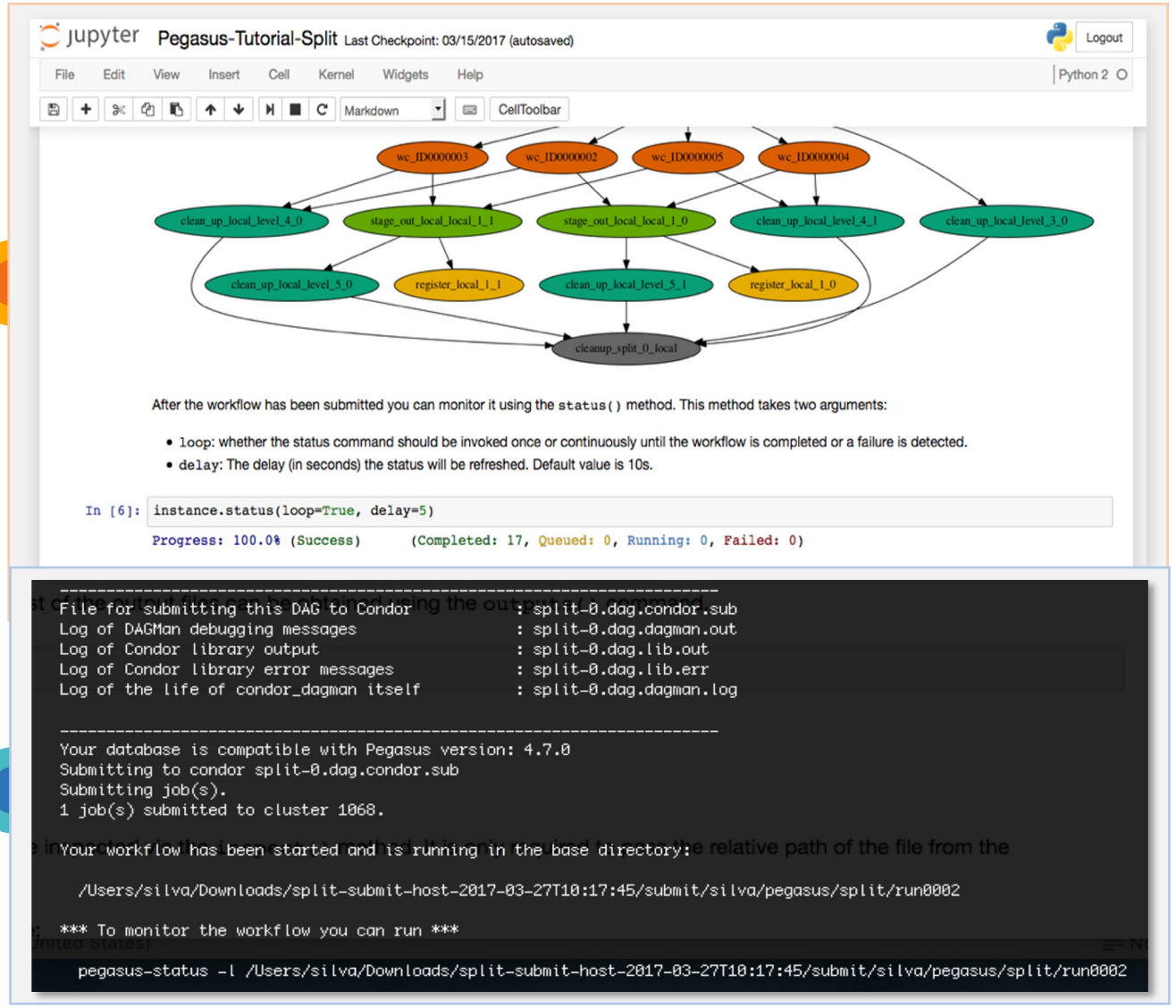






Running Pegasus workflows with Jupyter





Conclusion: Pegasus in a Heterogeneous World

1. Resource-independent specification
2. Submit locally, run globally
3. Flexible data staging configurations
4. Flexible data movement
5. Up and down integration with diverse CI, use of common languages

Pegasus mapping onto target resources
- Optimizing for performance
- Improving reliability
- Supporting scalability

- **Exploring ML techniques to improve workflow execution: performance prediction and anomaly detection**