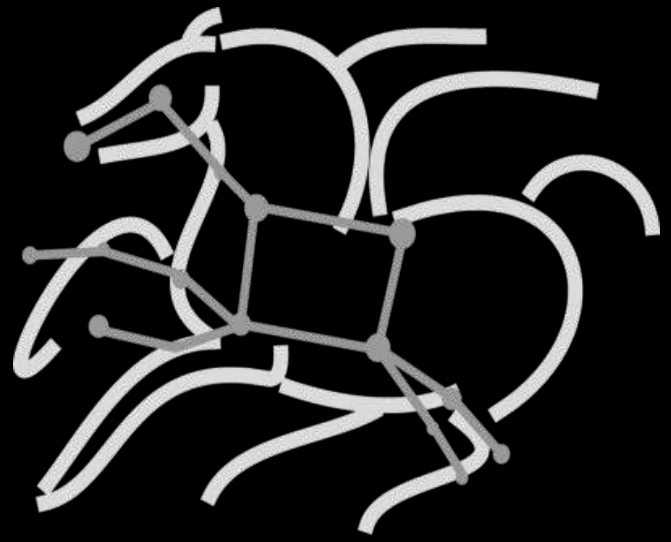
Pegasus Scientific Workflows on the Open Science Grid



https://pegasus.isi.edu

Automate, recover, and debug scientific computations.

Mats Rynge rynge@isi.edu

Outline

1. Quick introduction to Open Science Grid and OSG Connect
2. Use the “tutorial” command to get a sample workflow
3. Discuss site catalog and dax generator
4. Submit and monitor the workflow
5. Discuss the generated jobs
6. Containerized workflow

This is an interactive session. Please interrupt at anytime to ask questions.

Why Pegasus?

Automates complex, multi-stage processing pipelines

Enables parallel, distributed computations

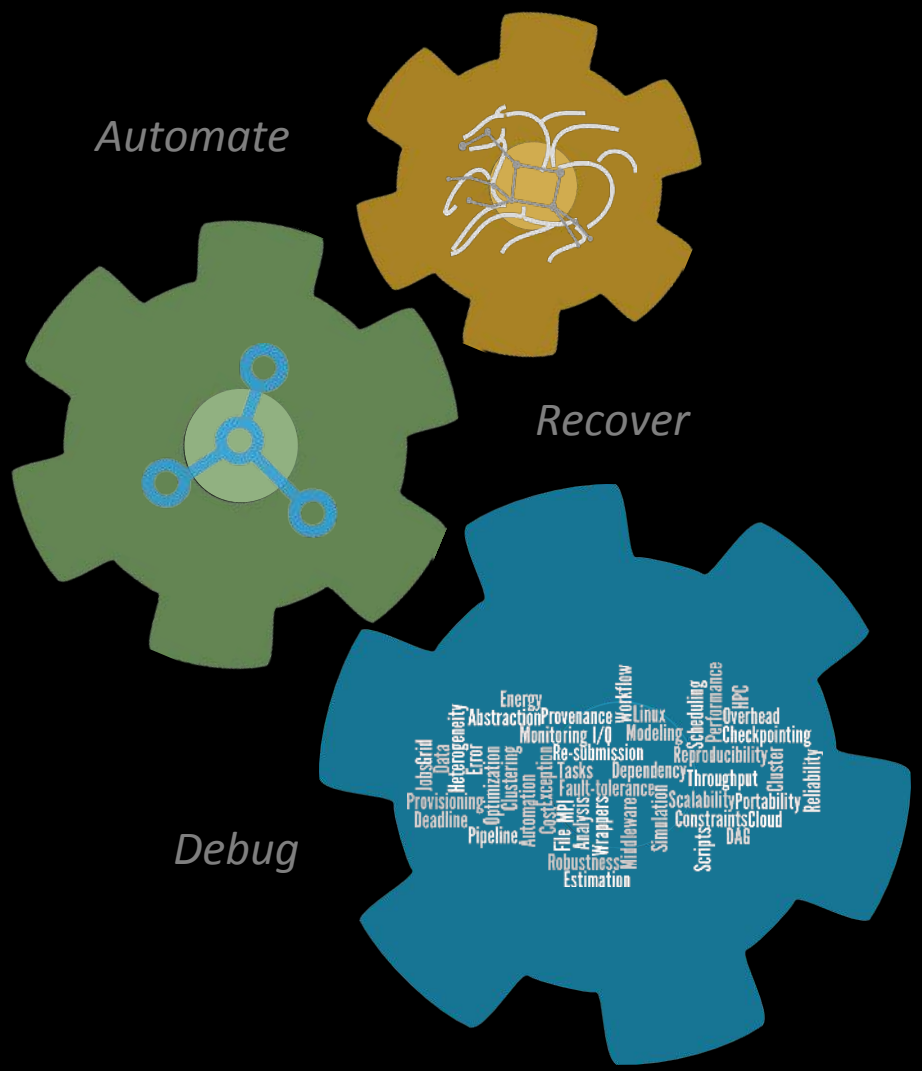
Automatically executes data transfers

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

Open Science Grid

A **framework** for large scale distributed resource sharing addressing the technology, policy, and social requirements of sharing computing resources.

OSG is a **consortium** of software, service and resource providers and researchers, from universities, national laboratories and computing centers across the U.S., who together build and operate the OSG project. The project is funded by the NSF and DOE, and provides staff for managing various aspects of the OSG.

Integrates computing and storage resources from over 120 sites in the U.S.

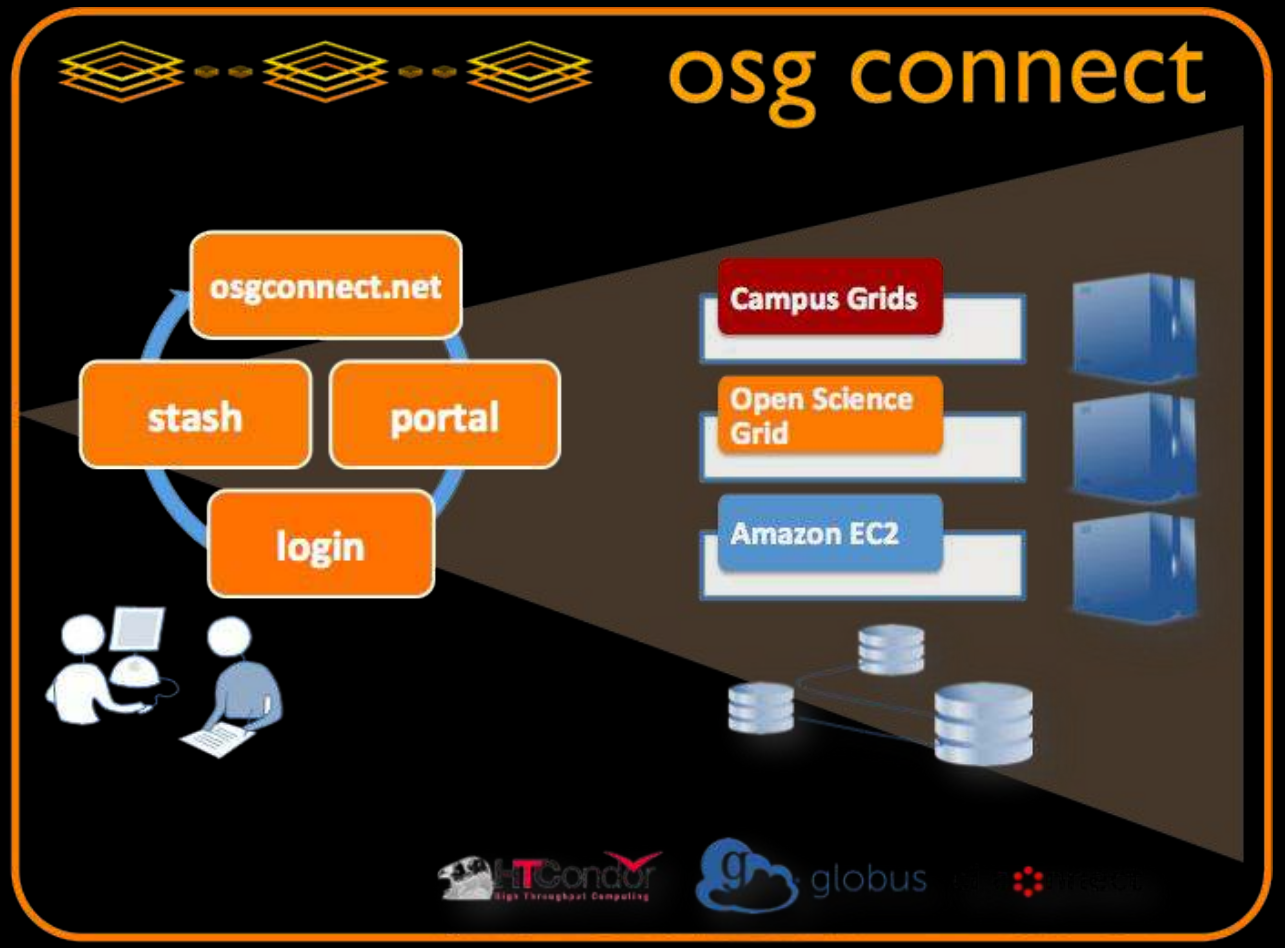
Open Science Grid

In the last 24 Hours	
278,000	Jobs
3,614,000	CPU Hours
13,859,000	Transfers
499	TB Transfers
In the last 30 Days	
8,586,000	Jobs
127,120,000	CPU Hours
0	Transfers
0	TB Transfers
In the last 12 Months	
121,418,000	Jobs
1,623,768,000	CPU Hours
681,704,000	Transfers
68,777	TB Transfers

OSG delivered across 118 sites

~ 3.5 million CPU hours delivered per day

OSG Connect Service



OSG Connect Provides:

- ★ Login host
- ★ Job scheduler
- ★ Software
- ★ Storage

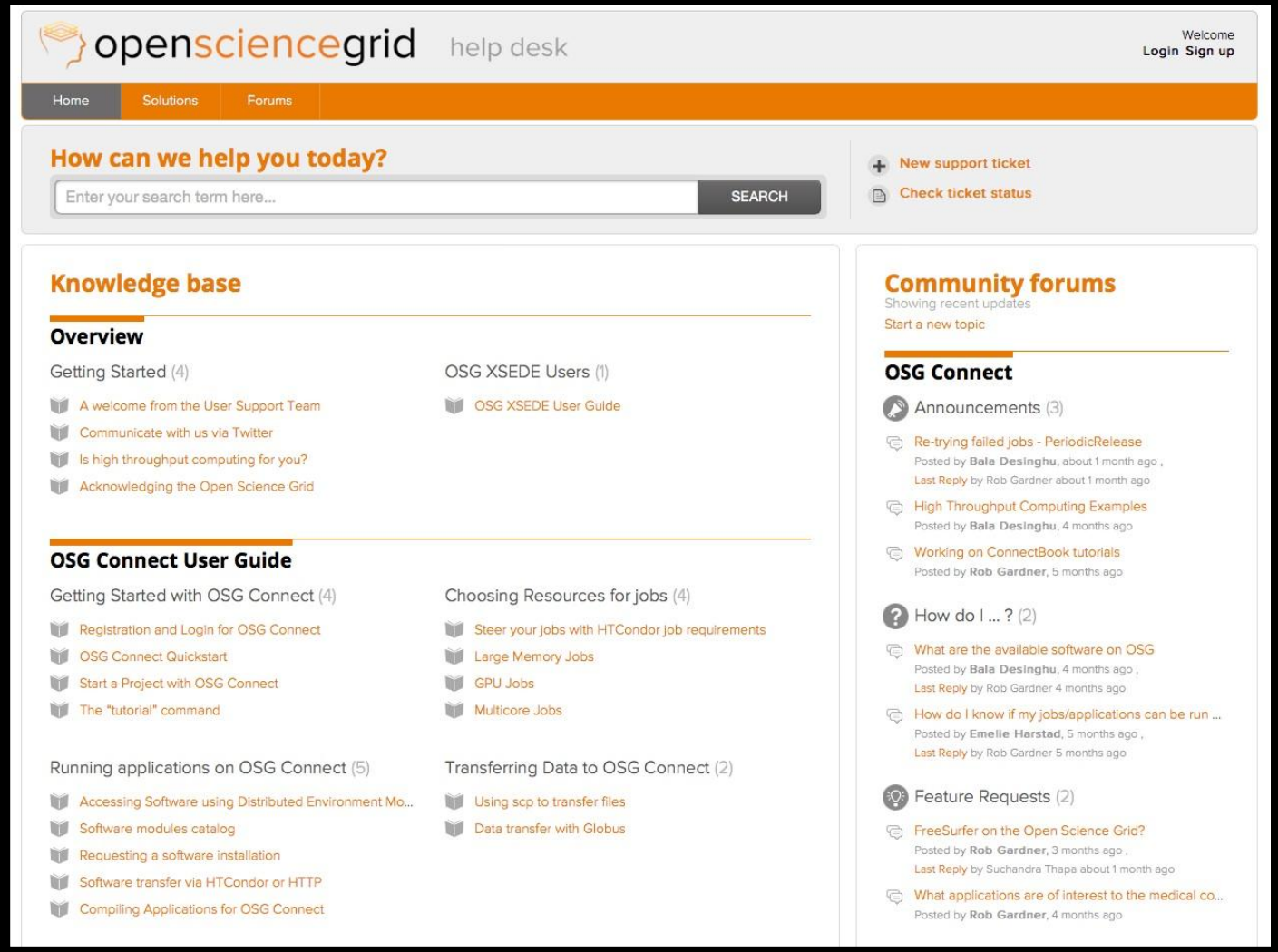
http://osgconnect.net/

OSG User Support

Helpdesk:
https://support.opensciencegrid.org

- Knowledge Base
  - User guides/tutorials
  - HTC Recipes
- Forums
- “How do I…?” articles
- Interactive online chat

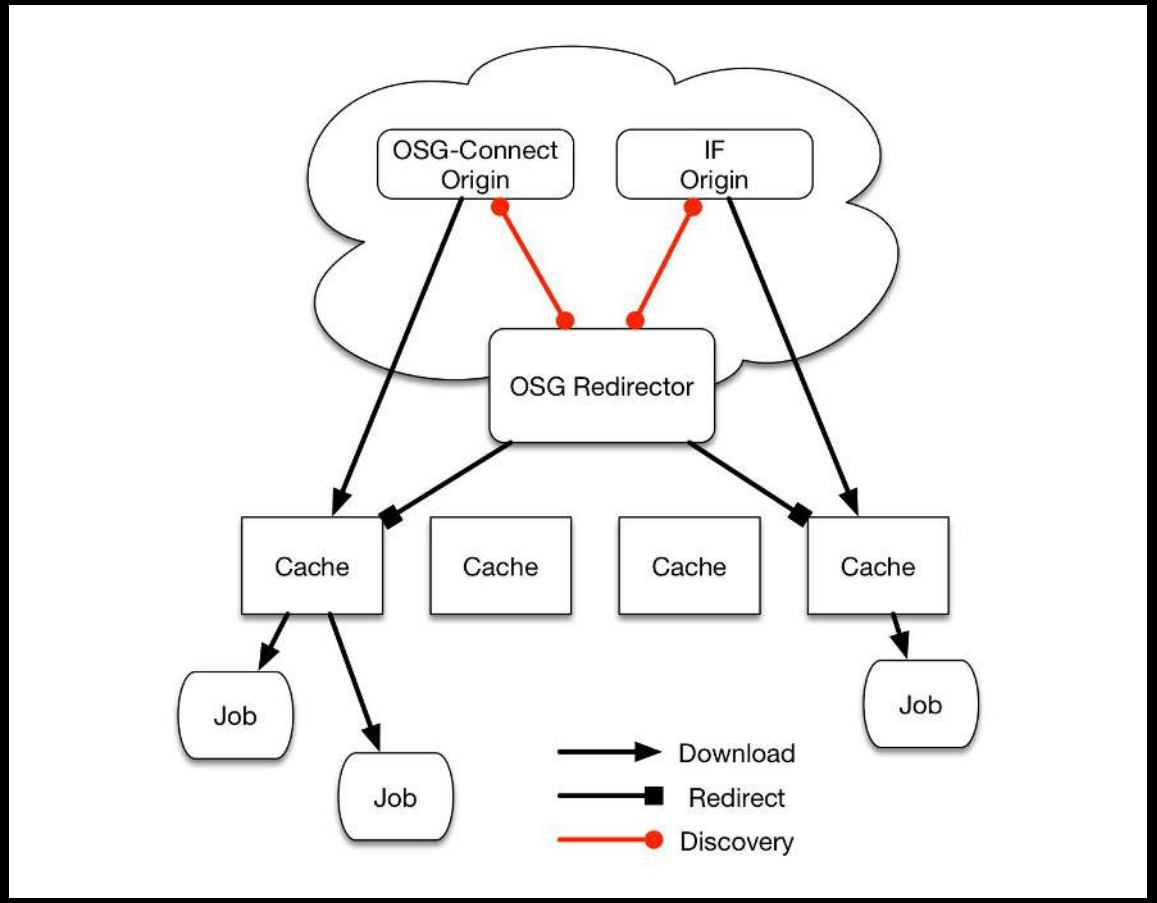
Support email:
user-support@opensciencegrid.org

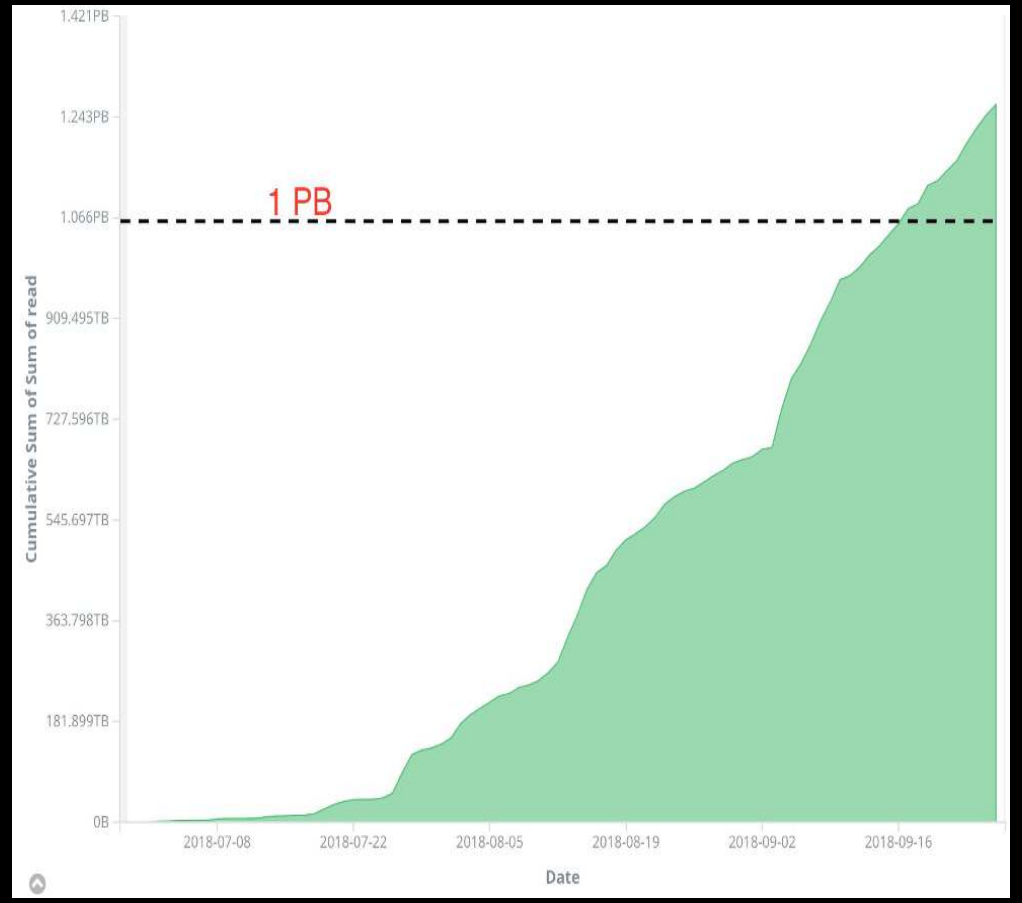


Storage service: “Stash”

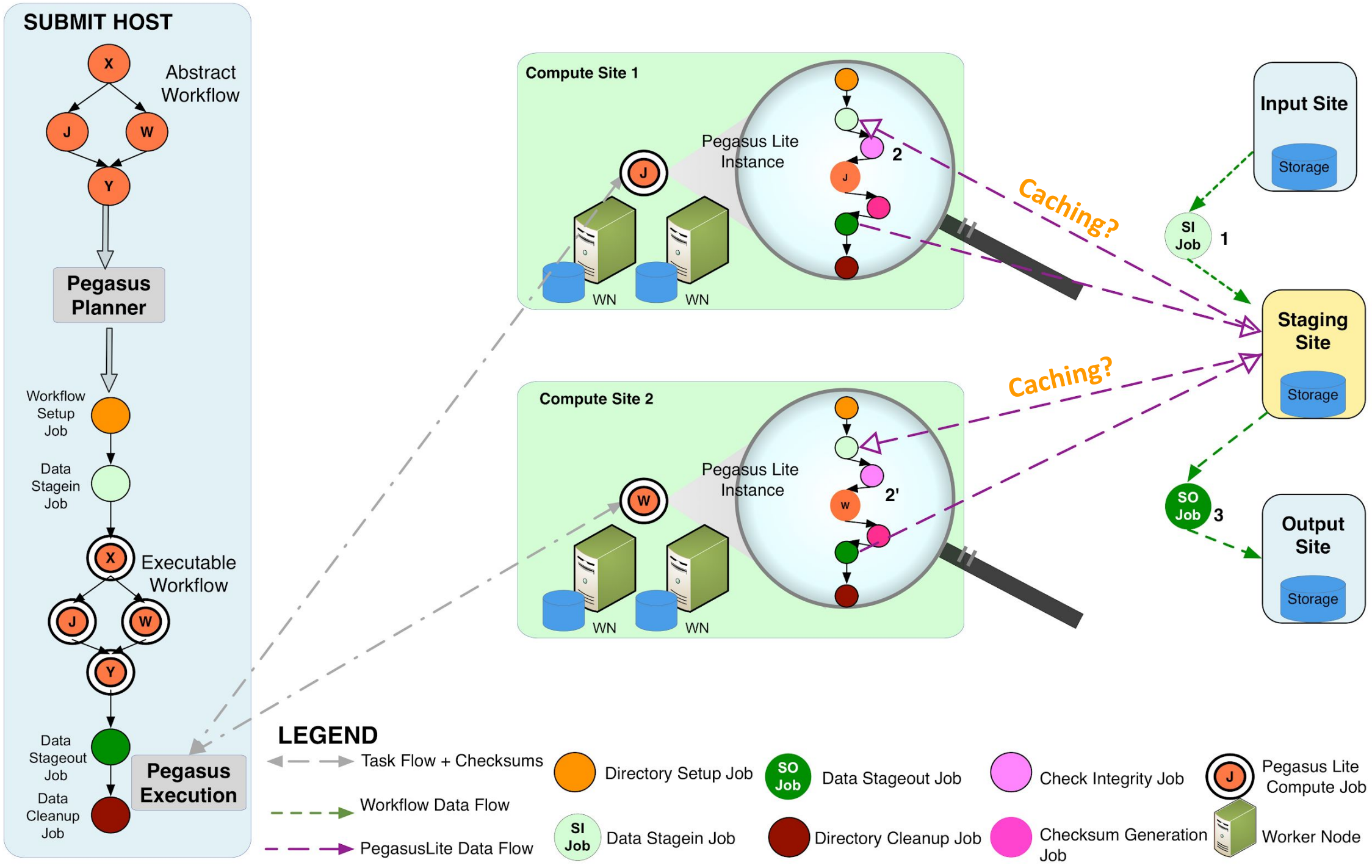
- Provide a quasi-transient storage service for job input/output data
- **POSIX** access provided to the login host
- **Globus Online** Server for managed transfers from campus data services
- Personalized **http** service endpoint
- Can now handle writes!
- Connected to 100 Gbps SciDMZ (I2, ESnet)

StashCache





https://derekweitzel.com/2018/09/26/stashcache-by-the-numbers/


SUBMIT HOST
Abstract Workflow
X
J
W
Y
Pegasus Planner
Workflow Setup Job
Data Stagein Job
Executable Workflow
Data Stageout Job
Data Cleanup Job
Pegasus Execution
Compute Site 1
Pegasus Lite Instance
WN
WN
Compute Site 2
Pegasus Lite Instance
WN
WN
Caching?
Caching?
Input Site
Storage
SI Job
1
Staging Site
Storage
SO Job
3
Output Site
Storage
Stash
LEGEND
Task Flow + Checksums
Workflow Data Flow
PegasusLite Data Flow
Directory Setup Job
Data Stagein Job
Data Stageout Job
Directory Cleanup Job
Check Integrity Job
Checksum Generation Job
Pegasus Lite Compute Job
Worker Node

Tutorial

https://support.opensciencegrid.org/support/solutions/articles/5000639789-pegasus

$ tutorial pegasus

$ cd tutorial-pegasus