Pegasus: Automating the Execution of Computational Workflows

Ewa Deelman, Karan Vahi, Gideon Juve, Mats Rynge, Rajiv Mayani, Rafael Ferreira da Silva – *USC Information Sciences Institute*
Miron Livny, Kent Wenger – *University of Wisconsin, Madison*
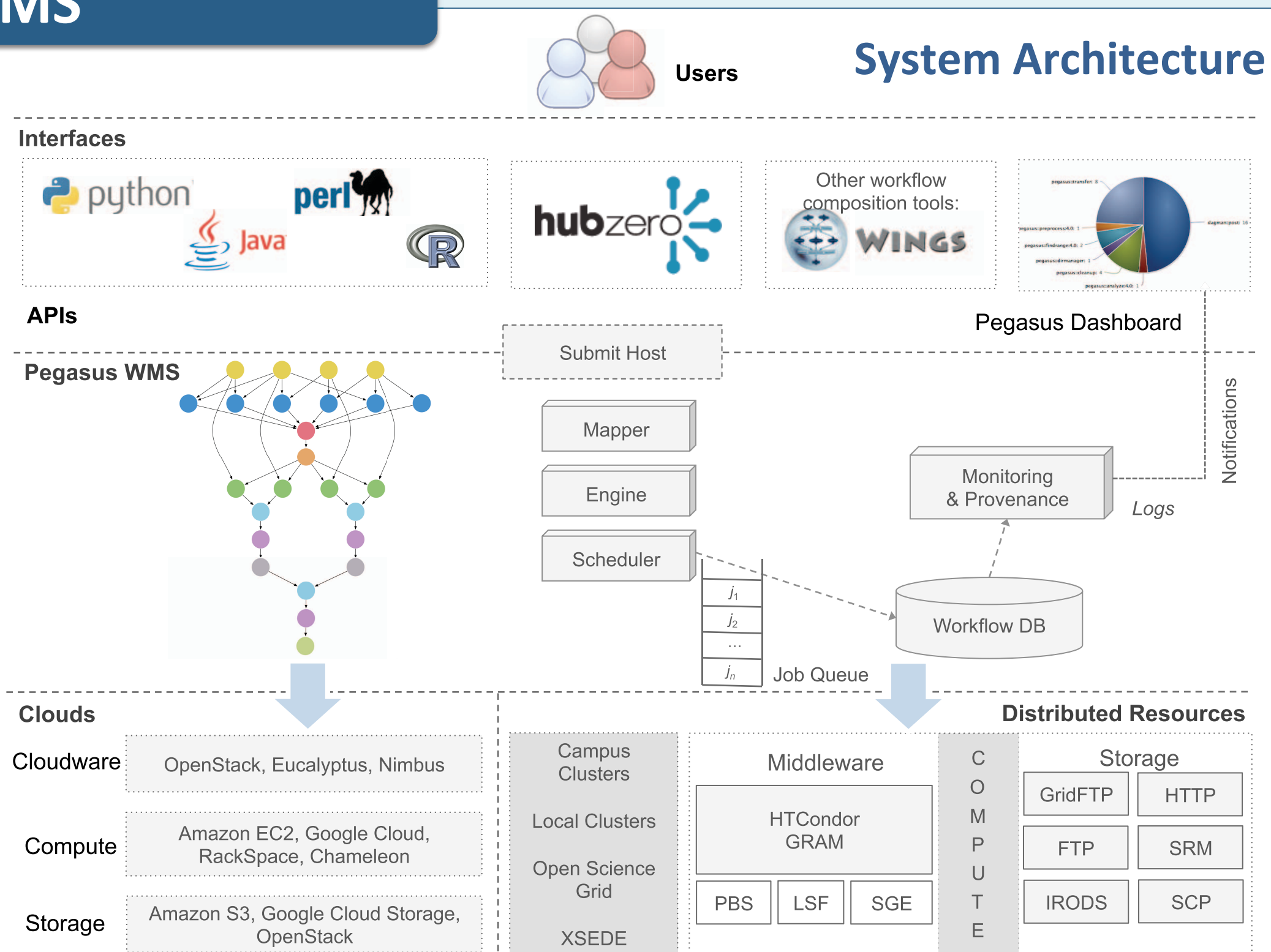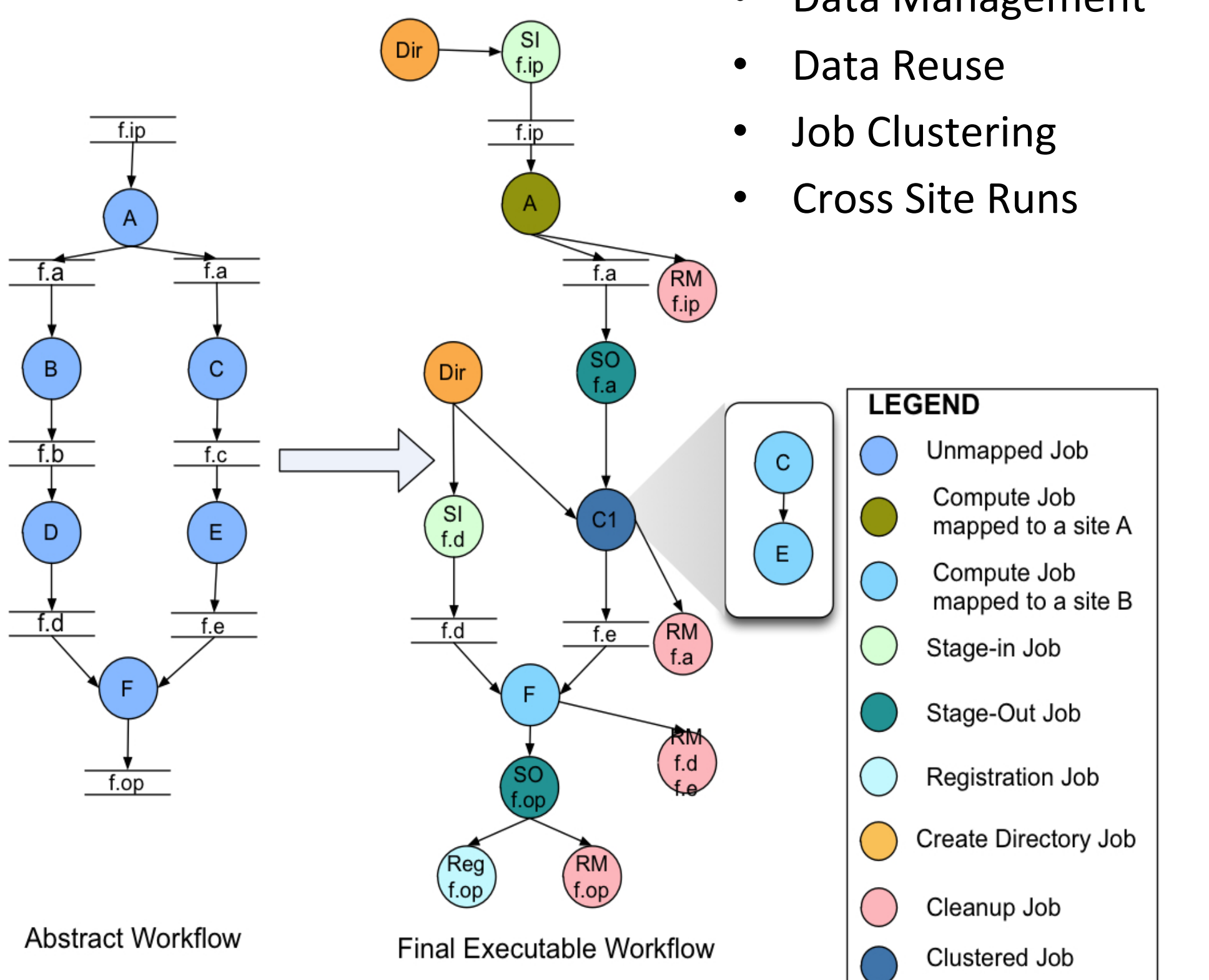
USC Viterbi School of Engineering

Pegasus WMS

- Pegasus is a system for mapping and executing abstract application workflows over a range of execution environments.
- The same abstract workflow can, at different times, be mapped different execution environments such as XSEDE, OSG, commercial and academic clouds, campus grids, and clusters.
- Pegasus can easily scale both the size of the workflow, and the resources that the workflow is distributed over. Pegasus runs workflows ranging from just a few computational tasks up to 1 million.
- Stores static and runtime metadata associated with workflow, files and tasks. Accessible via command line tools and web based dashboard.
- Pegasus-MPI-Cluster enables fine-grained task graphs to be executed efficiently on HPC resources.

System Architecture

Canonical Workflow Example

Abstract Workflow — Final Executable Workflow

Capabilities Highlighted

- Data Management
- Data Reuse
- Job Clustering
- Cross Site Runs

LEGEND
- Unmapped Job
- Compute Job mapped to a site A
- Compute Job mapped to a site B
- Stage-in Job
- Stage-Out Job
- Registration Job
- Create Directory Job
- Cleanup Job
- Clustered Job

Software Availability

Release Schedule

- Major Release every 9 months. Minor releases every 4 months
- Continuous Integration Testing with **Bamboo**
- Issue Tracking via **JIRA**

Download Options

- Source Code publicly hosted on GitHub
- Binary packages for Linux and MAC
- YUM/APT repositories with RPM/DEB packages
- Nightly Developments builds also available

Documentation / Training Materials

- Tutorials - Virtual Machine, EC2, and Docker images
- Support - Email lists and online chat rooms
- OSG Software Carpentry
- Online User Guide

Downloads & Usage Since 2013

- Workflows - 751,901
- Tasks   - 3,983,841,408
- Jobs   - 581,857,859

Astronomy and Physics

Pegasus powered LIGO analysis workflows to detect gravitational waves

LIGO
pyCBC

-0.02s

0.2 Second before the black holes collide. Image credit: SXS/LIGO

Periodogram workflows help detect extra solar planets

Galactic Plane workflow generates mosaics for astronomy surveys

Bioinformatics

Quality control workflows for data submissions to NRGR repository and PAGE consortium
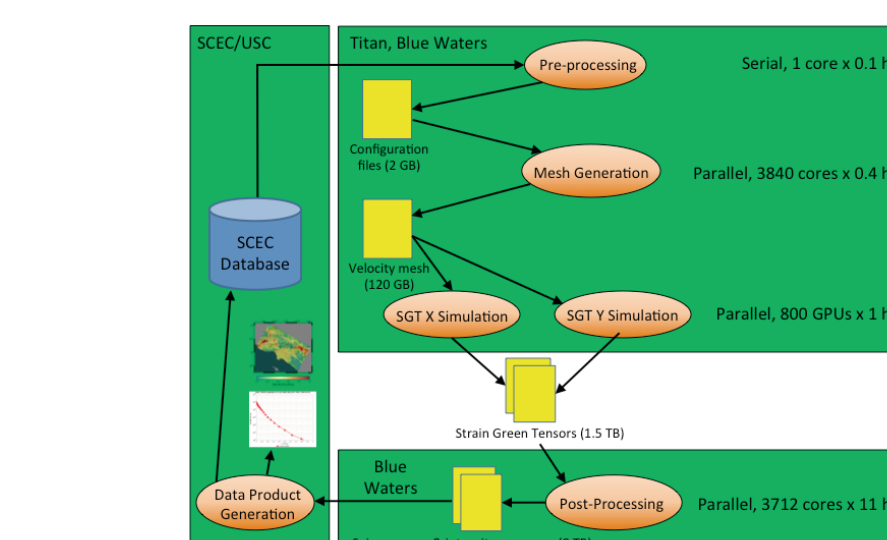
Imputation workflows on PAGE data

Workflows for Genome and Transcriptome free analysis of RSEQ

Brain span workflows help study gene expression in the brain

RNA Sequencing workflows for generating Cancer Genome Atlas

Soybean Knowledge base (SoyKB) workflow for re-sequencing soy-bean germplasm lines

Seismology

CyberShake workflows for seismic hazard analysis of LA basin

Broadband workflows for accurate predictions of ground motions

Applications Using Pegasus

http://pegasus.isi.edu/applications

Others

ACME Climate Modeling Workflows study interaction between climate change and societal energy requirements

Spallation Neutron Source Workflows study molecular dynamics and neutron scattering intensity calculations