Gearing the DECam Analysis Pipeline for Multi-Messenger Astronomy using Pegasus Workflows

Karan Vahi*, *Danny Goldstein[¶], George Papadimitriou*, Peter Nugent[$],*
*Ewa Deelman**

*USC Information Sciences Institute
[¶]California Institute of Tehcnology
[$]Lawrence Berkeley National Observatory

Poster: P10.39

# DECam Analysis Pipeline

Motivation

- Verify gravitational waves (GWs) discoveries with their electromagnetic (EM) optical counterparts in near real-time.

Pipeline

- Employs sophisticated machine learning algorithms to sift the data and identify events for scientists to follow up on.

- Processing begins by splitting the packed images into chips, applying crosstalk corrections, and performing standard bias/overscan subtraction and flat fielding.

- Subsequent operations are performed on a chip-by-chip basis in parallel.
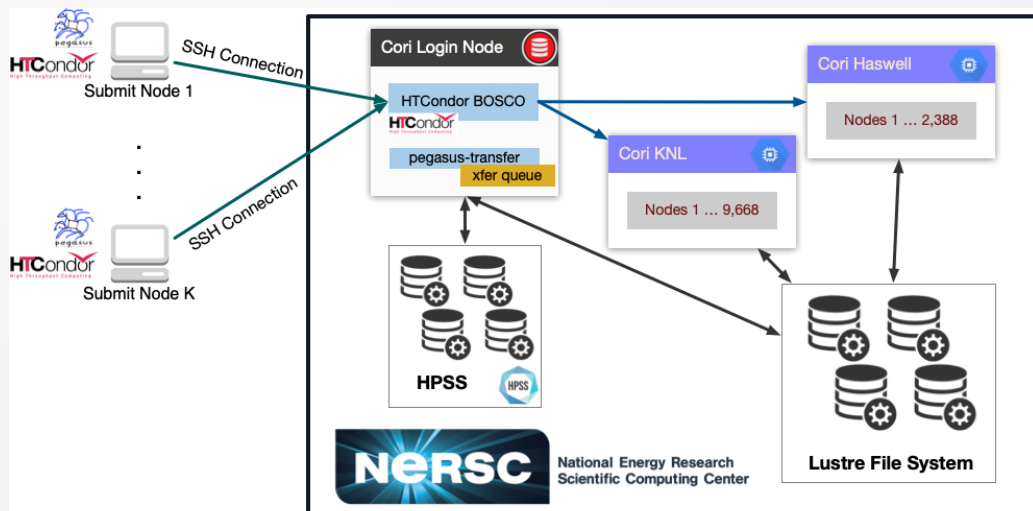
· **Input Files**
· **Intermediate Files**
· **Output Files**
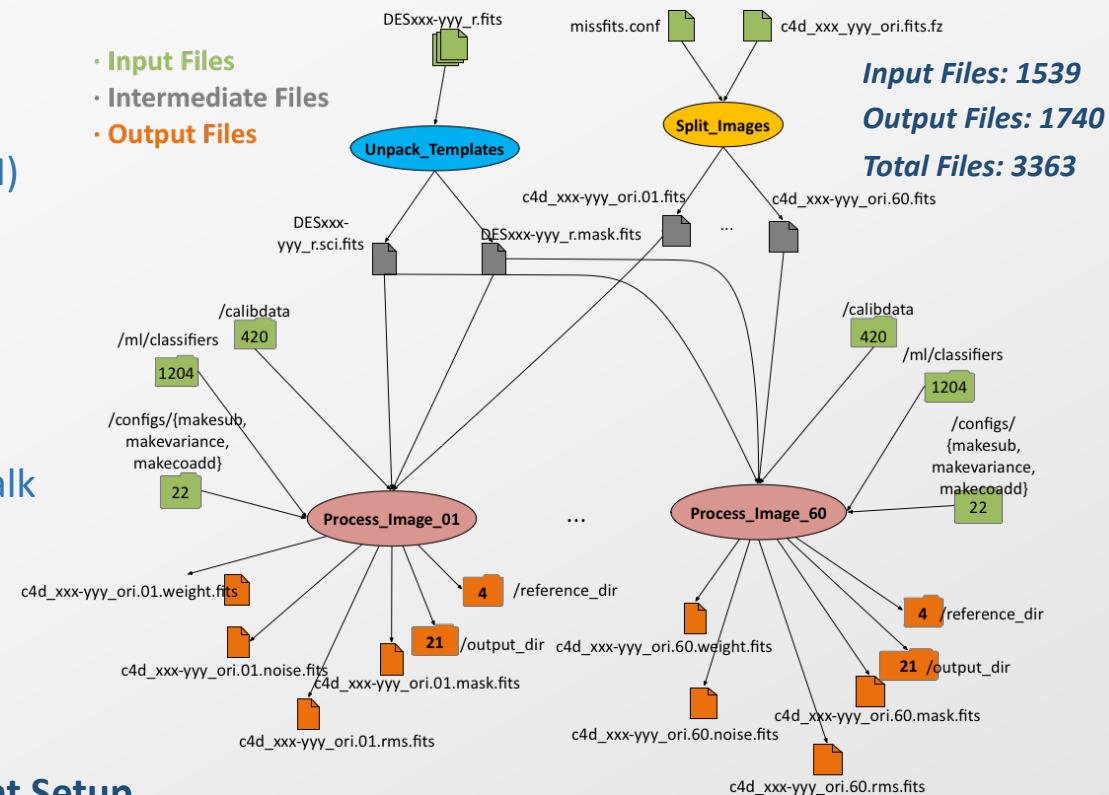
Input Files: 1539
Output Files: 1740
Total Files: 3363

Deployment Setup

- Workflows are submitted to CORI (CRAY XC40 Super Computing Cluster at NERSC).

- Each job executes a Shifter/Docker container containing the science codes.

- Historical data collected from the Victor Blanco telescope + DECam imager. Most of the dataset(~500 TB) archived at NERSC.

- Each LIGO trigger requires analysis of hundred of GB's of data

Pegasus

Poster: **P10.39**

ADASS XXIX

2

Pegasus Workflow Management System

Automates complex, multi-stage processing pipelines

Automate

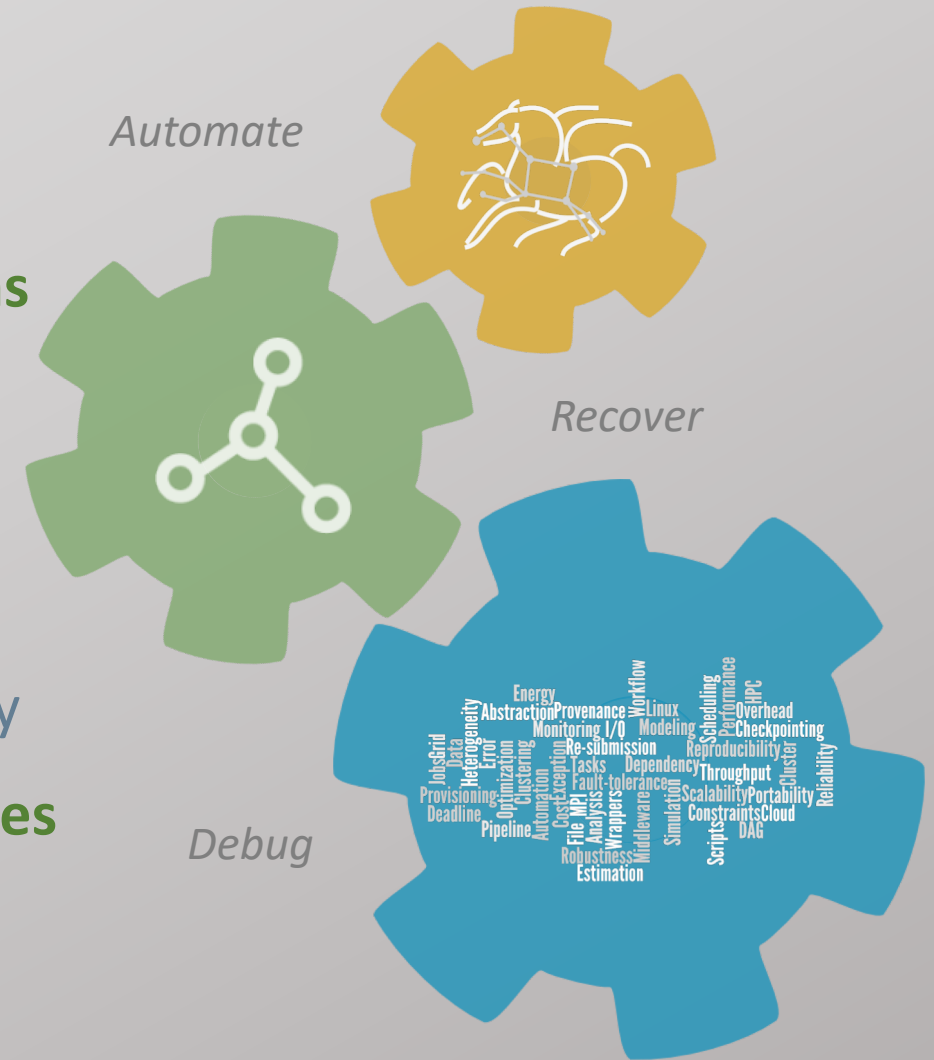Enables parallel, **distributed computations**

Automatically executes data transfers

Recover

Reusable, aids **reproducibility**

Records how data was produced (**provenance**)

Handles **failures** with to provide reliability

Keeps track of data and **files**

Debug

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

HTCondor
High Throughput Computing

Pegasus

- Users describe their pipelines in a portable format called Abstract Workflow, without worrying about low level execution details.

- Pegasus takes this and generates an executable workflow that
  - has data management tasks added
  - transforms the workflow for performance and reliability

Abstract workflow

logical filename (LFN)
platform independent (abstraction)

transformation

executables (or programs)
platform independent

stage-in job
Transfers the workflow input data

executable workflow

cleanup job
Removes unused data

stage-out job
Transfers the workflow output data

registration job