Producing a Multiwavelength Galactic Plane Atlas Using Montage, Pegasus and Amazon Web Services

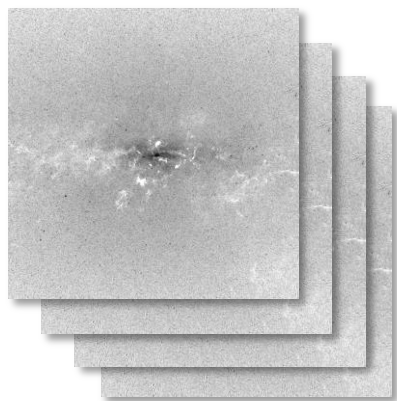Mats Rynge[1], Gideon Juve[1], Jamie Kinney[2], John Good[3], Bruce Berriman[3], Ann Merrihew[2], Ewa Deelman[1]

[1]USC Information Sciences Institute
[2]Amazon Web Services
[3]Infrared Processing and Analysis Center, Caltech

USC Viterbi
School of Engineering

Information Sciences Institute

Science Goal

- Multiwavelength image atlas of the Galactic Plane, with coverage of 360° along the galactic plane and ±20° on either side

- 16 different wavelengths from 1 µm to 24 µm

- Each output image is 5° by 5° in size, and have an overlap of 1° with neighboring tiles

- Processed so that they appear to have been measured with a single instrument observing all 16 wavelengths - Cartesian projection

- When complete, the data will be released to the community via an API

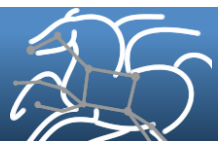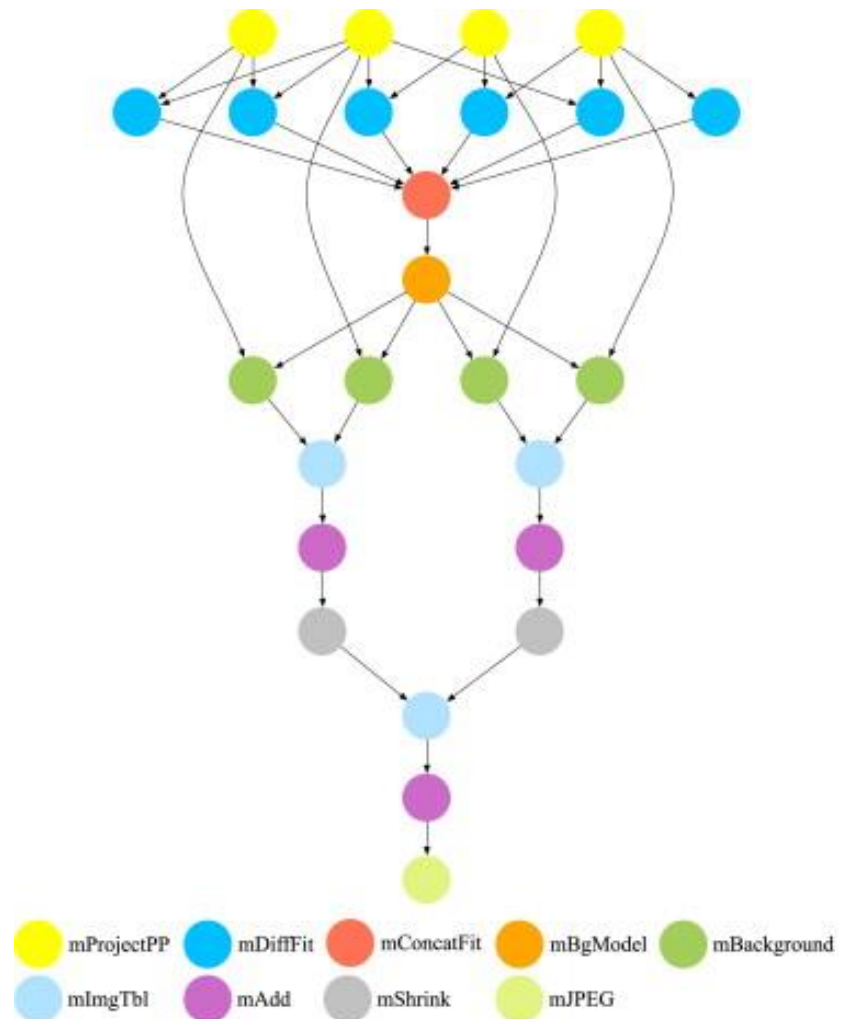| Survey / Bands (µm)	Coverage of 360°x40° area	Output Size (TB)	Compute time (1,000s core hours)
2MASS (1.2, 1.6, 2.2)	100%	14.4	87
GLIMPSE (3.6, 4.5, 5.8, 8.0)	11%	2.0	60
MIPSGAL (24)	8%	0.4	3
MSX (8.8, 12.1, 14.6, 21.3)	35%	6.8	36
WISE (3.4, 4.6, 12, 22)	100%	19.2	132

The Montage Image Mosaic Engine

- Toolkit written in ANSI-C for creating and managing image mosaics in FITS format.

- Portable and scalable – runs on desktops, grids and cloud computing platforms under *nix platforms.

- Code available through clickwrap license at Caltech.

- Widely adopted by astronomy and IT communities: used on desktops, integrated into processing pipelines, used in development cyber-infrastructure

Pegasus Workflow Management System

- **Builds on top of HTCondor and 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 planner/mapper ("compiler")**
  - **Transforms the workflow for performance and reliability**
  - **Automatically locates physical locations for both workflow components and data**
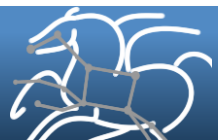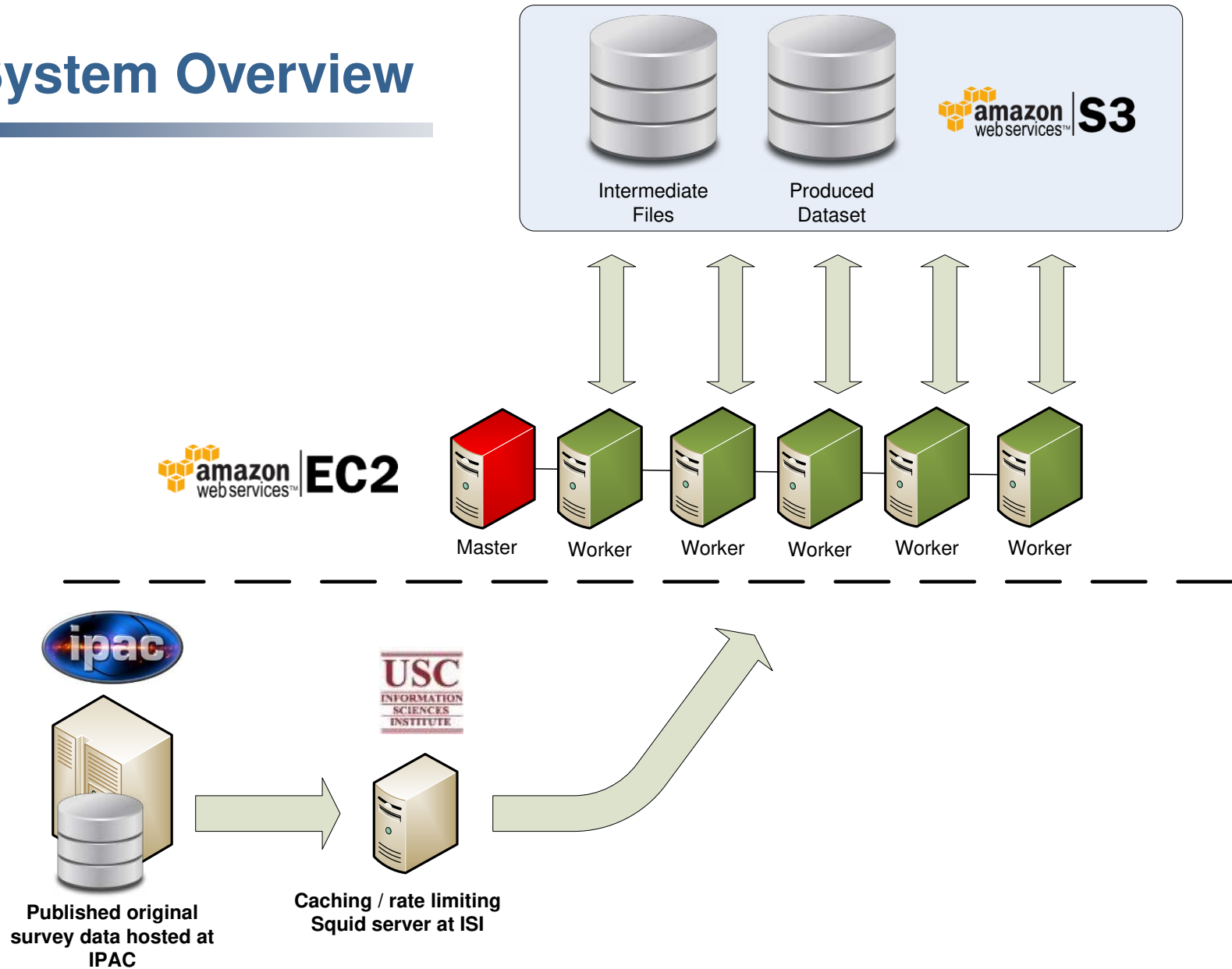 - **Collects runtime provenance**

Legend: mProjectPP, mDiffFit, mConcatFit, mBgModel, mBackground, mImgTbl, mAdd, mShrink, mJPEG

USC Viterbi
School of Engineering

Galactic Plane Workflow

16 hierarchal workflows
Each one with 1,001 subworkflows
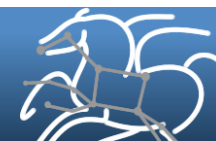**Over 10M input files**
45 TB output dataset
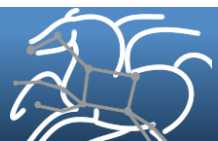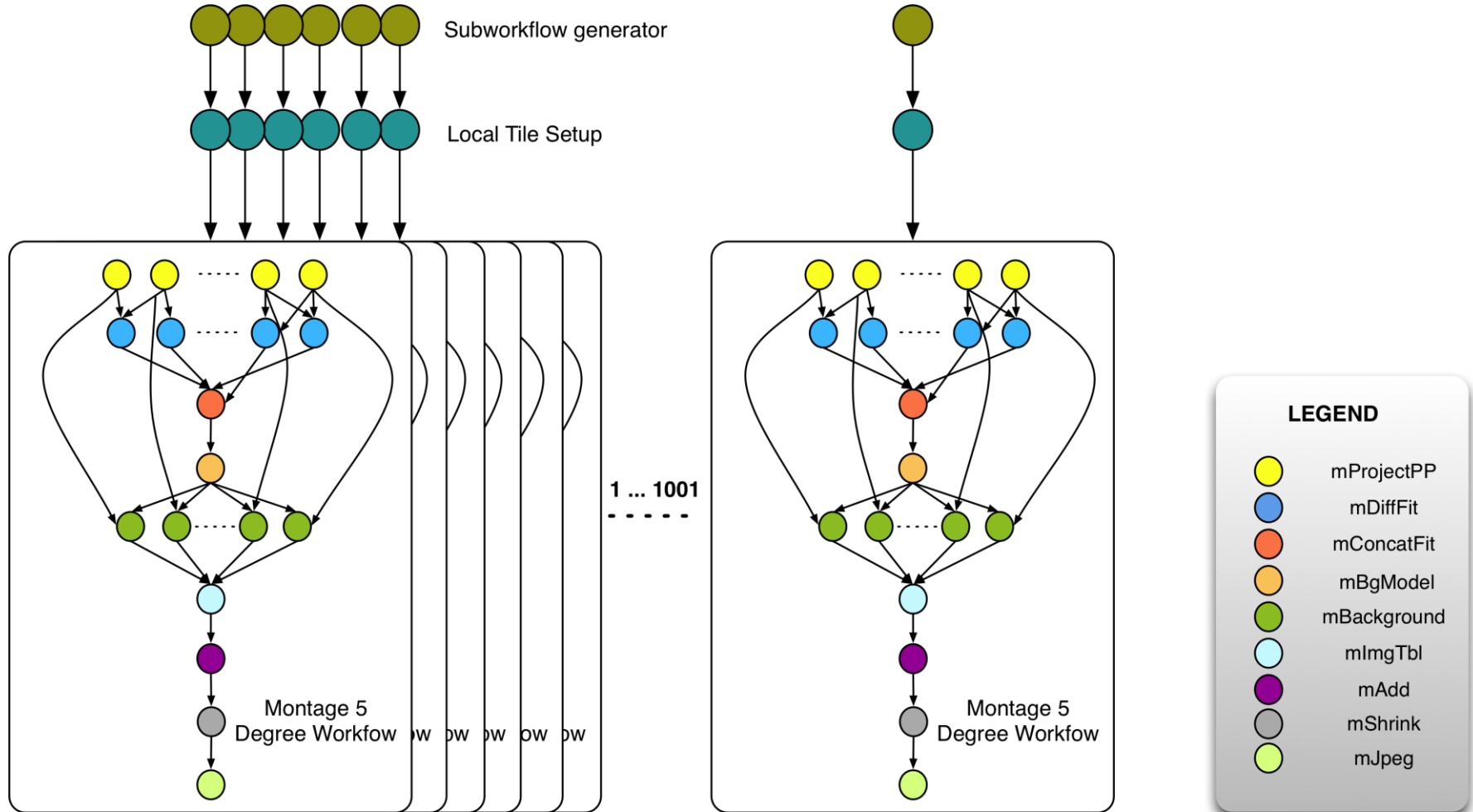
System Overview

Numbers

- **Amazon Web Services contributed the computations and storage**

- **hi1.4xlarge instance (the one we used)**
  - **Memory optimized, with 2 x SSD ephemeral drives**
  - **318,000 core hours**
  - **Spot instance price: $5,950**

- **cc2.8xlarge instance (benchmarked)**
  - **Compute cluster optimized, with 4 ephemeral drives (2 used)**
  - **274,000 core hours**
  - **Spot instance price: $2,200**

Survey / Bands (μm)	Coverage of 360°x40° area	Output Size (TB)	Compute time (1,000s core hours)
2MASS (1.2, 1.6, 2.2)	100%	14.4	87
GLIMPSE (3.6, 4.5, 5.8, 8.0)	11%	2.0	60
MIPSGAL (24)	8%	0.4	3
MSX (8.8, 12.1, 14.6, 21.3)	35%	6.8	36
WISE (3.4, 4.6, 12, 22)	100%	19.2	132

Questions?

rynge@isi.edu

- **Pegasus - http://pegasus.isi.edu/**
  - NSF funded
  - Open Source
  - Documentation, tutorial, and support available on website

- **Montage - http://montage.ipac.caltech.edu/**

- **Amazon Web Services - http://aws.amazon.com/**