*Astronomical Data Analysis Software and Systems XVI* **P4-20**
*ASP Conference Series, Vol. XXX, 2006*
*R. Shaw, F. Hill and D. Bell, eds.*

# The NOAO High-Performance Pipeline System: The Mosaic Camera Pipeline

Robert A. Swaters

*Department of Astronomy, University of Maryland, College Park, MD 20742*

Francisco Valdes

*National Optical Astronomy Observatory, 950 N. Cherry Avenue, Tucson, AZ 85719*

**Abstract.** The NOAO Mosaic Camera Pipeline produces instrumentally calibrated data products and data quality measurements from all exposures taken with the NOAO Mosaic Imagers at the KPNO and CTIO telescopes. We describe the distributed nature of the Mosaic Pipeline, the calibration data that are applied, the data quality metadata that are derived, the data products that are delivered by the Mosaic Pipeline, and the performance of the system.

## 1. Introduction

The NOAO Mosaic Camera Pipeline is designed to produce calibrated data products for all exposures takes with the NOAO Mosaic Imagers at the KPNO and CTIO telescopes. These imagers consist of 8 $2048 \times 4096$ CCD detectors arranged in a $4 \times 2$ pattern in the focal plane to provide a field of view of approximately $0.5^\circ \times 0.5^\circ$ (at the CTIO and KPNO 4 meter telescopes) or $1^\circ \times 1^\circ$ (at the KPNO 0.9m telescope).

The NOAO Mosaic Camera Pipeline is a data reduction pipeline designed to run on a cluster of computers. This pipeline system allows easy and efficient use of CPU resources in problems with inherent parallelism such as the processing of observations from mosaic cameras. The system makes use of the NOAO High-Performance Pipeline System (NHPPS) described in a companion paper (see Cline et al. 2007).

## 2. A Distributed Pipeline System

In the NHPPS, the data reduction process has been organized into a hierarchical structure of different individual pipelines. Each of these pipelines deals with one aspect of the reduction process. Some pipelines construct calibration data (such as bias, dome flats, pupil ghost and fringe correction, and dark sky flats), and others apply these calibration data. Some pipelines collect the final data products, and others control and organize the flow of data. All of these pipelines, in turn, are subdivided into a set of modules, each of which carries

out a small step of the data reduction process. These modules can be written in any language, although, because the pipeline system uses IRAF, most of them are IRAF scripts.

All of the pipelines and associated modules can be configured to run on all available nodes, or any subset thereof. This ensures maximum use of the available resources. Whenever a pipeline is to be started, a node selection algorithm is called, which first determines the nodes that are available to run this particular pipeline, and from those determines the best node by considering issues such as load and minimizing network traffic.

The distribution of the work across the nodes in the cluster also depends on the nature of the steps in the data reduction process. Given that the MOSAIC camera data consists of multi-extension FITS (MEF) files, each comprising data from the 8 CCDs, it may be more efficient, in some cases, to process the data by distributing the individual CCD FITS files across the cluster nodes, whereas in other cases it will be more efficient to distribute the complete MEF files. An example of distribution by CCD is the calculation of the fringe templates or master sky flats from a group of observations. In this case, the data from the 8 different CCDs in all of the MEF files are sent to 8 different nodes, each of which will process the data from one unique CCD. When applying calibration data to MEF files, however, it is more efficient to leave the individual CCDs on the same node, because this significantly reduces transfer of data between nodes. In this case, the data are distributed across nodes by MEF files.

To make sure available resources are used efficiently, any node can run multiple instances of a pipeline. Thus, on a node with two CPUs, two instances of a pipeline that apply calibration data to science observations will run in parallel.

## 3. Calibration

The NOAO Mosaic Camera Pipeline applies all major calibrations to the raw data:

- Removal of the signature of electronic cross-talk between individual amplifiers in the array
- Individual bias, dome flat, and twilight flat observations are checked (see also Sections 4 and 5) and combined
- The average bias and domeflat (or twilight flat) are applied
- Saturated pixels and bleed trails are flagged
- For object exposures, the world coordinate system and a rough photometric zeropoint are determined by matching objects in the field against the USNO-B catalog

The pupil ghost, fringe, and dark sky flat calibration data are derived by combining and filtering large groups of exposures. This optimal grouping is determined as part of the pipeline. The default is to group exposures by night are preferred, but groups spanning shorter or longer time intervals are used as needed, depending on the number of available exposures. Not all exposures are suitable to determine the pupil ghost, fringe and dark sky calibration data. How suitable exposures are selected is described in detail in a companion paper (see Valdes & Swaters 2007).

- The pupil ghost is caused by reflections off the secondary mirror. A pupil ghost template is constructed by combining suitable science exposures with the objects masked out. This template is then scaled for each exposure and subtracted
- The fringe template is also created by combining suitable science exposures with objects masked out, but large scale structures are also removed by subtracting a median filtered version of the combined images. The resulting template is then scaled for each exposure and subtracted
- The construction of a master sky flat to correct for differences in illumination of the detector and differences in color between the dome flats. It is created by combining suitable science exposures with objects masked out and divided into the data

All calibration images that are created by the pipeline are stored in the pipeline calibration database, along with a date range over which these calibration data are useful. These calibration data can be retrieved by the pipeline if it is not possible to construct the necessary calibration data from a given dataset. Calibration data from this database can be used, for example, if a dome flat for a particular filter is not available, or if there are too few exposures to construct e.g., a fringe template or a dark sky flat.

Because construction of the pupil ghost template, fringe template, and dark sky flat are CPU intensive, they are only carried out as part of the normal pipeline operations. In a "quick reduce" environment (e.g., at the telescope), these calibrations are not calculated, but instead the existing calibration data from the pipeline's calibration database are applied.

## 4. Data Quality

The NOAO Mosaic Camera Pipeline verifies, checks, and characterizes the data being processed throughout the reduction process, and the resulting metadata is stored in a database. This is being done for several reasons:

- Immediately after ingesting the data, the fits files and headers are verified and data that cannot be processed are rejected. Examples are: corrupt or incomplete fits files, or fits files with missing critical header keywords. Whenever possible, missing keywords are reconstructed.
- Statistics and other numerical characterizations of individual calibration exposures are compared against expected values and outliers are rejected.
- Calibration products created by the pipeline are characterized
- Science exposures are characterized; the metadata includes e.g., seeing, photometric depth, sky brightness, and WCS accuracy

The metadata derived from the raw data, intermediate steps, and pipeline data products are all stored in the pipeline metadata database. This database can be queried by the pipeline itself to carry information from one module to the next, but it also provides a means to investigate long term trends in the instrument's performance.

## 5. Rules

During the reduction process the pipeline needs to make decisions on the fly. In the NOAO Mosaic Camera Pipeline, these decisions have been separated from the code in the modules. Configuration files control syntactically simple decision (e.g., boolean variables, or comparing variables against a threshold). These files can be used to configure the pipeline in a global sense, and also for specific datasets. Rules can be more complex in nature, and can consist of complex scripts that take several variables as input. An example is the evaluation of dome flats, in which the statistics are evaluated differently depending on the filter.

## 6. Data products

After the data have been fully calibrated, the data are reprojected to the same orientation and pixel size, and to the closest tangent point on a predefined grid (thus ensuring that spatially close exposures have the same tangent point). Both the original reduced image and the resampled one are end products of the pipeline. In addition, the pipeline produces masks for both the original and resampled data. The pipeline also produces 2x2 block averaged versions of the data in PNG format, and thumbnail versions.

## 7. Performance

The NOAO Mosaic Camera Pipeline can fully reduce and create all data products at a rate of on average 35 Mosaic images per hour on a cluster of 8 nodes (each with dual 3.0 GHz Xeons), and also 1 node (with a dual Xeon 2.8 GHz) which hosts the calibration and metadata database. The processing rate depends on the nature of the observations and ranges from 30 exposures per hour to as high as 50 images per hour. The average of 35 Mosaic images per hour corresponds to approximately 5.0 GB per hour, or 0.12 TB per day.

## References

Cline, R. et al., 2007, this volume, [P4.19]
Valdes, F., & Swaters, 2007, this volume, [P4.21]