Astronomical Data Analysis Software and Systems XII
ASP Conference Series, Vol. 295, 2003
H. E. Payne, R. I. Jedrzejewski, and R. N. Hook, eds.

CARMA Data Storage, Archiving, Pipeline Processing, and the Quest for a Data Format

Raymond Plante[1], Marc W. Pound[2], David M. Mehringer[1], Stephen L. Scott[3], Andy Beard[3], Paul Daniel[3], Rick Hobbs[3], J. Colby Kraybill[4], Melvyn Wright[4], Erik Leitch[5], N. S. Amarnath[2], Kevin P. Rauch[2], Peter J. Teuben[2]

Abstract. In 2005, the BIMA and OVRO mm-wave interferometers will be merged into a new array, the Combined Array for Research in Millimeter-wave Astronomy (CARMA). Each existing array has its own visibility data format, storage facility, and tradition of data analysis software. The choice for CARMA was to use one of a number of existing formats or devise a format that combined the best of each. Furthermore, it had to address three important considerations. First, the CARMA data format must satisfy the sometimes orthogonal needs of both astronomers and engineers. Second, forcing all users to adopt a single off-line reduction package is not practical; thus, multiple end-user formats are necessary. Finally, CARMA is on a strict schedule to first light; thus, any solution must meet the restrictions of an accelerated software development cycle and take advantage of code reuse as much as possible. We describe our solution in which the pipelined data passes through two forms: a low-level database-based format oriented toward engineers and a high-level dataset-based form oriented toward scientists.

The BIMA Data Archive at NCSA has been operating in production mode for a decade and will be reused for CARMA with enhanced search capabilities. The integrated BIMA Image Pipeline developed at NCSA will be used to produced calibrated visibility data and images for end-users. We describe the data flow from the CARMA telescope correlator to delivery to astronomers over the web and show current examples of pipeline-processed images of BIMA observations.

[1]NCSA/University of Illinois

[2]University of Maryland

[3]Caltech/OVRO

[4]University of California, Berkeley

[5]University of Chicago

1. Data Storage

The AIPS++ Measurement Set 2 (MS2) format will be the canonical format for astronomical data products. This will allow CARMA software components requiring high-level science-oriented access to take advantage of the existing functionality of the AIPS++ toolkit. An example of this would be automatic data quality evaluation.

In addition to the visibility data obtained by astronomical observing, the CARMA antennas also produce fast streams of telemetry data, called monitor points. These streams are sampled every half-second and the array as a whole will ultimately contain thousands of monitor points. The monitor data are important for tracking the health of the array, diagnosing problems, and assessing long-term trends. As such, they must be stored in a way that allows easy access and comparison among subsystems. A relational database is a natural solution.

The visibilities from the telescope are initially written as a "binary brick," and subsequently combined with the monitor data to create the MS2 (Figure 1).

2. Archiving

The CARMA Data Archive will be an extension of the BIMA Data Archive currently in use. Within the archive, data are organized in hierarchical collections that reflect how astronomers interact with their data. The broadest collection is a Project covering all data resulting from a single proposal and which can contain a number of different Experiments. Within each Experiment are a number of Trial collections. Data from each observing track will be in its own Trial collection. Processed data are also collected into their own trial collections. As with most archives, users can search and browse the collections through the web.

High level metadata describing the observational experiments are very important for driving the pipeline (see below). These ultimately come from the astronomer during the planning stage and will include science-related information such as the spectral lines of interest and target sensitivity. This information will be used to fill the Observational Programs database (see Figure 1). In addition to being used to schedule the telescope, the metadata will be packaged up with the science (MS2) datasets and shipped to the archive.

New features of the archive include the ability to request the data in any one of the three formats for off-line processing in the AIPS++, Miriad, or Mir packages. This conversion can take place on-the-fly. The converted version will be temporarily cached in the archive in case that format is desired again later. New searching capabilities will also be added to support the searching and downloading of historical engineering data.

3. Pipeline Processing

The CARMA Pipeline will be an extension of the existing BIMA Image Pipeline; Figure 2 illustrates its different components. Processing is triggered automatically whenever new data arrives in the archive. The pipeline analyzes the metadata associated with the collection to determine what needs to be done. This in-

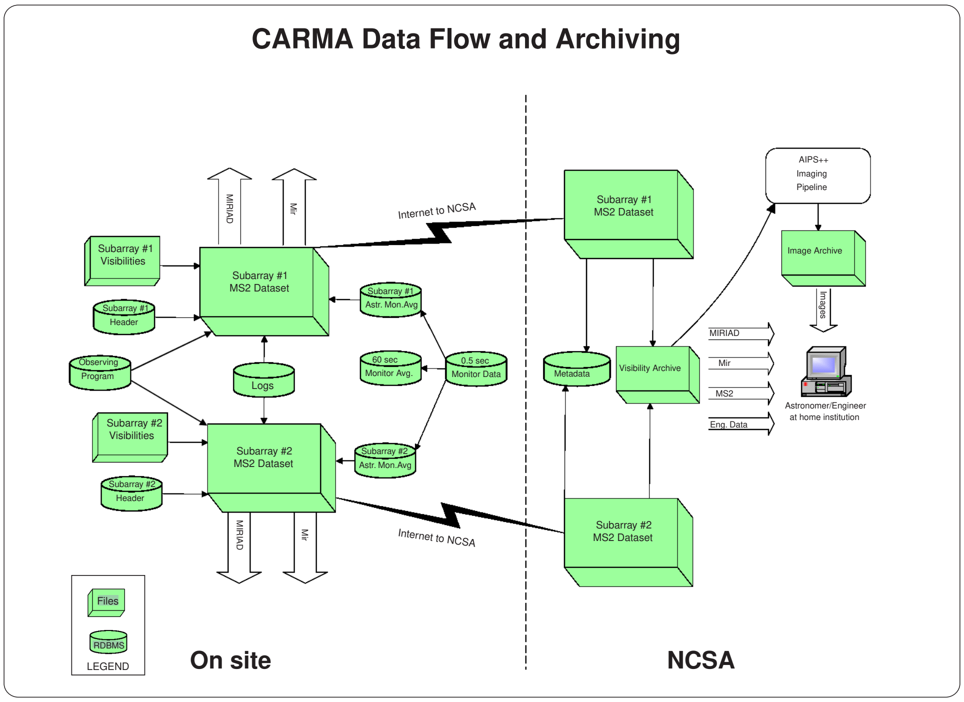

Figure 1. *Left)* The 15 CARMA antennas may operate as one array or, as pictured above, two independent subarrays. For each subarray, visibilities are written out as a binary "brick" with header values stored in databases. Monitor points are stored in 3 databases: at the full half-second rate, in 1 minute averages, and averaged to the astronomical integration time. The relevant pieces are put together into an MS2 data file before shipment to the CARMA/NCSA archive, which happens in near real-time. *Right)* When the data arrive at NCSA, metadata are extracted for entry in the searchable archive. Visibility data are calibrated and imaged using the BIMA Imaging Pipeline (see Figure 2). The astronomer can download the unprocessed visibilities, the calibrated visibilities, and the processed images. The CARMA/NCSA archive will support MS2, MIRIAD, and Mir as export formats for visibilities as well as engineering tables of monitor data. Converters will also be available on-site to allow observers to inspect or analyze the data locally using the respective packages.

cludes special processing parameters and science-related information provided by the astronomer during the planning stage. After processing, the new products—the calibrated visibilities and deconvolved images—are sent back to the archive to be ingested and made available to astronomers. These new data can trigger additional processing; for example, after all requested observing tracks have been calibrated, new processing is triggered to image and deconvolve from all tracks into a single image cube.

The actual processing is done with AIPS++, enabled for parallel processing, using NCSA SGI and Linux clusters. Users not only have access to the processed

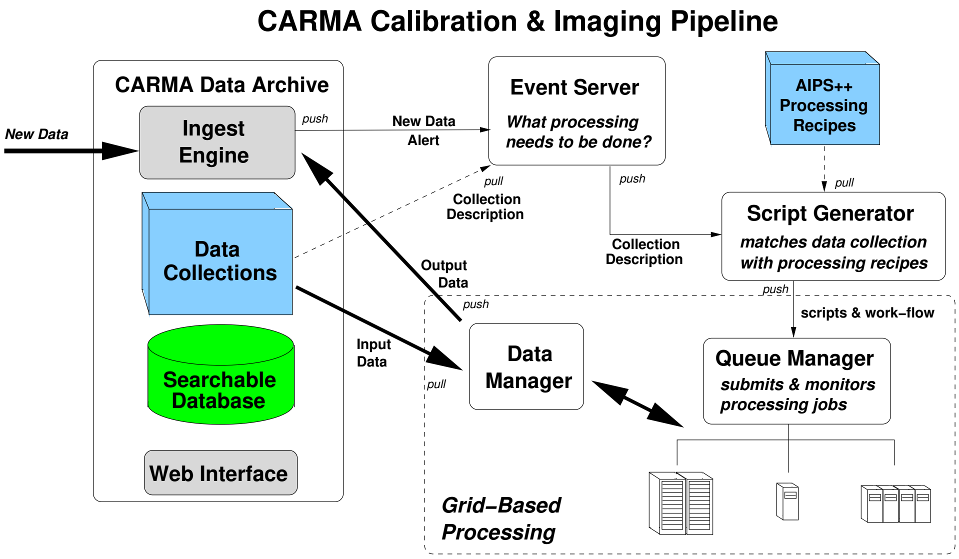

Figure 2. When a new data collection arrives in the archive, a message is sent to the Event Server which figures out what processing needs to be done. This is done by retrieving and analyzing metadata about the collection. The metadata is forwarded to the script generator to prepare the scripts by drawing on "recipes" in a recipe library. The scripts and instructions on what order they should be run (i.e., "workflow") is sent to the Queue Manager. Through the Data Manager, it retrieves the input data from the archive and submits the scripts and data to the Grid for processing. In practice, serial processing (e.g., calibration) is done on different machines from the parallel parts (e.g., imaging). The resulting data products are then sent back to the archive to be ingested.

data, but also the AIPS++ (Glish) scripts used; this allows them to alter and redo the processing off-line.

The use of Grid-based computing technologies will open up interesting opportunities for distributed computing. For example, we plan to use the Teragrid—a national Grid of distributed tera-flop computing linked via high-speed backbone—to process the data. This will allow processing to be distributed between CalTech and NCSA. We can also use the Grid to set up partial mirrors of the archive at the other consortium sites as well as give users greater access to the Pipeline for reprocessing of data.