Archives and Data Access

PASEO Meeting - July 15, 2010

Bryan Butler

EVLA Computing Division Head

Atacama Large Millimeter/submillimeter Array
Expanded Very Large Array
Robert C. Byrd Green Bank Telescope
Very Long Baseline Array

NRAO

Main Elements

Archiving of science data taken with the EVLA contains three main elements:

1. Collection and writing to disk, including organizing the interesting metadata (that which can be searched on - often called the archive "index").

2. Searching the archive.

3. Retrieving data from the archive.

Getting Data on Disk

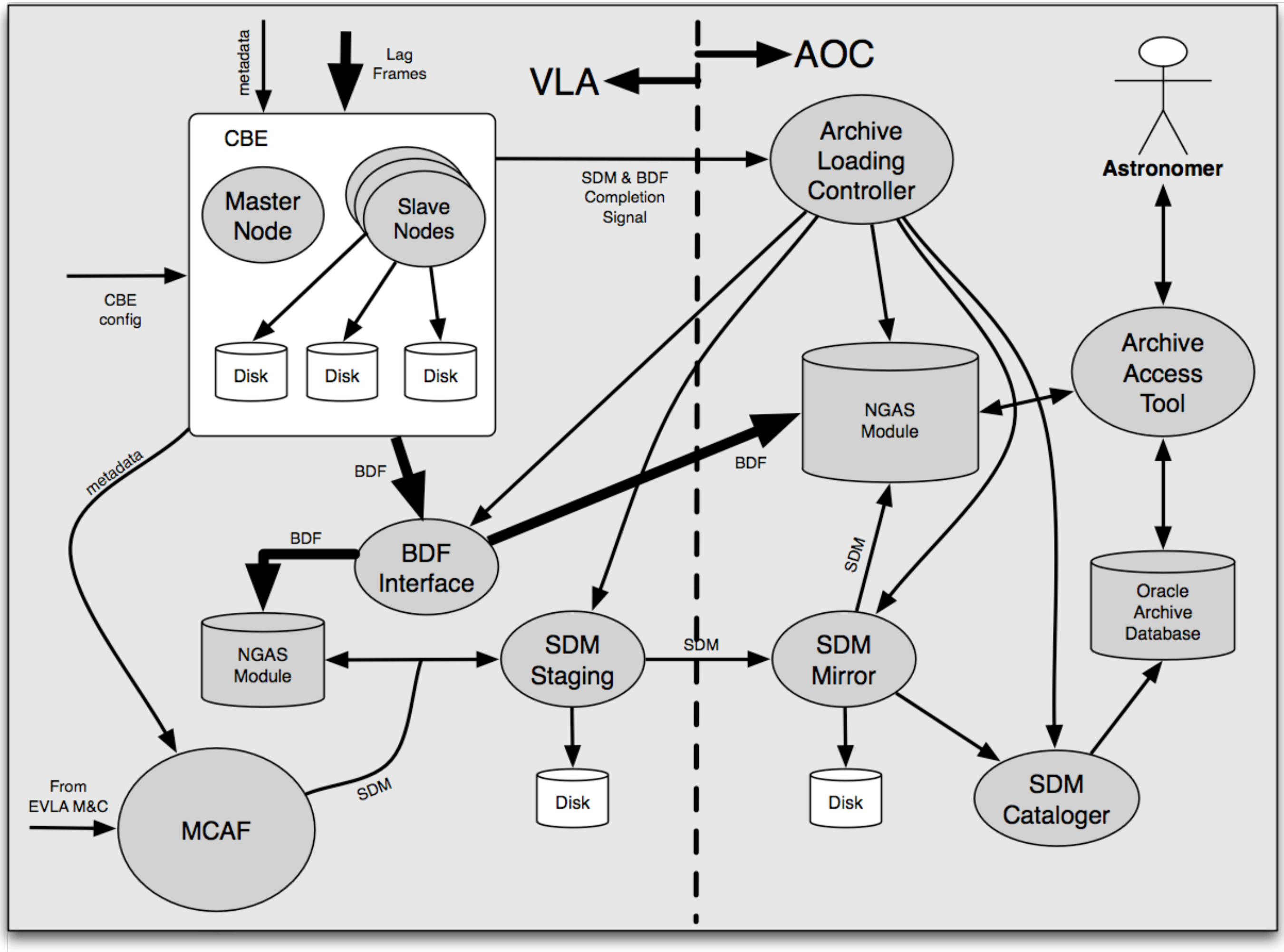
The science archive data is made up of two primary products:

- Visibility data, conforming to the Binary Data Format, a format in common for both EVLA and ALMA. The CBE is the software subsystem that collects the data from the Baseline Boards and organizes it into this format, and writes it to disk. It writes first into a "staging area", a Lustre filesystem at the site.
- Metadata (can be thought of as "headers"), conforming to the Science Data Model, also in common for EVLA and ALMA. The MCAF is the software subsystem that collects the data from various other parts of the M&C system and organizes it into this format, and writes it to disk. This is a number of "tables", which are written as XML files, with links to other tables and the binary data. These are also written into the staging area at the site.

Once data is completely written to the staging area (each subscan is a complete BDF+SDM), it is then copied over to the AOC, to an NGAS storage system, and the "index" entries are created (a task we call the SDM Cataloger).

EVLA

Getting Data on Disk



Searching the Archive

Once the SDM Cataloger has created the entries for an observation in the index database, the observation will show up in the search portion of the AAT. This is the same tool we have been using for almost 10 years for searching the VLA science archive, and works in the same way as it has in the past.

Retrieving Data from the Archive

Once the AAT has found data of interest, it can be retrieved in a number of ways:

- As raw SDM+BDF files;
- As a CASA Measurement Set (MS);
- As a UVFITS file (suitable for ingest into AIPS, for instance - though this will not be true for complex WIDAR configurations).

There is also some time and frequency averaging that can be applied during conversion into MS or UVFITS.

Retrieving Data from the Archive

- Data is written into an ftp area, as it has been in the past.
- Direct copy from the archive to a disk in the AOC takes roughly the same amount of time as observing, for large datasets.
- It is clear that this will not scale to full EVLA dataset sizes (potentially of order terabytes in size) which need to be distributed to observers outside the AOC.
- A plan is being devised to deal with this situation; the current baseline plan is to have datasets up to a certain size retrieved via ftp, larger datasets will be distributed on externally mounted disk (an open question is whether to have these returned or not).

VAO Support

- We currently host several VLA VAO repositories, and support queries against them:
  - FIRST survey
  - NVSS survey
  - “image archive”
- We will continue to host VAO repositories of pipeline-produced EVLA images (cubes and slices therefrom)
- An NRAO-wide approach to VAO support has been developed by the OSO Archive WG which describes the basic infrastructure for how VAO will be supported at NRAO; it implies little additional work beyond what we are already doing for the EVLA science archive
- Support from the NSF VAO grant is forthcoming - personnel and hardware will be supplied from this resource

Pipeline Support

- Information is now written into the SDM to allow pipelines to distinguish "intents" of different scans (in the Scan and SubScan tables). Typical intents are: Observe Target; Calibrate Complex Gain; Calibrate Flux Density Scale; Calibrate Bandpass; etc.
- That information is currently transferred to the CASA MS (into the State table).
- Such information could be used for pipeline reduction of "standard" SBs, but is not currently, because:
  - We need development of the pipeline itself (in CASA), which means we need to understand more fully the best practices for data reduction in general, and need some extensions or improvements to things like automatic flagging.
  - We need OPT "templates", to determine what is "standard".
  - We need a "trigger" for when the pipeline is to be run. ALMA does this through use of what is called an ObsUnitSet; we will need something similar but this is a significant enough change to our PDM that it cannot be undertaken lightly.
- Note that there is support for pipelines in AIPS via "calcodes", which are a proxy for intents, but this is ephemeral.