Data Rates for the ALMA Archive and Control System

Steve Scott & Steve Myers

Revised 03 November, 2000

1) Visibility and Image Data Rates

The ALMA data rate is a combination of the visibility data rate and the image data rate and affects the correlator, control system, pipeline, and archive. The correlator and control system are directly affected by the visibility data rate while the archive size is determined by the visibility data rate and the image data rate. The pipeline capabilities are affected by both the visibility and image data rate, but also by other factors that won't be examined here.

2) Scaling of Data Rates

The data rates discussed here assume a 64 antenna array. The visibility data rates scale with the number of baselines and will thus scale with the number of operational baselines as the array is constructed. The image data rates will scale less rapidly.

3) Atmospheric Phase Correction

Visibility data that has been corrected for atmospheric phase effects shall be available as well as the uncorrected data. In the early years of ALMA, both will be archived. When ALMA is a mature instrument, it will ideally automatically choose the best, on an antenna or perhaps baseline basis. The user shall in principle be able to select whether to archive corrected data, uncorrected, both, or an automatic choice of the best. This selection shall apply to all baseline data. Images will in general only be created on a single data set, not on both.

4) Average Data Rate Specification

A *visibility* is defined to be a single measurement from one baseline in one spectral channel – one complex number. The average visibility data rate for ALMA shall be one million visibilities per second (1.0 MVPS). The average image data rate shall be one-half million pixels per second (0.5 MPPS). These average data rates shall be the average over long periods of time and can be used to determine archive size. Some projects will use data rates higher than the average and some lower.

5) Peak Data Rate Specification

The peak data rates shall be 10 times the average; specifically 10 MVPS and 5 MPPS. It may be necessary to stage high rate data to intermediate storage before archiving.

6) Visibility Data Rate Tradeoffs

The user shall be able to specify the number of spectral channels (including sidebands for double sideband receivers), and integration time to meet the science goals. These choices combined with the corrected/uncorrected selection will yield a total visibility data rate. The user shall specify the recipe for the creation of the archive images, thus specifying the image data rate.

	Visibility Average		Visibility Peak	
	Both	Single	Both	Single
1 sec	250 chans	500 chans	2500 chans	5000 chans
10 sec	2500 chans	5000 chans	25000 chans	50000 chans
30 sec	7500 chans	15000 chans	75000 chans	150000 chans

Tradeoffs between integration time, channels, and phase correction for visibilities. *Both* and *Single* refer to the atmospheric corrected and/or uncorrected data selection.

	Image Average		Image Peak	
	256x256	512x512	256x256	512x512
30 sec	230 chans	60 chans	2300 chans	600 chans
5 min	2300 chans	570 chans	23000 chans	5700 chans
20 min	9200 chans	2300 chans	92000 chans	23000 chans

Tradeoffs between image frequency, image size, and channels.

7) Data Format and Volume

The volume of the archive shall be determined by implementation design decisions. These decisions include the selection of visibility and pixel element storage size and the possible use of compression algorithms. The following table shows the data volume if we assume, for illustration only, 4 bytes per complex visibility and 4 bytes per image pixel.

Data Volume				
	1 sec	1 hour	1 day	1 year
Average Visibility	4 MB	14 GB	350 GB	126 TB
Average Image	2 MB	7 GB	170 GB	63 TB
Average Total	6 MB	21 GB	520 GB	190 TB

8) Archive Contents

The archived data shall consist of the visibility data, images, monitor data, and the scripts used to collect and reduce the data. The visibility data and images shall comprise the majority of the data (>95%). When long integrations are used for the images, the images will be stored in the archive. When shorter integrations are used, the images shall be generated on the fly from the visibilities upon extraction from the archive. The break point between these two techniques shall be determined by the computing capability of the archive extraction pipeline, and may evolve over time. To ensure that images are always available from the ALMA archive for all projects, images must always be archived if the pipeline cannot generate them upon extraction. Images created during a project for feedback or quality control may be stored in the archive to take advantage of existing mechanisms to store and retrieve data. However, these temporary images shall eventually be purged from the archive, recovering the space and ensuring that only final project images exist in the archive.

9) Integration Times

Integration times determine when data is written to the archive. The corrected and uncorrected visibility data shall be integrated over the same time periods. All baselines shall be integrated over the same time periods. Different spectral windows may have different integration times. The average visibility across each window shall be archived on a timescale comparable with the atmospheric fluctuations (approximately one second).

10) Proposal Preparation and Data Rates

The proposal preparation tool shall calculate data rates and total data volume for a project.

11) Administration of Data Rates

The science, scheduling, or operations group shall determine the policies and methods (if any) of allocating and enforcing data rates for projects. There may be restrictions on the allowed combinations of corrected and uncorrected data (for example, recording of both may always be required), and these restrictions may change over time.