Data Rates for the ALMA Archive and Control System

Steve Scott, Steve Myers, & Munetake Momose

Revised 02 May, 2002

This document was produced in response to a request from the ALMA High Level Analysis group to the Software Science Requirements group for clarifications on the ALMA data rates requirements.

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 Pathlength Correction

Visibility data that has been corrected for atmospheric pathlength shall be available as well as the uncorrected data. When ALMA is a mature instrument, it will ideally be capable of automatically choosing the best, on an antenna or perhaps baseline basis; but initially, both corrected and uncorrected data will be archived. The system 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 data from all baselines and the selection shall be a matter of observatory policy to preserve uniformity in the archive. For integration times shorter than the atmospheric coherence time, only one will be recorded. Images will in general only be created on a single data set, not on both.

4. Visibility Data Definition

A *visibility* is defined to be a single measurement from one baseline in one spectral channel in a single sideband or polarization– a single complex number.

5. Average Data Rate Specification

The average visibility data rate for ALMA shall be two million visibilities per second (2.0 MVPS). The average image data rate shall be one million pixels per second (1.0 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.

6. Peak Data Rate Specification

The peak data rates shall be 12 MVPS and 6 MPPS. It is acceptable to stage high rate data to intermediate storage before archiving.

7. 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. The peak data rate for longer integration times cannot be achieved with the Baseline Correlator.

	Visibility Average		Visibility Peak	
	Both	Single	Both	Single
16 msec	N/A	16 chans	N/A	96 chans
1 sec	N/A	1000 chans	N/A	6000 chans
10 sec	10000 chans	20000 chans	30000 chans	60000 chans*
30 sec	15000 chans	30000 chans	90000 chans*	180000 chans*

Tradeoffs between integration time, channels, and pathlength correction for visibilities. *Both* and *Single* refer to the atmospheric pathlength corrected and/or uncorrected data selection.
*Exceeds 32000 maximum number of channels in the Baseline Correlator

	Image Average		Image Peak	
	256x256	512x512	256x256	512x512
30 sec	460 chans	110 chans	2800 chans	660 chans
5 min	4600 chans	1100 chans	28000 chans	6600 chans
20 min	9200 chans	4500 chans	112000 chans*	92000 chans*

Tradeoffs between image creation interval, image size, and channels.

8. 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	8 MB	28 GB	700 GB	252 TB
Average Image	4 MB	14 GB	340 GB	126 TB
Average Total	12 MB	42 GB	1 TB	380 TB

Data volume based on assumption of 4 bytes per complex visibility and 4 bytes per image pixel.

9. 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.

10. 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). The specification for On The Fly Interferometric Mosaicing is 10 msec, but the fastest interferometric readout of the baseline correlator is 16 msec.

11. Proposal Preparation and Data Rates

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

12. 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.

13. Previous Specification and Ramifications

This revision of this document represents a two fold increase in the average data rate and a twenty percent increase in the peak rate. *Adoption of this recommended increase is conditional upon adequate resources being available to the computing division to effect the increase.*

14. Assumptions and Justifications

The ASAC recommended a minimum of 8000 channels for the 2G correlator, with a goal of 4000 to 8000 channels in each of 16 sub-bands. The Baseline Correlator has an output of 4000 to 8000 channels for many interesting configurations, and double sideband receivers can produce twice the number of channels. Assuming a mix that hits the midpoint of each gives 9000 channels. Because the Baseline Correlator can produce 32000 channels in certain narrow band high resolution configurations we have adopted 10000 channels to be a “typical” target. The typical integration time was chosen to be 10 seconds for two reasons. The first is that it matches the fast switching calibration cycle that would be necessary if the atmospheric pathlength correction is not adequate. The second is that it is inbetween the requirements of high temporal resolution projects (sub-second) and the UV cell sampling time for simple imaging (50-100 seconds). It is difficult to forecast the mix of required integration times, so our estimate is not extremely accurate. Our typical number of channels and integration time (10000 channels at 10 seconds), define the average data rate for recording both corrected and uncorrected data.

15. Open Questions

This revision was prompted by an examination by the ASAC of the specifications for the proposed 2G correlator. The data rates proposed here are not adequate for the 2G correlator, therefore proposals for a 2G correlator should include additional computing components for the necessary increase in data rates.

On The Fly Interferometric Mosaicing has a specification of a 10 msec minimum integration time which demands a very high data rate. While the Baseline Correlator limits the integration time to 16 msec, the present peak data rate yields only 96 channels. Further increases to the data rate will of course require additional resources, but modest increases in the peak rate (> 20%) may run into technological thresholds where the incremental resources do not increase linearly and also where additional engineering would be required.