Alma Memo 501
Estimation of ALMA Data Rate

R. Lucas (IRAM), J. Richer (MRAO), D. Shepherd (NRAO),
L. Testi (Oss. Astrofisico di Arcetri), M. Wright (BIMA), C. Wilson (McMaster University)

2004-06-07

Abstract

In order to assess more firmly the estimated data rates on which the ALMA Computing IPT has based the design of software and hardware, the SSR has calculated the data rate for a sample of programmes in the DRSP prepared by the ALMA Science IPT. Due to the limitations in the process, the results are lower limits to the required data rates, which are compatible with the present ALMA specifications.

1 Introduction

The rate at which data are produced is naturally an important design parameter of ALMA Software: the average data rate controls mainly the size of the Science Archive, while the peak data rate sets performance needs for key elements such as the correlator/archive interface, and the Science Pipeline.

The Science Software Requirements Committee (SSR) had evaluated the data rates and included then it the requirements: see the main requirement document[1] and a specific note on data rates [2]. The data rate had been estimated as 6 MB/s and 60 MB/s for the average rate and the peak sustained rate, respectively. The images were tentatively considered to account for one third of this rate.

The recent availability of the Design Reference Science Plan (DRSP) developed by the Science IPT [3] has enabled the SSR to refine the evaluation of the data rates. The preliminary results have been available to the software developers since december 2003. The purpose of this memo is to make them available to the larger ALMA community.

2 Procedure

In order to give a rapid answer the SSR has chosen to sample only a fraction of the DRSP programmes, and for efficiency the SSR members were assigned programmes in their best field of competence. They did a first evaluation of the data rate for these programmes; on this basis a spreadsheet was written to perform the calculations using a common set of rules. These rules are based on the SSR requirements, and on our general understanding of the planned procedures used for observation and calibration.

This study, based on the DRSP, applies to ALMA as a mature instrument like this set of programmes.

The rules are the following:

Visibility format: The standard is 4 bytes per visibility. We do not need to store weights for every baseline visibility as the weights are easily factored into antenna-based parts: $W_{ij} = \sqrt{W_i * W_j}$. The cost to keep baseline-based weights would be a factor 1.5 on the visibility data rate. This does not preclude introducing baseline-based weights later in pipeline/off-line data reduction when it is needed there (as e.g. the AIPS++ measurement set has baseline-based weights).

Number of spectral channels: We take the ratio of the bandwidth required in km/s to the resolution in km/s, and we multiply by two to account for a full Nyquist sampling of the spectra. This assumes that the unnecessary parts of the correlator output is dropped in real time by correlator software. We only keep the spectral part required by the observer.

Number of sidebands: When they are separated by 90-deg phase switching, i.e. for bands 8-10 we assume we keep both sidebands if required (the image sideband may be dropped).

Atmospheric Path length Correction: This study applies to ALMA as a mature instrument, when we should be able to determine in quasi-real time whether the WVR-based path length phase correction improves the results (SSR Requirement 2.3-R6). So we take one result only here (corrected or uncorrected), not both as will be the default in the early science years.

Integration time: For adequate sampling of the visibility plane, the integration time should be smaller than $82/b$ if b is the maximum baseline in kilometers. Nevertheless we set a maximum of 45s, as phase calibration will be needed (see below).

Calibration: We will have reference phase calibration with a cycle time ranging from several minutes to 20 seconds, depending on the atmosphere conditions. According to Mark Holdaway the time efficiency is 80% for 20s cycles. For 60s cycles it is 95%. We have planned to have a self-tuning cycle time in standard observing procedures, to get the phase rms below the value specified by the observer. We thus assume a time efficiency TE of 0.9 on average (corresponding to an average on-source time of 45s). Now the calibration also increases the data volume, by:

$$1 + \texttt{visibilitiesonthecalibrator/visibilitiesontarget}$$

That is, for continuum projects, the data volume is doubled, while for spectroscopy, it is only moderately increased. We assume we keep 32 channels per 2GHz baseband on the phase calibrator, so $8 * (32 + 5) = 296$ visibilities per baseline (we add five 0.1sec sub-integrations on the calibrator, for which channel averages only are recorded). So to take calibration into account we multiply the data rate by: $0.9 * (C * P * S + 296)/(C * P * S)$, where C is the number of channels, P that of polarization products, S the number of sidebands (when separated by 90-deg switching). In other words we add 296 to the product $C * P * S$ and we multiply by 0.9.

Observing Modes: We have considered the following main observing modes:

1. Single-field: Apart from calibration as above, all time is on-source.
2. Pointed-Mosaic: We add 3s per pointing to the total time; sampling is assumed at 0.45 HPBW in both directions.
3. OTF-Mosaic: We add 5% to total time due to turn-around time; we assume a spacing of 0.45 HPBW between rows; and 4 integration times per HPBW crossing time.

We did not include single-dish observing, as the data rate is expected to be much lower than for interferometry, and as the actual observing modes were not well specified.

Images: The following assumptions were made:

- We store deconvolved "final" science images (not dirty images or dirty beams). The pixel separation is assumed chosen as one third of the synthesized beam width.
- We store the images in all cases (ignoring requirement 7.2-R12 for the time being, as the criteria for choosing between image storage and image on-the-fly reconstruction from the archive depend on pipeline performance and are difficult to estimate at this time).

Following those rules an Excel spreadsheet was set up (see AllDRSP.xls on the SSR Twiki, and its pdf version AllDRSP.pdf) to estimate the data rates based on user input. In the spreadsheet:

- The first page contains the summary of results and the recipes outlined above.
- The second page contains the observer's input from the DRSPs
- The third page contains the detailed calculations.

3 Results and Discussion

Table 3 lists the programmes we have included. Table 3 gives the results in terms of data volumes and rates, average and maximum, for the visibilities and images.

We obtain an average visibility data rate (4.4 MB/s) well within the current specification (6 MB/s). The image rate is 23 times lower than the visibility rate. This ratio had been conservatively estimated to be a factor of two. This is due to the relatively long programmes that have been proposed, with a large number of deep searches.

However several points must be noted, all leading to higher estimations of the data rate:

- Our sample (32 DRSPs) may not be fully representative. We believe that we have sampled the main observing modes present in the DRSPs.
- The maximum data rate is highly dependent on the integration time for the interferometric on-the-fly, currently set to 0.5 second. This number is no more than a guess, as this observing mode needs to be experienced on a real telescope. Nevertheless there appears to be no technical reason to perform this sort of observing at specifically high scanning rates; in most cases the integration time per *uv* point will have to be larger, due to the sensitivty constraints. The two programs with the largest data rates are an interferometric on-the-fly observation (133 MB/s), and a spectral survey of galactic absorption (145 MB/s). These programmes account for 87h together or 1.7% of the total. The programme with the next highest data rate reaches only 13.3 MB/s.

  We think both projects can actually accomodate the current specification without reduction in scope.

  In these two cases it is useful to try and estimate the time extent during which the peak data rate must be sustained.

  · For the galactic absorption programme there is no real constraint, as the angular resolution is little relevant (a point source is observed, much smaller that the synthesized beam in any antenna configuration). So there is no requirement that such a program should be performed in a short time period.
  · For the OTF maps, the sensitivity is obtained by adding individual maps of about 4 hours duration each; each map should be performed preferrably without moving the antennas, which probably means inside a 2-4 day period. Now while all maps are executed, the configuration may change without e.g. modifying the synthesized beam by more that $\sim 30\%$; which means the 80 hours of data should be taken in a period of approximately one week.
- In this study we have limited the spectral coverage to the proposer's proposed bandwidth. In general the correlator will analyze a wider bandwidth (as the correlator modes generally step the number of channels by factors of two), and more channels than required will often be available at the proposed resolution. This policy is sparing the size of the computing system (in particular the archive), but limits the chances of serendipity discoveries (a new. unexpected line appearing in a spectral band expected to be empty of emission). When writing an actual proposal, scientists usually tend to make a more complete use of the hardware to get the maximal science return, by e.g. including spectral bands set up to measure lines which, while not needed for the main science proposed, are sometimes leading to serendipitous discoveries, and quite often give hints for further studies; or simply by extending the spectral coverage of the main science bands for the same purpose. Note

Table 1: DRSP included.

Project	Ref Title
1.1.1	Unbiased survey of submm galaxies - 1
1.1.2	Unbiased survey of submm galaxies - 2
1.1.3	Unbiased survey of submm galaxies - 3
1.1.4	Unbiased survey of submm galaxies - 4
1.3.1	Spectral line survey in high-z molecular absorption systems
1.3.2	Deep search for new molecular absorption line systems
1.4.1	Sunyaev-Zel'dovich Effect of Proto-Clusters
1.7.4	Structure of the ISM in irregular galaxies
1.7.5	Low Frequency of Free-Free Emission in Nearby Starburst Galaxies
1.7.10	Searching for Proto-Super Star Clusters in the Antennae
2.1.1	Small scale structure of molecular clouds
2.1.2	Density and temperature profile in pre-stellar cores
2.1.3	Kinetic Temperature Structure in Protostars and YSOs
2.1.4	Density and temperature profile in high-mass cores
2.1.5	Spatial Density Probe Comparison in Protostars and YSOs
2.1.6	The Connection Between Cloud Structure and the IMF
2.1.7	Physical Structure of Low-mass Star-Forming Cores
2.1.8	Infall velocity structure of starless cores
2.1.9	Envelope Structure of Intermediate-Mass YSOs
2.2.1	Mapping the turbulence in a molecular cloud
2.2.3	Structure and collapse of protostellar envelopes
2.2.5	Magnetic field in molecular outflows
2.2.6	Energetics of the HH80-81 molecular outflow
2.2.10	The internal structure of the BHR71 outflow
2.3.6	Survey of interstellar HCO+ absorption
2.3.7	Surveys of interstellar molecular absorption
2.4.4	Disks in the sub-stellar regime
2.4.6	Transition disks around CTTs/WTTs & near ZAMS stars
2.4.8	Structure of disks around high-mass protostars
3.3.2	Line surveys in evolved stars
3.5.4	The populations of relativistic particles and magnetic field structure in the Crab Nebula and other plerions
3.5.5	ToO Observing of Radio Supernovae

Table 2: The main results of this study.

Time:		Visib.:		Images:	
Total (h):	4935.8	Total Size (TB)	78.5	Total Size (GB):	3471.
Average (h):	69.5	Average Size (TB):	1.1	Averagel Size (GB):	49.6
Maximum (h):	450.0	Maximum Size (TB):	38.5	Maximuml Size (GB):	699.8
		Average Rate (MB/s):	4.4	Average Rate (MB/s):	0.195
		Maximum Rate(MB/s):	145.5	Maximum Rate (MB/s):	2.43

however that the DRSPs include a number of systematic unbiased searches, both in the spatial and in the spectral domain.

An additional uncertainty remains: the full usage of the correlator, including the proposed tunable filter, was not fully available the scientists at the time of the redaction of the proposals (we have tried to correct for this in obvious simple cases).

4 Conclusions

We thus consider the numbers in table 3 as lower limits. The current specification (6 MB/s) is about a factor of two higher, and we conclude that it should be kept, as it is adequate given what we know of the science programmes at this time, and given the assumptions we have made. The lower limits to the average data rate in table 3 (4.4 MB/s) imply that PIs will be able to get more science than they were anticipating. Reducing the specification on the basis of this study would be highly risky and we strongly discourage any such attempt. We also note that with the new tunable filters designed for the correlator, the current peak data rate specification will cause a small minority of programmes to be data rate limited.

The Science IPT has a plan to upgrade the DRSP by getting more detailed input from the proposers [3]. The data rates should be computed in detail at this occasion.

References

[1] SSR Committee, ALMA Science Software Requirements and Use Cases, ALMA Software memo 11 , also known as ALMA-70.10.00.00-002-I-SPE.

[2] Steve Scott, Steve Myers, and Munetake Momose, Data Rates for the ALMA Archive and Control System

[3] E.F. van Dishoeck, A. Wootten, S. Guilloteau, The ALMA Design Reference Science Plan: Version 1.0, December 10 2003