

Data Rates for the ALMA Enhanced Baseline Correlator

Stephen Scott
10 February 2003

1. Status

The current data rate is specified in the SSR Requirements in terms of visibilities and pixels per unit time to intentionally avoid implementation issues. However, for convenience we have assumed a reasonable conversion to bytes, without overhead, to get the table shown below. The data rate affects the design and cost of the correlator data processing system, the pipeline, the realtime network, and the archive.

	Visibilities	Images	Total
Average	4 MB/sec	2 MB/sec	6 MB/sec
Peak	40 MB/sec	20 MB/sec	60 MB/sec

Current data rate from SSR Requirements

There was an SSR request in May 2002 to double the average total rate and marginally increase the peak that is shown in the table below.

	Visibilities	Images	Total
Average	8 MB/sec	4 MB/sec	12 MB/sec
Peak	48 MB/sec	24 MB/sec	72 MB/sec

SSR proposed data rate increase May 2002

This request has not been acted on at this time and will be replaced with a new request from the SSR that is matched to the proposed enhancements to the baseline correlator. This new request must be consistent with the ALMA scientific program and guidance is sought from the Science IPT. The rest of this report should provide enough information to translate science programs into data rates.

2. Assumptions

Most of the factors necessary to fully examine data rates have been covered in "Data Rates for the ALMA Archive and Control System" by Scott, Myers, and Momose. We extract some simplifying assumptions from that report without repeating the justifications and add a few others. Some of the assumptions apply to any data rate, others just to average rates.

- Both atmospheric pathlength corrected and uncorrected data will always be recorded for integration times longer than 1 second (SSR requirement)
- A complex visibility takes 4 bytes of storage (SSR requirement)
- An image pixel takes 4 bytes of storage
- Only one of pathlength corrected or uncorrected data will be imaged by the pipeline
- Average integration time is ten seconds
- Average number of sidebands equals 1.4. Justification: about half the receivers are double sideband and 80% of the time when a double sideband receiver is used both sidebands are recorded, giving an average of 1.4 sidebands
- Average image is 440x440 (average of 256x256 and 512x512)
- Average final image creation interval is 20 minutes (duration of an experiment)
- Ignore autocorrelation spectra
- Ignore 1 second sampled band average data
- Ignore headers

The items that have been ignored could contribute as much as 10% to the data rate but are difficult to accurately quantify.

3. The enhanced correlator

ALMA Memo 441 describes the proposed enhanced correlator. The 32 filter bank option gives 8K channels from each of the four basebands for all bandwidths. The latest report is that the 32 filter bank option does look feasible, giving a channel increase of a factor of 32 for the 2 GHz wide band. To achieve the increased number of channels, the band is decomposed using FIR filters and then reconstituted at the end. There will be slight dips in the bandpass shape where the filters meet. These should be stable and removed by normal calibration, although the signal to noise will be affected in these regions.

Bandwidth	Baseline	Enhanced
2 GHz	256	8192
1 GHz	512	8192
500 MHz	1024	8192
250 MHz	2048	8192
125 MHz	4096	8192
62.5 MHz	8192	8192

Number of channels per baseband for the 32 filter bank enhanced option

Polarization processing in the correlator increases the number of data sets per baseband but proportionally decreases the number of channels so that the total number of channels summed across all data sets is still given by the table above.

4. Data rate equations

The equations for the visibility rate and the image pixel rate in Bytes/sec are given below and only rely on the assumption of the storage size for a data item.

$$visRate = \frac{4bytes \times 2016baselines \times N_{chan} \times nSidebands \times apcFactor}{t_{integ}}$$

$$pixRate = \frac{4bytes \times N_{chan} \times nSidebands \times size^2}{t_{image}}$$

where

- N_{chan} total number of channels, summed over all basebands
- $nSidebands$ number of sidebands; 1 for ssb, 2 for dsb
- $apcFactor$ factor for atmospheric pathlength correction; 1 for $t_{integ} < 1$ second, 2 for $t_{integ} > 1$ second
- t_{integ} integration time in seconds
- $size$ number of pixels on each side of the image
- t_{image} image creation interval in seconds

Using the assumptions of section 2, the data rate equations for average conditions become

$$avgVisRate = 0.002 \text{ MB/sec} \times N_{chan}$$

$$avgPixRate = 0.0009 \text{ MB/sec} \times N_{chan}$$

$$avgTotalRate = 0.0029 \text{ MB/sec} \times N_{chan}$$

5. Scientific and technical data rates

All of the previous specification of data rates by the SSR has been done in terms of visibilities and pixels. With these units there is no doubt about how much science data you will get. The computing team of course needs to add on headers and data packaging and transmission overheads, and in the end they must deal with real bytes rather than visibilities

and pixels. It is convenient to use the term “scientific data rate” when referring to the data rate in bytes without overhead. The term “technical data rate” refers to the data rate including overhead that must be used to size networks and archives. In the past a distinction has probably not been made between these two rates, but should be in the future. It is not clear at this time what factor should be used to get the technical rate from the scientific rate. We only deal with scientific data rate here.

6. Scaling laws and technical breakpoints

The channel parallel nature of ALMA means that in many cases increases in data rate may be handled simply by scaling the amount of computing hardware or archive media needed. Operational costs scale as well as capital costs, particularly for archive media. In some cases scaling fails as a breakpoint is reached at which the system needs to be redesigned. An example would be increasing the data rate beyond the capabilities of the chosen network hardware. These breakpoints should be identified by the computing team to help guide any modifications to the data rate.

7. Peak rates

Much of the computing hardware and software that has been recently designed has been done with a view to the increase proposed by the SSR in 2002, and this design is most sensitive to the peak rate. There is a strong sentiment in the computing group to limit the peak rate to the proposed 72 MB/sec. This is a reasonable constraint as the peak rate allows the full 32K channels of the correlator at 10 second integration times or longer with single sideband data. If this peak rate is not exceeded, then increases in the average rate should mainly involve more hardware and media for parallel processing and archiving, without major impacts on the fundamental design. A peak scientific data rate of 72 MB/sec is recommended.

8. Average rates

The main unknown in estimating an average rate is the number of channels used across the mix of ALMA projects. The approach of using the maximum available resolution over the bandwidth dictated by the science guarantees no loss of information while generating the highest data rate. Possible side effects of the large data sets produced are long download times and the lack of resources to reduce the data at the home institution. Interacting with data sets with very large numbers of channels may also be challenging. The 2002 SSR average rate translates to a little more than 4100 channels. Given the capabilities of the enhanced correlator, an increase to the average rate of a factor of two (or more) might seem reasonable. But this is clearly an area where the expertise of the Science IPT will prove very beneficial.

9. Other issues

Identification of any technical breakpoints would be very beneficial for future work on data rates. Any request for an increase in data rates will first have to be studied for cost implications by the Computing IPT and is of course ultimately subject to change control and approval by senior project management.