VLBA ARRAY MEMO No. 47

December 31, 1981

To: VLBA Study Group

From: R. C. Walker

Subject: Data storage requirements for the VLBA

The amount of storage space required for a VLBA data set is a function of several variables. Rather than trying to provide tables showing the effects of these variables, I will give the equations for the number of bytes needed for both spectral line and continuum observations as a function of all relevant parameters. A few sample cases will be evaluated.

The number of bytes in a data set is given by:

$$\# \text{ Bytes} = 3600\ w\ H\ B\ C / t$$

where:
- w = Number of bytes per complex data point (Minimum 4, probably much more. VLA export format uses 48 for u,v,w,4 polarizations and bookkeeping. Fewer are needed for spectral line as u,v,w and bookkeeping can be shared.)
- H = Hours of data in observation (typical 10)
- B = Number of baselines (probably 45)
- C = Number of delay or spectral channels (see below)
- t = Integration time in seconds (see below)

The number of channels in the continuum case is determined by the delay range needed for the desired field of view and by the uncertainties in the clocks. Note that the field of view may be set by the degree of uncertainty in the a priori source position rather than the size of the source. This should not normally be a problem because accurate positions can be found very quickly with the VLA. The number of channels required by the field of view is:

$$C = 6.46E\text{-}5\ l\ b\ x$$

where:
- l = length of longest baseline in km (usually 8000)
- b = bandwidth in MHz (up to 50)
- x = maximum offset from phase center for good data. in arc seconds (typically 0.1 to 1.0)

The number of channels required by the uncertainty in the clocks is:

$$C = 2\ b\ T$$

where: b = bandwidth in MHz
T = clock uncertainty in microsec (usually 0.25)

The total number of channels is the sum of the above:

$$C = 6.46E\text{-}5\ l\ b\ x \quad + \quad 2\ b\ T$$

For spectral line data, the number of channels is determined by the desired velocity range and resolution. If the number of channels is fixed by the correlator, it probably will have a maximum value of 512. If the correlator is of a recirculating design, the number of channels could be very large although we should not attempt to support post-processing of absurd cases.

The integration time will usually be set by the need to avoid time average smearing for sources in the field of view. To give an idea of the magnitude of the problem, the fringe period for an object one arc second away from the phase center for observations at 22 GHz (reasonably common H2O maser case) can be as short as 4.8 seconds. To avoid smearing, the integration time should allow several points per fringe period. A large amount of space could be saved by making the average time a function of baseline length but that possibility is not included in the following formula. The longest safe integration time is given by:

$$t = 8.5E5 \ / \ (\ x\ f\ l\ p\ )$$

where: x = The maximum offset from the phase center for good data. This usually will be set by the a-priori's although in many spectral line sources it will be set by the separations of features.
f = Frequency (GHz)
l = Maximum baseline length in km (usually 8000)
p = The minimum number of points per fringe (57)

The above formulae can be combined to give a composite formula for the number of bytes needed:

$$\#\ bytes = .00423\ w\ l\ B\ f\ b\ H\ x\ p\ (\ 6.46E\text{-}5\ l\ x \quad + \quad 2\ T\ ) \qquad \text{(continuum case)}$$

or:

$$= 4.24E\text{-}3\ w\ l\ B\ f\ H\ x\ p\ C \qquad \text{(line case)}$$

Note the dependence on the square of the baseline length and the field of view in the continuum case as long as the number of channels is not set by the knowledge of the clocks.

The table attached to this memo shows several cases of interest. In general, the data sets are a few times larger than full track VLA data sets with the exception of the extreme H2O spectral line case. That case is clearly too large to be reasonably supported so I suggest that we not try. Observations of sources such as H2O in ORION will have to be done with very much less than full tracks and full resolution. I suspect that most of the science in such observations can be obtained with more modest observations. The requirements for the modest H2O case are still very large but not impossible. The number of 6250 bpi tapes required is similar to the number of 1600 bpi tapes used in H2O VLBI experiments in the past and the data set is only 4 times the size of a large spectral line VLA data set that might be produced when the current channel limitations are overcome.

I am not sure how we should deal with the disk space requirements. The above information indicates that we should be prepared to deal with quantities of data somewhat larger than what the VLA generates. For continuum observations, this should not be a serious problem, especially since the tabulated data set sizes are those required before fringe fitting only. The spectral line case presents more of a problem, although it is a problem that will have to be faced for the VLA too. Mark Reid advocates storing all of the data on disk (after considerable trouble using tapes with a very large H2O VLB experiment he is now working on). This could require as much as 100 gigabytes of disk which is excessive now but may not be in a few years. I can think of several ways to be clever and reduce this requirement dramatically but history has shown that we always think of ways to increase the requirements too.

One area where I suspect that our initial impressions that the VLBA is easier than the VLA are still correct is in mapping. The coverage of the VLBA for snap-shot observations consisting of a single short integration will not be nearly as good as that of the VLA (8 times fewer baselines) so I suspect that a larger percentage of the observations will involve long tracks, or at least several scans at different hour angles, so there will be fewer maps per unit observing time. Also there will be less input data for each map which reduces the sorting and gridding which are an important part of the computing load with large data bases. This will be somewhat offset by the need for many iterations of self-calibration for continuum data. Note that, except for the reference channel, the iterative self-cal will not be needed for spectral line data.

VLBA DATA BASE SIZES

Parameters:

W	!	Bytes per visibility record. -estimate includes u,v,w,time,baseline,stokes parameters
L	!	Maximum baseline in km.
B	!	Number of baselines.
F	!	Frequency in GHz
BW	!	Bandwidth in MHz.
H	!	Hours of data per baseline.
X	!	Maximum offset from center of field.
P	!	Minimum points per fringe period.

Results to be calculated in subroutine:

C	!	Number of channels (fixed for line) - for continuum, set by delay window
CU	!	Number of channels used (allows for clock uncertainty - usually .25 microsec)
T	!	Integration time. Set by smearing. Will not exceed specified maximum.
MBYTES	!	Megabytes required.
MVIS	!	Millions of visibility records.
TAPES	!	Number of 6250 bpi tapes (180 Megabytes).

Comment	W	L	B	F	BW	H	X	P	C	CU	T	MBYTES	MVIS	TAPES
Moderate continuum case	42	8000	45	10.6	28.00	10.0	0.20	5.0	3.	15.	10.03	101.085	2.437	0.56
Extreme continuum case	24	8000	91	43.0	56.00	10.0	0.20	5.0	6.	31.	2.47	979.505	40.813	5.44
Hot spots	42	4000	35	1.6	28.00	10.0	2.00	5.0	14.	26.	13.29	105.506	2.512	0.59
Fake data tests	24	8000	45	5.0	28.00	10.0	0.04	5.0	1.	1.	121.24	0.321	0.013	0.00
Moderate H2O case	10	8000	45	22.2	0.00	10.0	1.00	5.0	256.	256.	0.96	4330.816	433.082	24.06
Extreme H2O case	10	8000	45	22.2	0.00	10.0	15.00	5.0	512.	512.	0.06	129924.477	12992.447	721.80
OH case	10	8000	45	1.6	0.00	10.0	2.00	5.0	512.	512.	6.64	1248.524	124.852	6.94
VLA continuum case - export	48	35	351	5.0	50.00	10.0	10.00	5.0	1.	1.	20.00	34.329	0.715	0.19
VLA H2O case	10	35	351	22.2	0.00	10.0	2.00	5.0	256.	256.	30.00	1078.272	107.827	5.99