EVLA Memo 63
Scaling Relations for Interferometric Post-Processing

Rick Perley and Barry Clark

September 16, 2003

Abstract

We present simple scaling laws which will allow estimation of computing post-processing imaging costs, parameterized by array scale, antenna size, antenna number, and wavelength. Full-beam post-processing costs for large-N (small-D) arrays will be very expensive, and must be included in any total cost for an interferometric array. We present two simple cost formulations which include these post-processing costs.

1 Introduction

Interferometers are imperfect. Raw ('principal solution') images made from an interferometric array are nearly always limited by the sidelobes of the synthesized beam – typically a few percent of the peak – rather than by thermal noise. There are of course very effective and well-established methodologies for deconvolution of the imperfect image, but these are nearly always computationally expensive, easily dominating the total computing effort.

A major factor in judging array design and cost is the cost of data processing. In general, employing smaller antennas means there must be more of them to preserve point-source sensitivity, which thus generates a larger database. The problem is compounded because the larger primary beam necessitates more frequency channels and faster time sampling. All of these factors generate larger databases, and more post-processing.

The purpose of this memo is to generate rough scaling laws so that array designers can estimate the implications of antenna size. We have also attempted to estimate the coefficients of these laws, so that a rough cost equation can be made.

2 Basic Assumptions

We assume an array comprising N elements, each of diameter D, which spans a maximum dimension B, operating at wavelength λ with a fixed total collecting area, so that the product ND^2 is constant. This is appropriate when sensitivity is the prime design driver. We consider the 'cost' of imaging, and assume that this is proportional to the product of the number of images to be made, multiplied by the size of the database from which these images are derived:

$$\text{Cost} \propto N_{maps} V_{data} \tag{1}$$

3 The Number of Maps

The situation we consider is full-beam imaging, which implies that the entire solid angle subtended by the antenna element's primary beam must be imaged in order to remove the sidelobes of all the background objects to the level of the thermal noise. A simple analysis will illustrate the nature of the problem. Suppose within the primary beam there are N_s point sources with apparent flux densities of S_i Jy, and that the PSF has an rms sidelobe level of R, expressed as a fraction of the peak. Then, the rms noise level due to the sidelobes can be estimated as

$$\sigma_c = R\sqrt{\sum_{i=1}^{N_s} S_i^2} \tag{2}$$

The sum is over the apparent flux densities of the sources as attenuated by the antenna primary beam. For simplicity, we assume there is a mean value, S_m. Then, for N_s such sources, we obtain

$$\sigma_c = \sqrt{N_s} R S_m \tag{3}$$

We can roughly estimate the sidelobe noise level from our experience with L-band data: $N_s \sim 100$, $R \sim 0.01$, and $S_m \sim 10$ mJy. For these conservative values (some would argue they should be much higher), the estimated sidelobe noise level is 1 mJy, or 100 times the thermal noise limit for the EVLA at this frequency.

The accumulated sidelobes from background objects cannot be ignored, and the number of background objects means that at low frequencies at least (say, below 8 GHz for EVLA sensitivities[1]), it will be necessary to image the entire primary beam to find and remove these background objects.

How can this best be done? The VLA/EVLA is a two-dimensional array, which means it is coplanar only instantaneously. As originally discussed by Clark[1], and expanded upon by Perley[2] the normal 2-dimensional FT relationship cannot be applied to visibility measures taken in a 3-dimensional volume without significant aberrations. The formal solution is a 3-dimensional transform, within which the sky emission appears on a sphere of unit radius. But this approach is highly inefficient – for an array with the resolution of the full EVLA (350 km baselines), the 'depth' of the necessary transform is large – $\sim 3\lambda B/D^2$ cells, roughly 100 cells at 1 GHz, and over 2000 cells at 74 MHz. Stated another way, over 99% of the cells in the '3-d' image volume are empty of sky emission, so their computation represents little more than wasted effort.

Other, more efficient approaches are discussed by Cornwell and Perley[3]. The widely adopted solution is to cover the celestial sphere with a mosaic of smaller 2-d tangent images, each of which is sufficiently small to minimize aberrations. The computations then cover (approximately) only the surface of the sphere. A distinct advantage of this approach is that direction-dependent calibration constants (such as atmospheric/ionospheric phase errors) can more easily be implemented. The disadvantage is that the entire data volume must be regridded for each of the sub-fields, with a (u,v,w) rotation to keep the image plane tangent to the celestial sphere. Both `AIPS` and `AIPS++` have extensive software to accomplish this, and the method is now well established and highly effective.

The number of subfields required turns out to be the same as the depth of the required 3-d transform: $\sim 3\lambda B/D^2$. For the full EVLA, this will vary from less than 10 at 40 GHz, to over 1000 at the lowest frequencies.

However, a single image per subfield is not sufficient. Each subfield can deconvolved individually only until the remaining emission is comparable to the sidelobe levels from objects outside that subfield. At this point, the effects of the sources discovered so far must be removed by u-v subtraction from the original database, and the entire mosaic of subfields recomputed. With current VLA sensitivities, each subfield will be be recomputed many, perhaps dozens, of times[2]. As the decision of when the entire set is to be re-imaged depends on the level of the sidelobes, having a 'good' PSF is clearly advantageous.

[1]The frequency above which this approach will no longer be necessary is keenly debated at the AOC. Some say as low as 4 GHz, other say as high as 18 GHz. All estimates so far are based on intuition, experience or prejudice. Eventually, somebody will have to do a proper sum over the beam-weighted number counts.

[2]With the much higher sensitivity of the EVLA, the number of 'major cycles' will increase dramatically – a factor not included in this analysis

The scaling law for the number of times the set of subfields must be recomputed is straightforwards: If there are N antennas in the array, the sidelobe levels scale as $1/N$. A wavelength dependence can be added easily – if the array has a flat sensitivity with frequency, (approximately true down to a frequency of approximately 500 MHz where the galactic background begins to dominate), the 'constrast' between background objects and thermal noise will rise with the mean spectral index – say, as $\lambda^{0.7}$. Thus, we estimate the number of re-imaging cycles to scale with $\lambda^{0.7}/N \propto \lambda^{0.7}D^2$, and the total number of images (and hence the number of times the entire data volume must be transformed) to scale as:

$$N_{maps} = N_{fields}N_{cycles} \propto \frac{\lambda B}{D^2}\frac{\lambda^{0.7}}{N} = \frac{\lambda^{1.7}B}{ND^2} \propto \lambda^{1.7}B. \tag{4}$$

where the last proportionality has utilized our assumption that the total sensitivity (and hence collecting area) is a constant.

However, if the antennas in the array are grouped, for the purpose of (say), saving on fiber costs, an extra factor is required. The reduction in the number of major imaging cycles assumed the N antennas each contribute equally to reducing the PSF sidelobe levels. Grouped antennas act as a unit, whether their signals are summed or correlated individually, and this reduction will not then occur. In this case, the scaling law above must be multiplied by a further factor, N_{ag}, the number of antennas per station. If significant grouping is contemplated, this becomes a very significant factor.

4 Data Volume

Calculation of this factor is straightforwards. An array of N antennas, amongst which all possible correlations including autocorrelations are made for each of N_p polarizations, for each of N_c spectral channels, using an integration time of t_d seconds, produces data at a rate

$$\dot{V} = \frac{5N_cN_pN(N-1)}{t_d} \qquad \text{Bytes/sec} \tag{5}$$

where it is assumed each visibility is written as two four-byte numbers, with a 20% overhead for 'meta-data'.

Since we are assuming a constant-sensitivity array, the number of antennas scales as $N \propto 1/D^2$. To avoid chromatic aberrations ('bandwidth smearing'), it is necessary to employ narrow frequency channels. The number required is shown by Perley[4] to be $N_c = k(BWR-1)B/D$, where $BWR = \nu_u/\nu_l$ is the bandwidth ratio, and the coefficient k is between 1 and 10, depending on the degree of tolerable error[3]. In the same memo, Perley shows that the time averaging interval required to limit time-smearing aberrations to negligible levels is given by $t_d \sim D/(B\omega)$, where ω is the angular rotation rate of the earth. Combining these factors, we obtain, using an intermediate value of k,

$$\dot{V} \sim 0.01\left(\frac{NB}{D}\right)^2\left(\frac{W}{\nu}\right) \qquad \text{Bytes/sec} \tag{6}$$

where W is the bandwidth, and ν is the band frequency. The desired scaling laws are thus:

$$V \propto N^2B^2/D^2 \propto B^2/D^6 \propto B^2N^3 \tag{7}$$

where we have omitted the bandwidth and frequency factors, and a linear factor of time.

[3]Chromatic aberrations can be offset by deconvolution, if the aberration is not severe, so a modest error can be tolerated

5 Processing Costs

The above relationships can now be combined to give the overall scaling laws for computing, under the assumptions given before. We find

$$\text{Cost} \propto \lambda^{1.7} N_{ag} \frac{NB^3}{D^4} \propto \lambda^{1.7} N_{ag} \frac{B^3}{D^6} \tag{8}$$

where N_{ag} is the number of antennas in a station, B is the maximum baseline, and D is the antenna diameter. Alternatively, for a constant-area array, we find

$$\text{Cost} \propto \lambda^{1.7} N_{ag} (NB)^3 \tag{9}$$

How much worse is this than what we currently deal with? A lot worse. For the EVLA, with ten times the maximum baseline of the present VLA, and 37 antennas all of 25-meter diameter, the increase from this relation is by a factor of $\sim$ 2500. However, this formulation has assumed the bandwidths are the same – in fact the EVLA will have at least twenty times the bandwidth of the current VLA in L-band, so the actual enhancement factor over the present computing requirement is more like 50,000. Indeed, it is likely to be considerably higher than this, since the effect of extra sensitivity, requiring more major cycles, has been neglected. If Moore's Law gives a doubling of capacity every 18 months, it will take at least 20 years to return to current relative processing speeds for full-field imaging. Replacing the EVLA with (say) 4 times as many antennas of one half the diameter will multiply this factor by another $4^3 = 64$.

6 A Cost Equation

As shown above, the cost to process correlator data can be substantial. This is especially true for high sensitivity arrays operating in the continuum, where the entire element beam must be deconvolved to high dynamic range to avoid the confusion caused by the sidelobes of sources distant from the tracking center. This cost has not hitherto properly been taken into account in array design. Array design usually proceeds on the basis of a cost equation, which is used to examine the tradeoff between the number of elements and the size of each element. Below we derive cost equations containing this term.

First, consider the simple case, in which we are designing for a single band, with a given receiver, and we are given a resolution and a sensitivity, so the product ND^2 is constant. Let us parameterize the cost equation by the number of elements.

The cost equation can be written (retaining its dependencies on both antenna diameter D and number N):

$$C = a + bN + dND^{2\alpha} + eN^2 + fND^{-4}, \tag{10}$$

where α where

- a is the fixed cost irrespective of the number of antennas (building, power substation, water supply, internet hookup, etc.).
- b is the cost of the equipment on each antenna (receiver, fiber tranceivers, correlator station signal conditioner, etc).
- d determines the costs of the antennas. The exponent α is somewhere between 1.3 and 1.4.
- e is the 'per baseline' part of the correlator cost, including general purpose computers to clean up and format the data stream. (The correlator also has a 'per element' cost, included in the coefficient b.)

- f sets the cost of post-processing. (The factor B^{-3} has been omitted for now.)

For a constant-area array, this can be rewritten as

$$C = a + b'N + dN^{\eta} + eN^2 + f'N^3. \tag{11}$$

The exponent $\eta = 1 - \alpha$ is typically -0.3 to -0.4, so that the cost of the antennas for a constant-area array actually decreases as the number of antennas increases.

With the current VLA, A configuration, and current computers, the coefficient f' is of order \$10. For larger configurations, this coefficient will scale roughly with the cube of the baseline. With an 18 month exponential decrease in the cost of processing power, by 2018, f' might decrease to \$0.01. Even so, with a few thousand elements, the last term would be a substantial part of the budget.

For the WIDAR correlator, the coefficient e is of order of \$2000. With techniques adapted for a larger number of elements, and using the electronics of the next decade, it would be surprising if this coeficient did not decrease by a factor exceeding ten. This is not a dominating cost for most reasonable configurations.

A complete cost equation would include several other components: whether the receiver front-ends are cooled or uncooled, the bandwidth of the receivers, and the size of the array. Below is a first attempt at taking these factors into account.

We still presume we are designing to a given desired sensitivity. Sensitivity is proportional to

$$\sigma \propto \frac{T_{sys}}{ND^2\sqrt{W}} = \frac{cT_c}{ND^2\sqrt{W}} \tag{12}$$

and is now the quantity to be kept fixed. In these expressions,

- W is the bandwidth
- T_c is the system temperature of a cryogenic receiver
- $c > 1$ is the ratio of system temperature of uncooled and cryogenic receiver systems.

The cost equation is now written as:

$$C = a + (b_a + b_c + b_w W)N + dND^{2\alpha} + eWN^2 + fND^{-4}W\nu^{-2.7}B^3 \tag{13}$$

where we have re-inserted the dependencies on baseline and wavelength, and have broken the antenna equipment costs into three components:

- b_a is that part of the cost of an antenna system that is independent of the diameter of the antenna, the bandwidth of the receiver, and whether the receiver is cooled or uncooled.
- b_c is the cost of cryogenics for cooled receivers (zero for uncooled receivers).
- b_w is the cost of components proportional to bandwidth (principally the correlator antenna based signal processor and some components of the digital transmission system.

Using the constant-sensitivity relation to remove the dependence on antenna diameter, we find

$$C = a + (b_a + b_c + b_w W)N + d^{\star}N^{\eta}(cW^{-1/2})^{1-\eta} + eWN^2 + f^{\star}\nu^{-2.7}B^3Wc^{-2}N^3. \tag{14}$$

The W factor in the correlator cost occurs because in the usual cases, the clock rate of convenient digital circuitry is substantially lower than the bandwidth we wish to process, so multiple copies of the circuitry are run in parallel. The W factor in the data processing is due simply to the increased number of delay channels we must process to cover the entire element beam.

It is interesting to note that the cost difference between a system with cryogenic receivers and one with uncooled receivers is (assuming the other parameters are held fixed):

$$\Delta C = -b_c N - d^\star N^\eta W^{(\eta-1)/2}(c^{1-\eta} - 1) + f^\star \nu^{-2.7} W B^3 N^3 (1 - c^{-2}). \tag{15}$$

A negative value means the uncooled array is cheaper than a cooled array for a given sensitivity. The increased antenna cost required for uncooled systems is traded off against not only the cost of the cryogenics, but against the cost of the data processing, which is higher with the smaller antennas permitted by the cryogenic system.

These equations are for a system designed for imaging single fields. For a survey system, the figure of merit does not go as sensitivity ($\sim ND^2$), but as the simple product of number of elements and element diameter, (ND). This may lead to quite different conclusions.

In any event, it seems necessary to more carefully consider the data processing requirements for future arrays of high sensitivity and long baselines before embarking on their final design. It is possible, but by no means certain, that more economic algorithms than those considered above will yield sufficient dynamic range.

References

[1] Clark, B.G. VLA Scientific Memo #107. 1973

[2] Perley, R.A., 'Imaging with Non-Coplanar Arrays', in 'Synthesis Imaging in Radio Astronomy', ASP Conference Series, Vol 180., 1999.

[3] Cornwell, T.J., and Perley, R.A., 'Radio-Interferometric Imaging of Very Large Fields.' Astron.&Astrophys. Vol 261, 353-364, 1992.

[4] Perley, R.A. EVLA Memo #64, 2003.