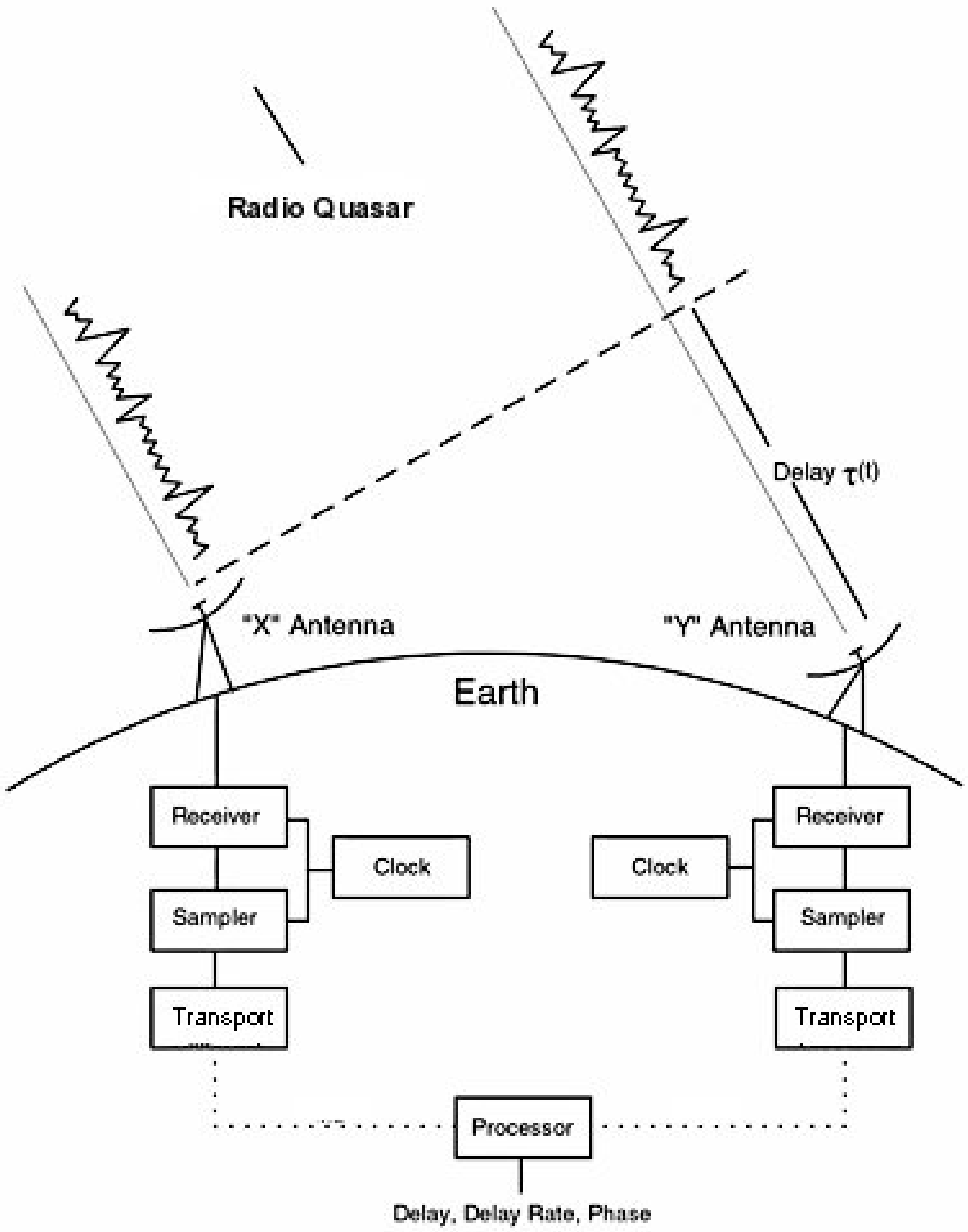
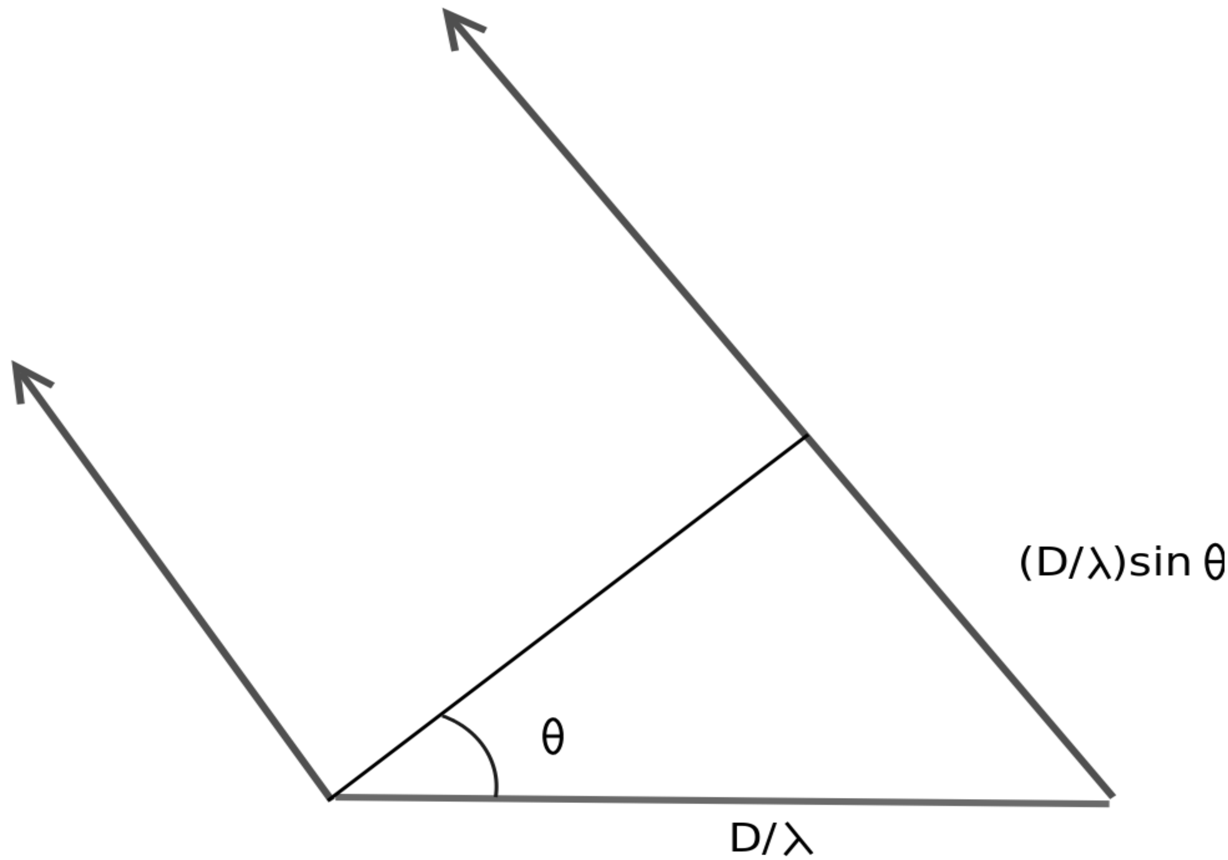
VLBI Basics:
From Station through Correlator

The Basic Diagram of VLBI:



We will follow this diagram from the Radio Quasar to the Observables coming out of the Processor. First, the basic quantity that we wish to observe is the Delay τ. Once we know the value of the delay, we can solve for the triangle formed by the baseline between the two stations, the direction to the Quasar and the angle between the baseline and the Quasar direction.

The Basic diagram in more schematic form:



We've replaced the time delay with the geometric distance (D/λ)sin θ.

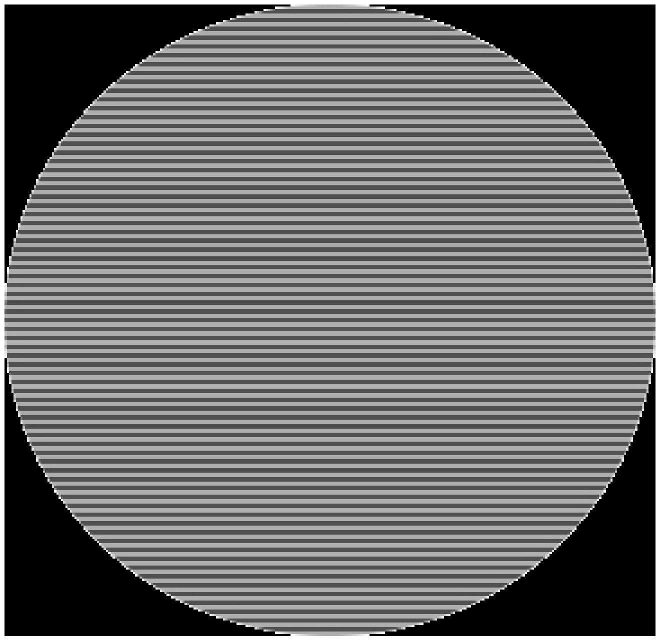
As the source moves due to the rotation of the Earth, the interferometric response (called the "Visibility" V) changes as :

$$V = \cos(2\pi(\frac{D}{\lambda})\sin(\theta))$$

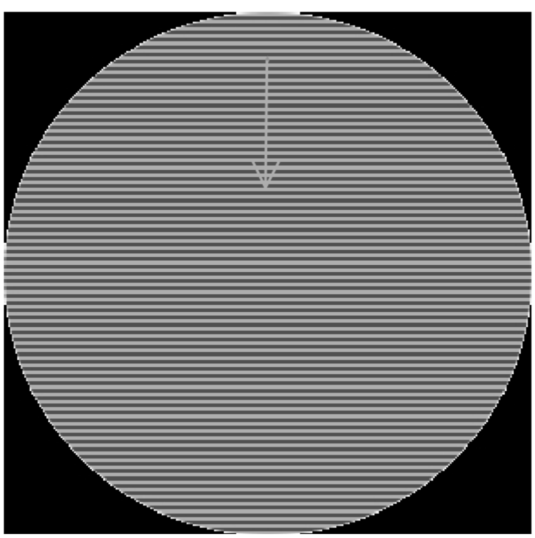
For small angles (such as when we're pointed perpendicular to the baseline D/λ) this reduces to:

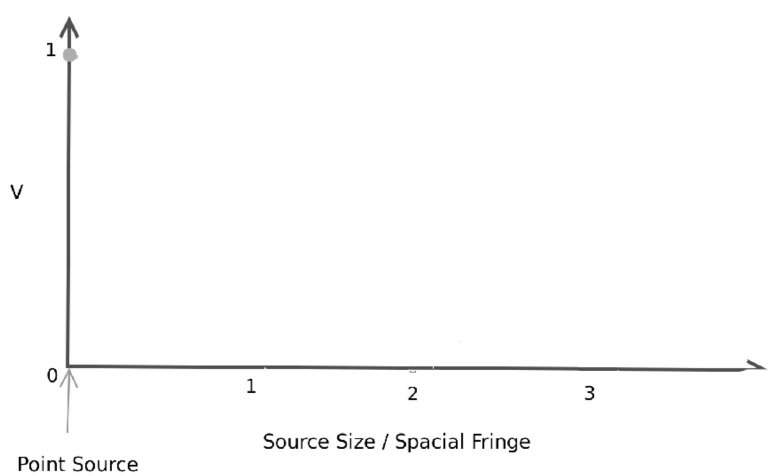
$$V = \cos(2\pi(\frac{D}{\lambda})\theta)$$

This means we get a cosine pattern projected on the sky:



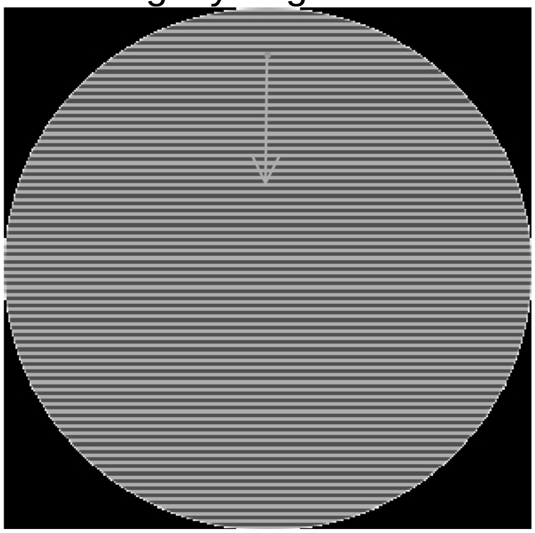
If we have a point source, and the baseline and equator are lined up vertically, a point source moves through the pattern:

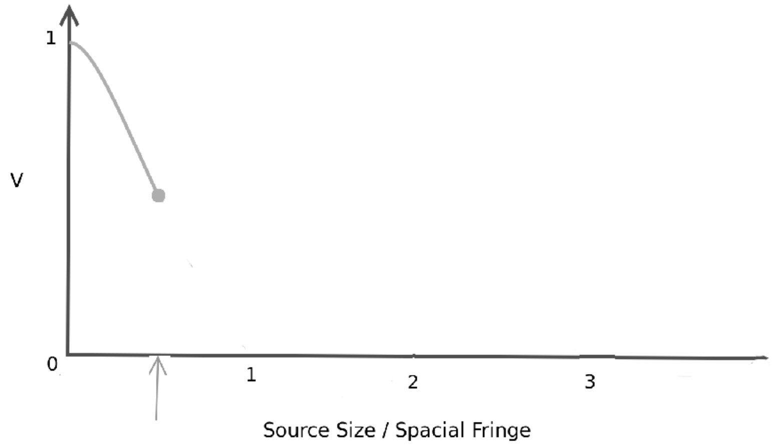




On the right we see that, because the point perfectly reproduces the cosine pattern, the Visibility is maximum and equal to the pattern amplitude (assuming we call the source strength 1).

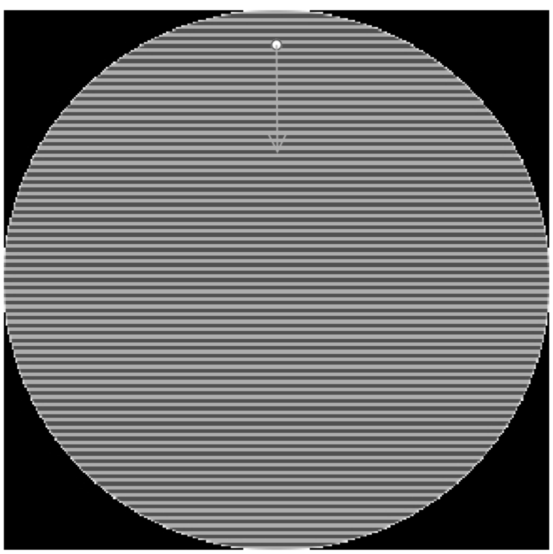
For a slightly larger source:

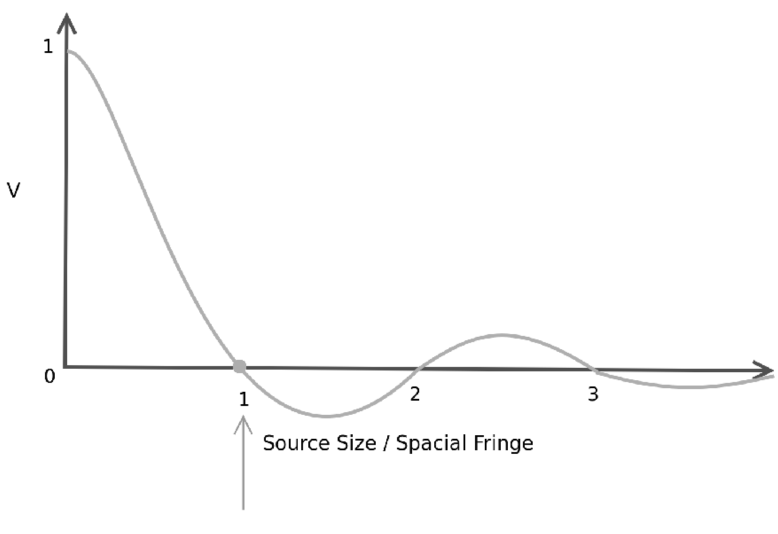




The source is smeared over more than a little of the cosine pattern and the negative part of the cosine function (dark) starts to subtract from the positive part of the cosine function (light) and the overall amplitude decreases due to the interference.

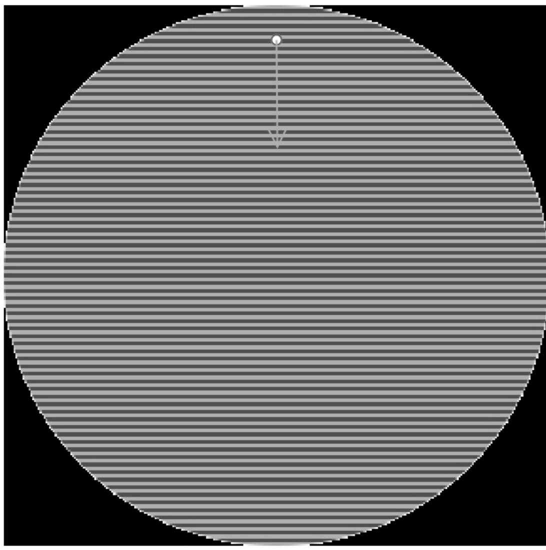
If the source is one fringe in size:

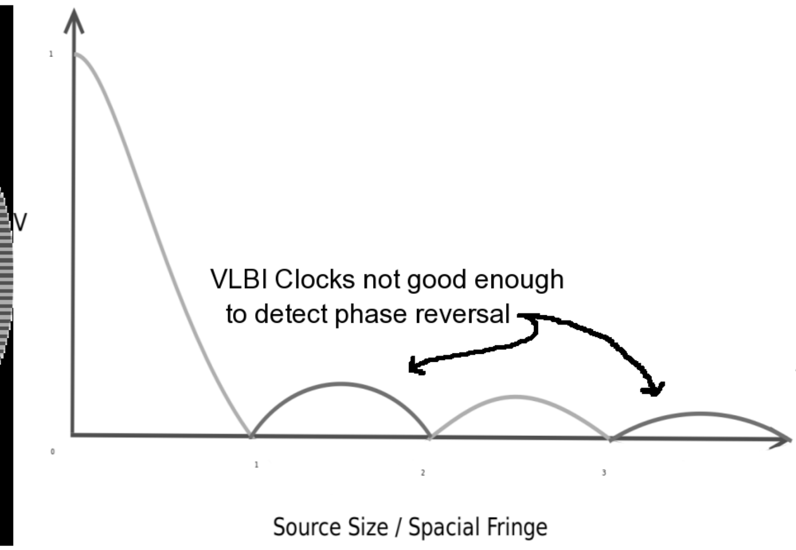




Here we see that for a source that is one fringe in size, the positive part of the cosine response is always equal to the negative part and the visibility V is always equal to 0! A source one fringe in size is invisible to our interferometer! Note that for a source larger than a fringe, there is a reduced response since the full-fringe part of the source cancels and only that part greater than a fringe has a response. Notice also that the visibility continues to decrease and will eventually be lost in the noise, and that the phase of the visibility reverses every fringe size.

Due to the fact that we use imperfect, separate clocks at each station, we cannot recover the phase. Thus the VLBI interferometer response is as follows:





If you measure a single source's visibility from many baselines in different directions, using the difference in baseline length (the fringe is proportional to λ/D: wavelength divided by baseline length) to measure the difference in source response in different size-scales and directions, you can derive an image. This is how astronomical VLBI works.

For geodetic VLBI we are mainly interested in the fact that we require a nearly point source so we don't have to worry about the source disappearing at the longest baselines, or having a large difference in response depending on which baseline we're observing with.

For measuring our baselines, we need a reference frame that is stable, so we also need sources that don't move. To summarize:

We need bright, nearly point-like sources that don't move!

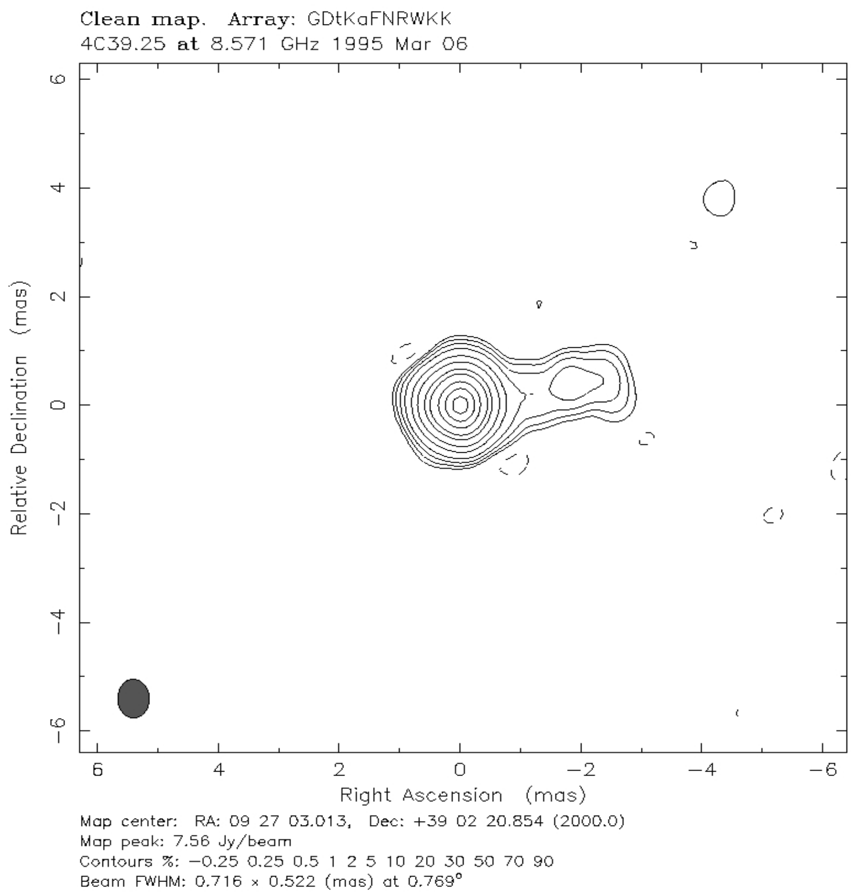
Quasars are among the brightest radio sources in the sky. They are also very far away, which means their apparent positions won't move for a very long time, even if their intrinsic velocity is very high. (Some quasars change their shape, though, so not all quasars are suitable. Some also have large scale structure and aren't point-like).

Some Quasars are compact.
Some Quasars are stable in brightness.
Some Quasars don't move.

Very few are all three. Less than 1000 over the whole sky.

The strange case of 4C39.25

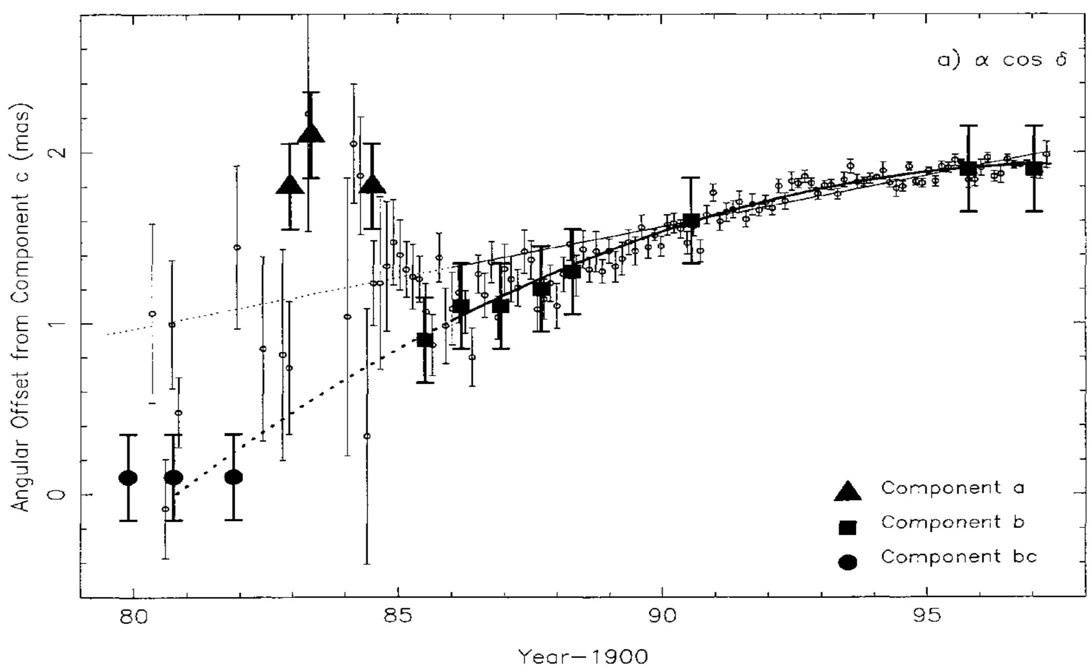
4C39.25 is a source we actually use in geodetic VLBI. An image made from geodetic observations is shown below:



Notice that there is a bright, compact component on the left (this is source "b"). There is a less compact, and less bright component on the right (this is source “c”). Source “a” is hidden under the left side of component “b”.

This image was made in 1995. In the 1980, the 4C39.25 was dominated by components a and c. At this time, the position of the source was complicated by the length and direction of the baseline being used. Since there we few baselines in the geodetic sessions, confusion was maximum!

By 1990, component b had appeared and began to move from component c towards component a. Component b remained compact and grew brighter. By 1995, the astrometric position for this source had “locked-on” to moving component b! In the late 1990s and early 2000s, component b also came to a stop near component a. It is now one of our brightest sources, point-like and moving very slowly. However, it illustrates the problems with non-point (and moving!) sources:



In the plot, image measurements are the bold, solid points and the small open circles with error-bars are astrometric measurements from geodetic sessions. Before 1985, the astrometric measurements are more or less in between the a and c components. Between 1985 and 1988 the astrometric measurements are between components a and b (and moving toward b). From about 1988 on, the position was following component b as it completed its apparent journey from near component c to near component a. The source has been only slowly moving since 1997.

This plot demonstrates the confusion that can result by using an extended source with changing structure.

We've talked about the basic VLBI geometry and a little about the sources we observe, now let's look at what we require at a station.

Station Requirements

When we point a radio telescope at the sky, we have the following noise sources in our beam:

Quasar noise (our signal!)
3° K Cosmic Background
Background and foreground sources in the beam
Interstellar medium
Ionosphere
Thermal noise from atmosphere and ground
other sources (RFI, etc)
Noise generated in the Observing System

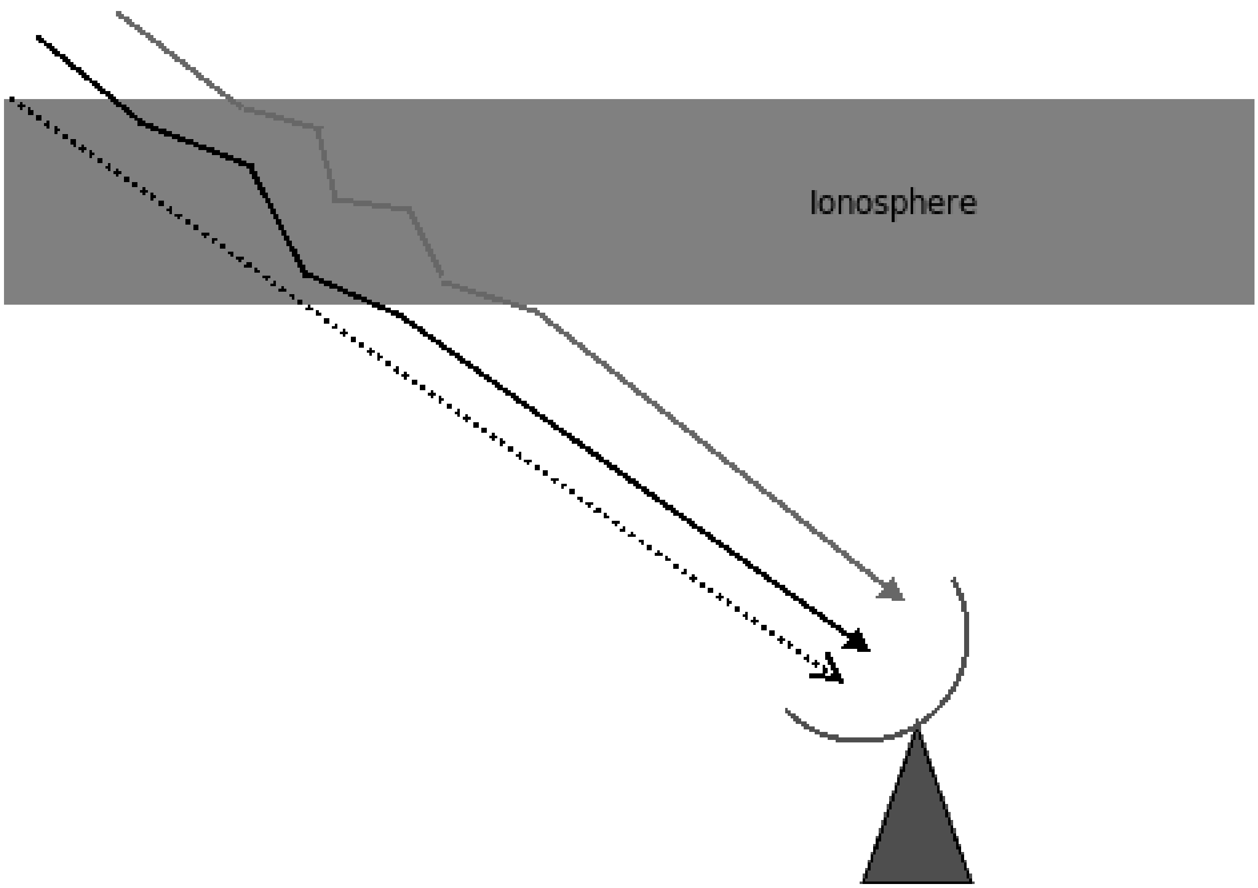
We are trying to measure a Delay which is a time measurement. This depends on our time resolution and our Signal to Noise Radio (SNR).

Time Resolution $\propto \quad \frac{1}{(spanned\ Bandwidth)}$

This is the total bandwidth spanned by our observation (not our recorded bandwidth!). So for better single measurement time resolution we require a wide band feed and receiver.

We also have an observing problem: the Ionosphere adds a delay as the signals pass through. This is dependent on the electron content of the ionosphere and varies with the observing elevation, the time of day and the state of the sunspot cycle.

We can calibrate this additional delay by using the fact that the delay through the ionosphere is dependent on frequency. The higher the frequency, the less the additional delay.



This is why we use X-band and S-band at the present time, and will use a wide range of frequencies in VLBI2010. Using the differential delay at the various frequency bands we can reconstruct the geometrical delay.

A second requirement is our SNR. The first part of SNR is the signal, which comes from the Quasar and is fixed. The second part is our local noise. The stations measure this when they determine the System Effective Flux Density (SEFD). The SEFD is a measure of the total noise from the entire observing system at the station in Janskys (10^{-26} W/m^2Hz).

SEFD $\propto \dfrac{T_{sys}}{(\eta_a * A)}$

Where:

T_{sys} = System Temperature

η_a = Antenna efficiency

A = Antenna Area

So we can see that we need a low System Temperature, which includes a low noise receiver, and a large, efficient Antenna to lower the SEFD. In geodesy there are conflicting requirements in that there also is a need for fast antennas, and big fast antennas are expensive, so compromises have to be made. Antenna efficiency includes such things as obstructions due to sub-reflectors and quadrapods as well as the smoothness of the surface.

The flux density of the typical geodetic source is on the order of 1 Jansky. The SEFDs of our typical station is on the order of 1000 Jy (they range from ~300 to ~3000). So the initial single sample SNR is on the order of 1/1000!

In VLBI:

$$\text{SNR} \propto \frac{S*\sqrt{N}}{\sqrt{SEFD_1 * SEFD_2}}$$

Where:

S = Source Flux Density

N = Total number of Samples

$SEFD_i$ = SEFD of the i^{th} station

$N = 2 * \Delta\nu * T$

Where $\Delta\nu$ is the bandwidth and T is the integration time

To increase the SNR we need to increase the recording rate:

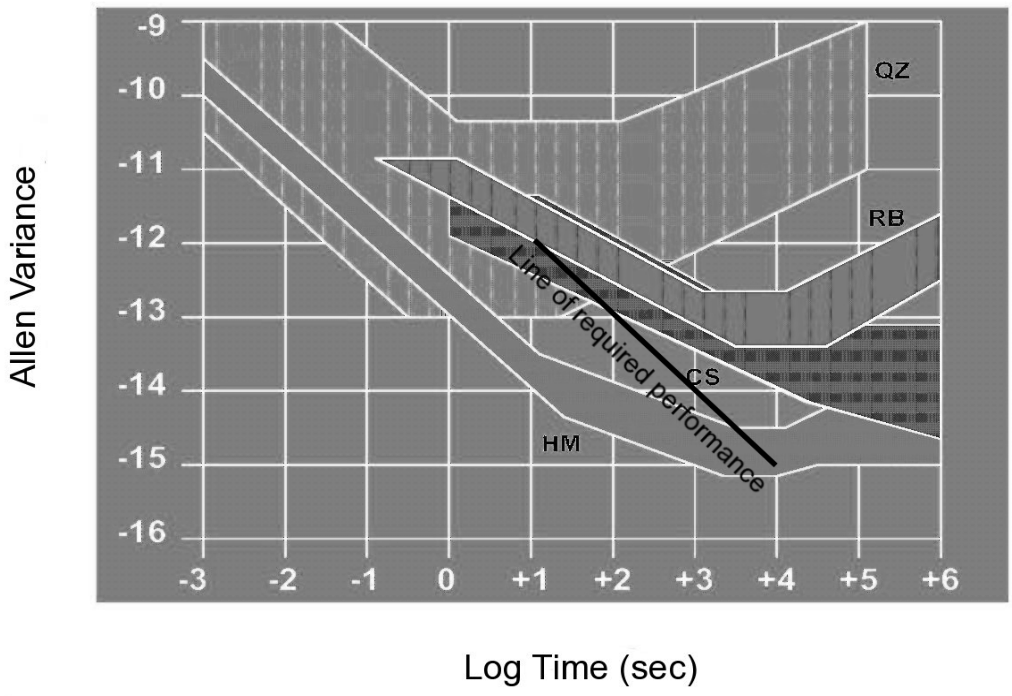
Bandwidth (Hz) = ½ * (samples/sec)

This is called Nyquist sampling

In VLBI we have another problem: We rely on a local clock to control our Local Oscillator since we cannot distribute a local oscillator signal with enough accuracy to maintain coherence at our widely separated stations (maybe someday we will be able to).

So we need a clock which can maintain timing between stations good enough to keep our offset to less than ~ 1 radian. One radian at 10GHz is $1.6 * 10^{-11}$ seconds. If we want to keep this accuracy over 1000 seconds, this requires a clock good to ~ 1 part in 10^{14} .

The following plot shows the behaviors of available clocks:



QZ is a high-performance Quartz Oscillator
RB is a rubidium standard
CS is a cesium standard
HM is a Hydrogen maser

The line of required performance for our application is indicated. Only a Hydrogen maser has enough of a margin to guarantee coherence over the whole range.

Finally, in order to calibrate the delays through the observing system, we need a delay calibration system. This consists of two parts: a cable calibrator which removes the delay from maser to LO at the receiver by measuring the time delay through the cable, and a phase calibrator which removes any frequency dependent delay through the signal path.

To summarize:

Station Requirements:

Wide-band feed and receiver
Low noise receiver

Large, efficient dish
High speed recording and/or transport system
Hydrogen Maser clock
Multiple frequency bands
Delay calibration system
    Cable cal
    Phase cal

Combining the Observations into a Single Array

We have incredibly weak sources being observed by 1000X more noisy observing systems

We have limited ability to expand the bandwidth (sampler/recorder limitations)

We have limited integration times (recorder limits and clock behavior)

How do we combine these observations into a single, Earth—sized array?

Mathematical Magic!

In fact, two types of mathematical magic:

Fourier Transforms and Cross-correlation

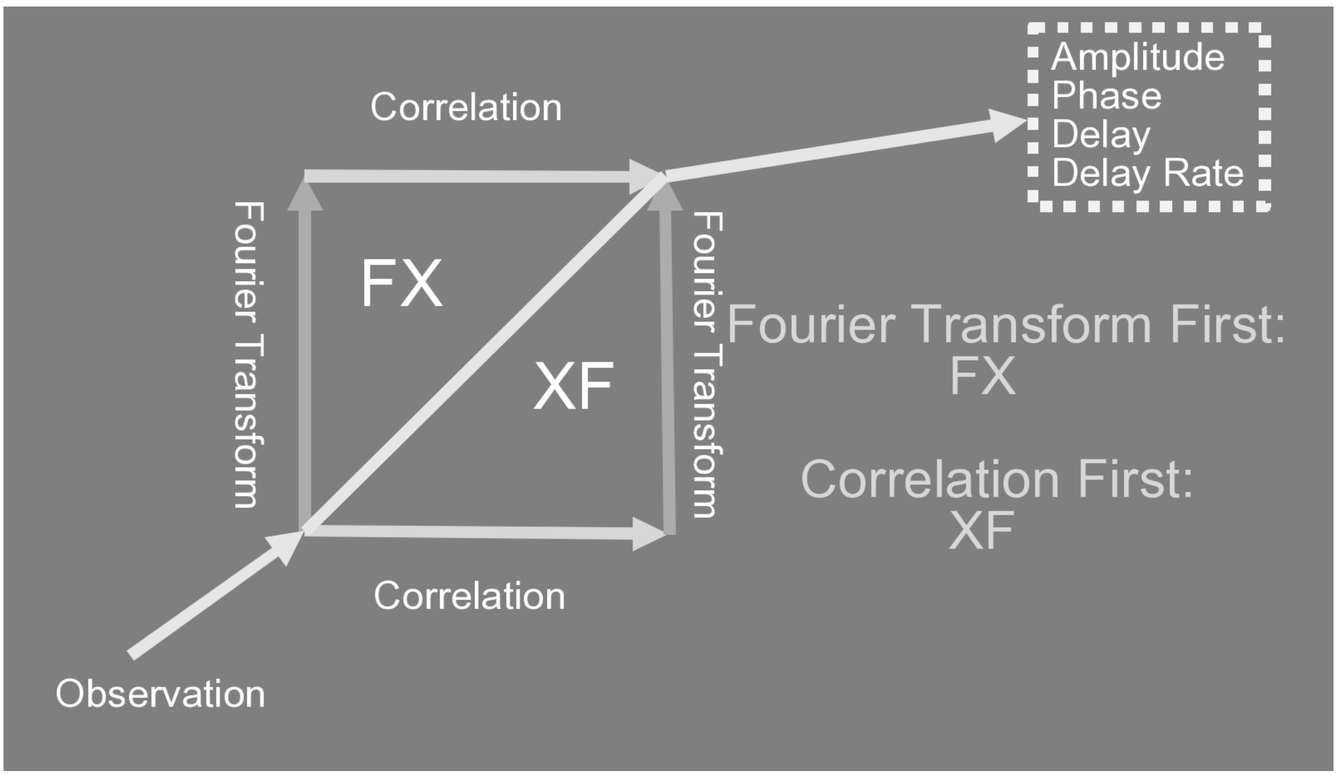
Fourier Transforms allow us to extract frequency information from data. They are the equivalent of an efficient least-squares fitting of sine waves to our data. As such, they allow us to not only pull frequency information out of our data, but allow us to do some least squares fitting if needed.

Fast Fourier Transforms are a realization of the Fourier Transform that is very fast, and easily implemented on computers. FFT techniques have been available since the 1940s.

Cross-correlation is a mathematical technique that is very powerful in pulling weak signals out of strong background noise. This technique is used by Global Navigation Satellite Systems (ie. GPS) and CDMA cell phones.

There are two flavors of processors used to combine the signals from geographically distant VLBI stations:

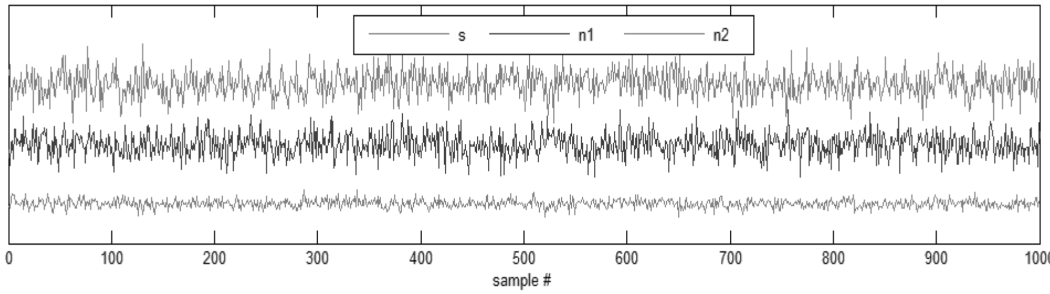
FX and XF



They both use FFTs and correlation, but they perform the tasks in reverse order. FX correlators include the VLBA hardware correlator and the DiFX software correlator, XF correlators include the Mark 4 and EVLA hardware correlators.

Why do we call them "Correlators"

If we had a high SNR, we could just measure the arrival times of the noise pulse from the quasars at the two stations. But our SNR is typically 10^{-3}, so this doesn't work. We need to pull the signal out of the noise and correlation is how we do it.



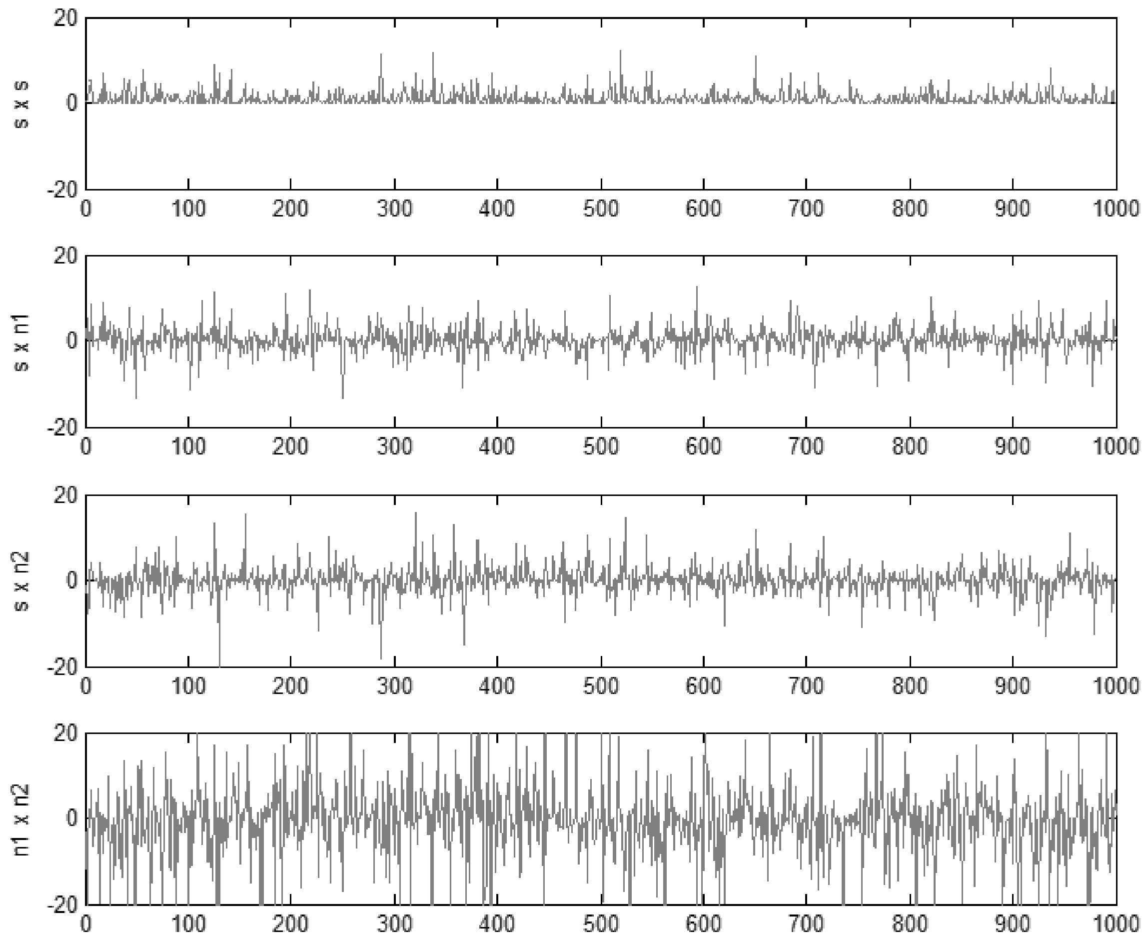
In this plot we show two noise “signals” n1 and n2, and a lower amplitude signal s (bottom). The signals reaching the correlator are of course:

s + n1 and s + n2

if we multiply these signals together we get:

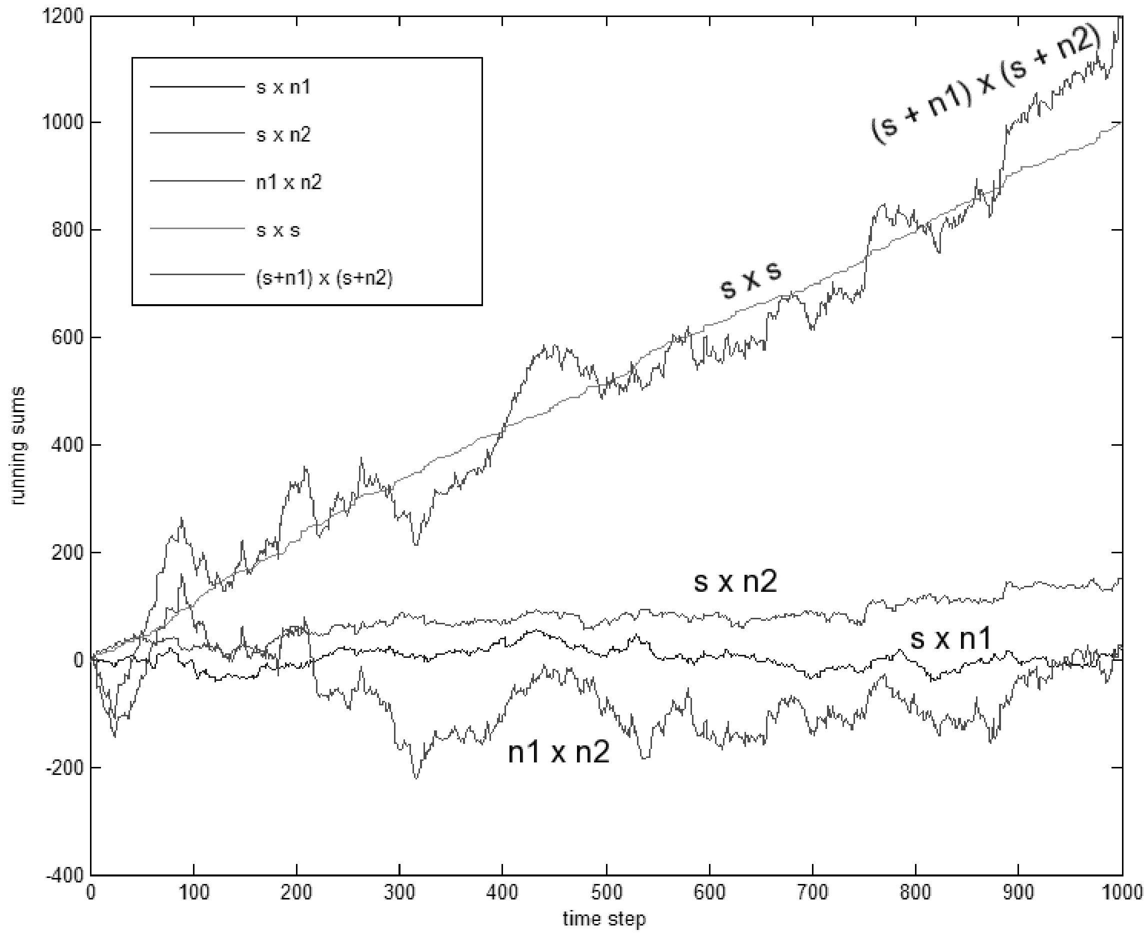
$(s + n1) * (s + n2) = s^2 + n1 * s + n2 * s + n1 * n2$

These components are shown below:



The secret of correlation is the fact that the s^2 is always positive (positive number X positive number = positive number, negative number X negative number = positive number) while the noise components multiplied by each other or by the signal will have negative components. You can see this above: the s x s plot has only positive going spikes while the rest of the plots have both positive and negative spikes.

If we accumulate a running sum of the various components and the (s + n1) * (s + n2) combinations we get the following plot:



As you can see, the (s x s) is clearly different from the random-walks of the combinations with the uncorrelated noise. The signal that we are actually dealing with (s + n1) x (s + n2), follows the (s x s)slope, and the slope of the (s x s) line can be fairly accurately estimated.

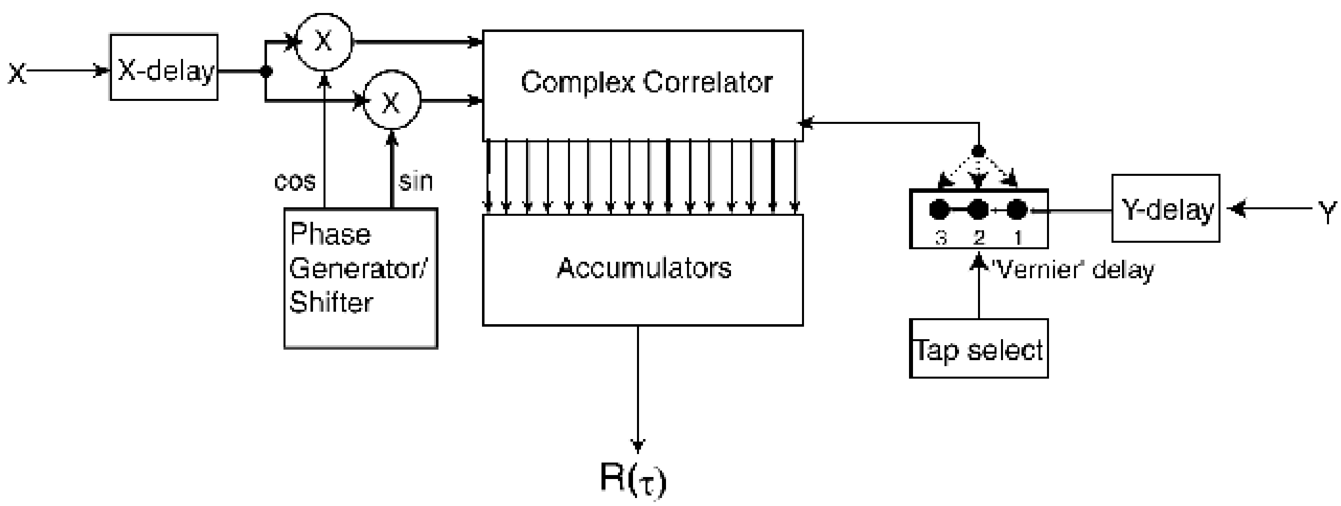
This is the power of cross-correlation: weak signals buried in noise will betray themselves in the summed output.

Correlation Details

There are a lot of small details that have to be taken care of before we will get successful correlation of VLBI observations. In the following, I will use the Mark 4 VLBI Correlators (which are XF correlators) as an example. The same processes will be applied to all correlators, but will differ in

details.
If we were correlating at RF frequencies, the models would account for all geometric effects such as the Doppler Shift. However, we correlate at baseband frequencies (after the RF has been mixed with the Local Oscillator) so we need to have separate delay and phase models.

We shift the phase of the observed bit stream by a model generated by the a priori geometry. This generates separate sine and cosine terms:

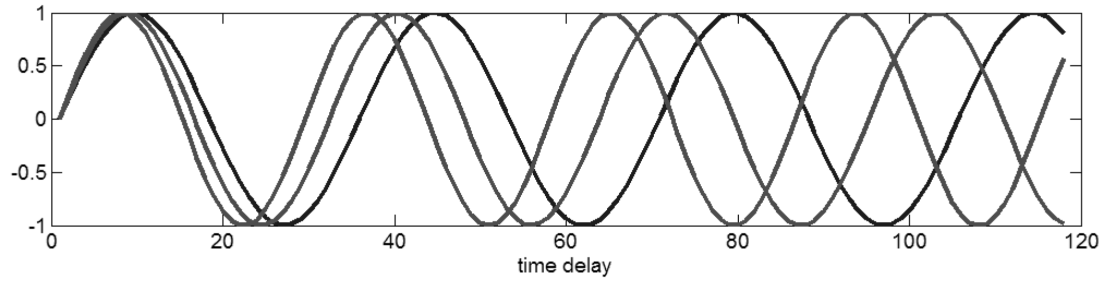


The sine and cosine terms are then correlated in a complex correlator section (because we must correlate both the sine and cosine terms) and the resulting amplitude and phase for each lag (typically 32 lags are done in a Mark 4 geodesy session) is obtained. This is performed for each channel observed.

Obtaining a fringe in each channel is not the final step. If we used the results for a single channel or even side-by-side channels, our delay resolution would not match the capabilities of our wide-bandwidth feeds and receivers. In order to improve the delay resolution we use:

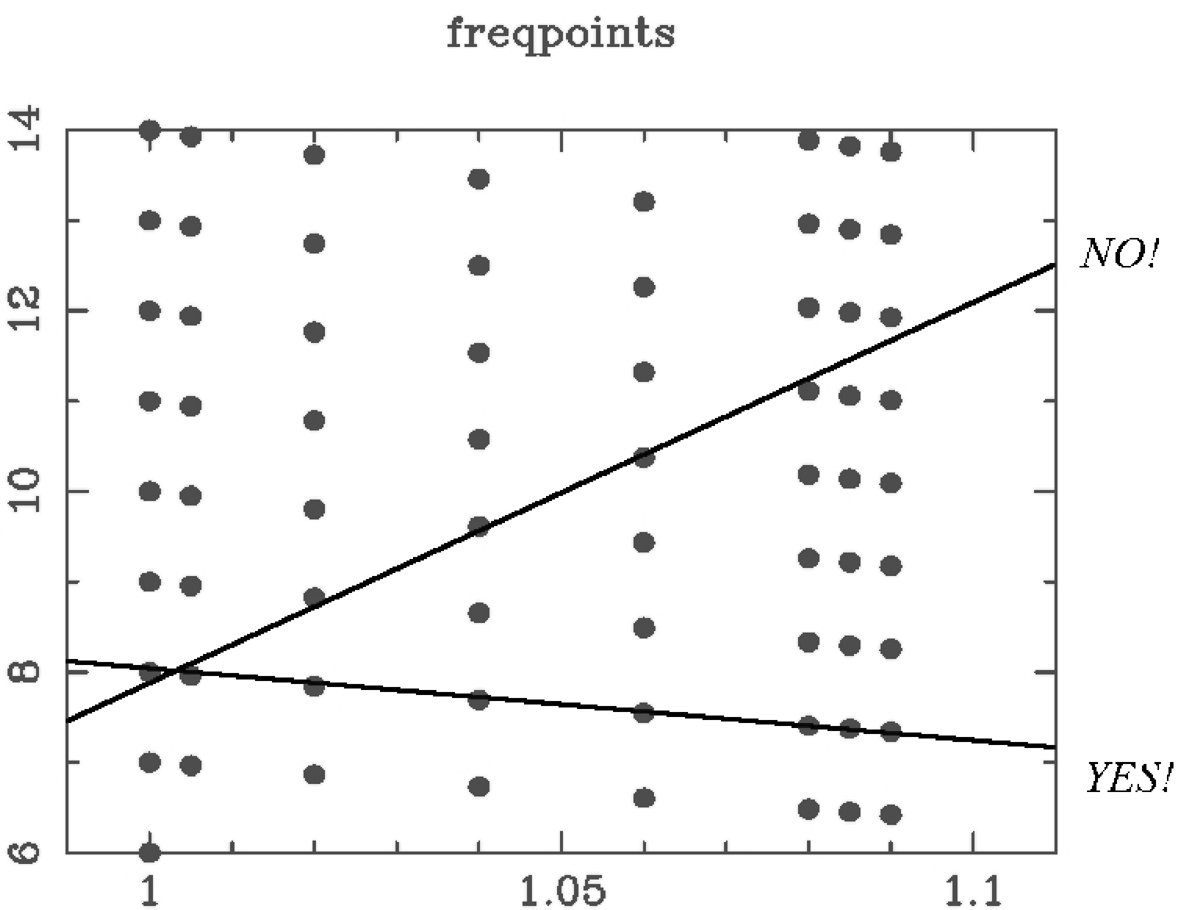
Frequency Bandwidth Synthesis

We are able to expand our resolution beyond our recorded bandwidth by spreading our channels out over the whole bandpass of our observing system. In the following diagram, note that the various frequencies all start at the same phase, but over the plot, they do not all align again. In fact, only the two highest frequencies get to 180° phase offset.



All three frequencies will only align at a delay that should be large enough different from the initial point that we can reject it.

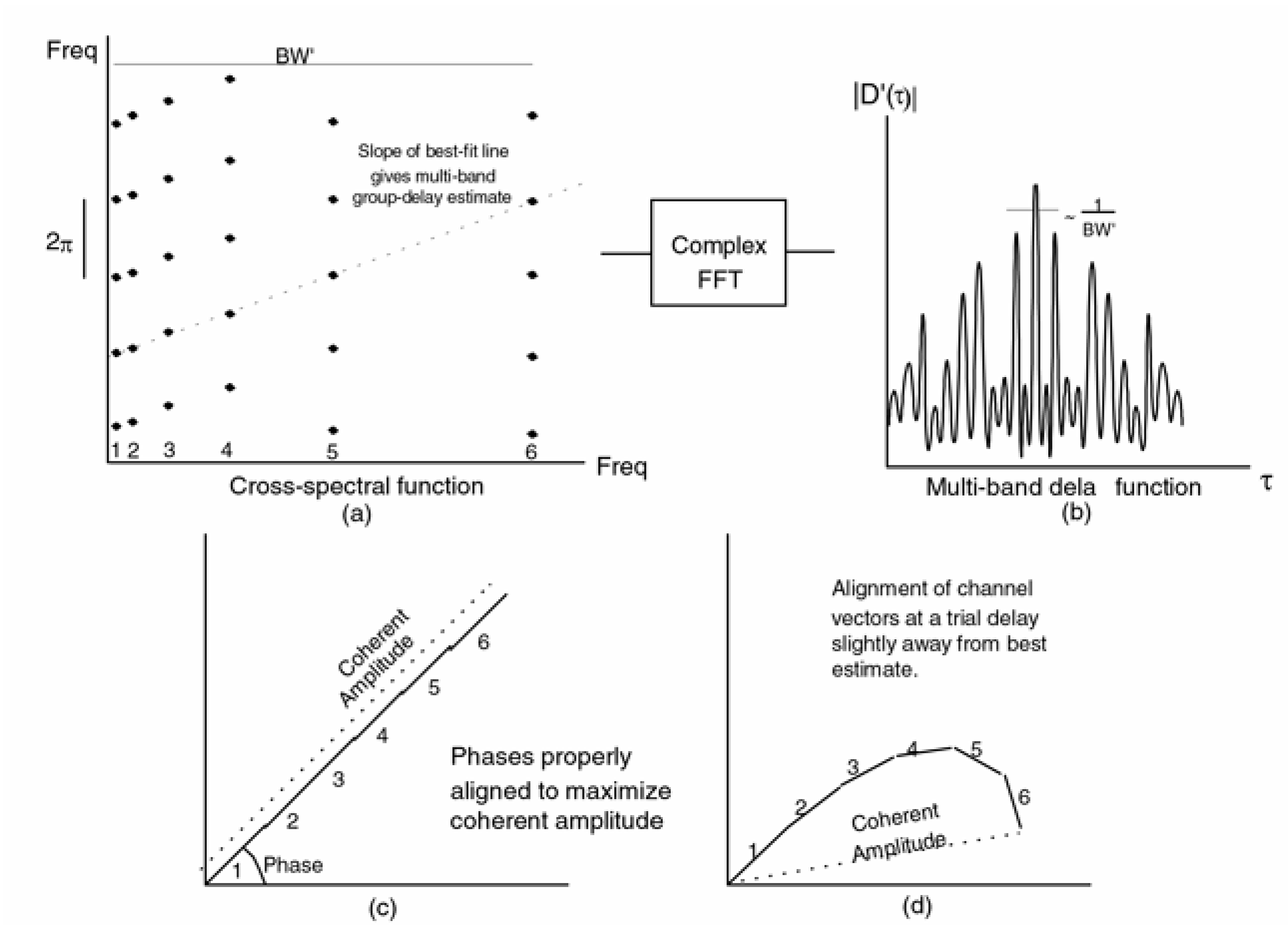
In practice we use more frequencies (6 in S-Band and 8 in X-band in the present sequences) to reduce the chance of an ambiguous delay as much as possible:



Note that the slope of the matching line will provide the delay and that the next possible match would be a very inclined line and thus a very different delay.

Losing channels reduces the ability to properly reject ambiguous fits, and can lead to erroneous results that have to be caught and fixed (if possible) before a final result can be determined.

We actually do the above fitting using FFTs :



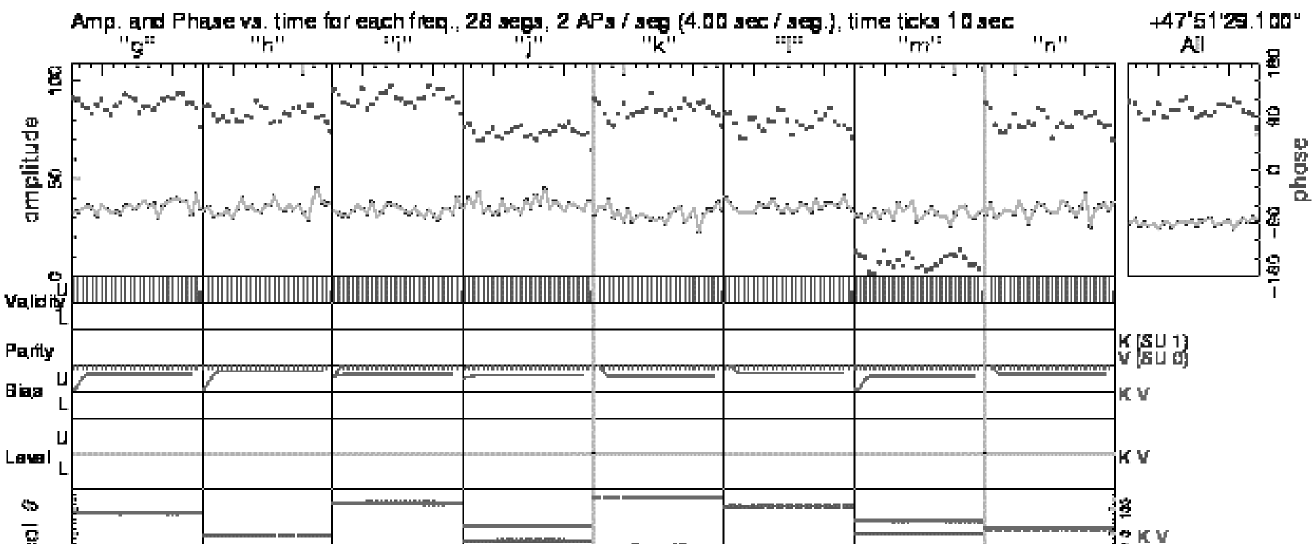
Note in (b) that the delay resolution function peak's width is proportional to the inverse of the spanned bandwidth of the observations. The wider the spanned bandwidth the narrower (better) the resolution. We hope that our observing frequency sequence gives better suppression of the sidelobes than show here!

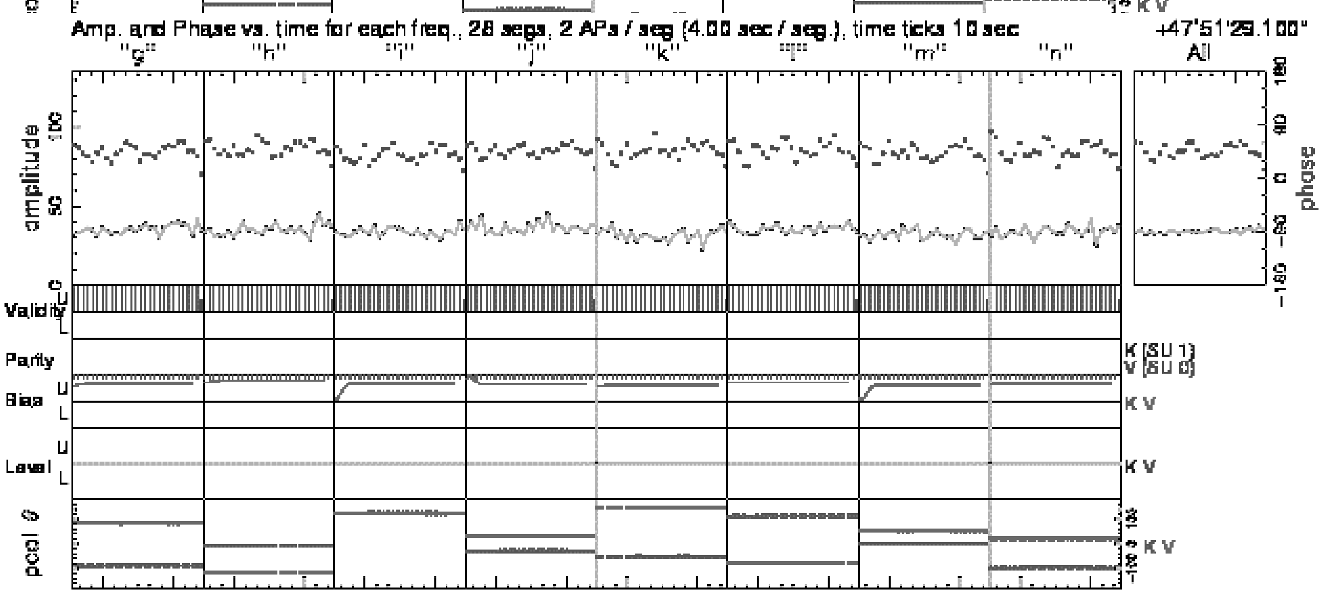
The (c) and (d) sections show an alternate way of thinking about frequency bandwidth synthesis, as an alignment of the phase vectors in the various channels. If we choose the delay correctly, we align all of the phases and we get a maximum amplitude as in (c), if we are not quite correct, the phases do not align an the derived amplitude is somewhat less as in (d).

Optimizing Coherence

At this point, we're almost done. One of the final steps is to correct for any frequency dependent delay differences caused by the observing system. The feeds, receivers, BBCs, cables, etc. can each contribute a linear change of phase across our observing bandwidth. For example, each BBC's filters could have slightly different delays. To correct for this, as we have already seen, we inject a series of phase calibration tones (one each MHz) into the front end. These tones have the same phase at the start of each second.

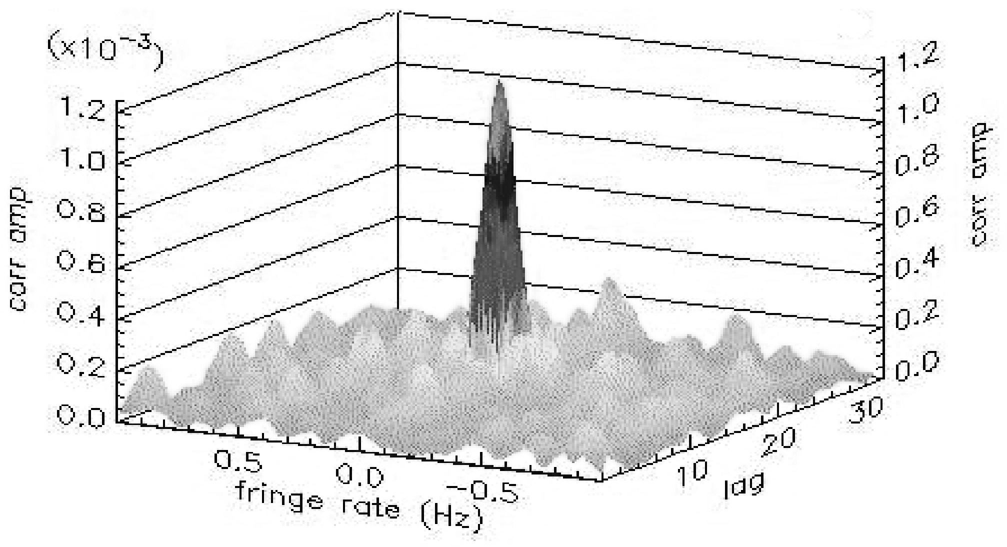
The correlator detects these tones, measures the phases and applies a correction to each channel:





Notice that the upper trace in the two plots is not aligned in the first plot and is aligned in the second... this is the result of applying the phase cal.

At this point, we should have fringes:



And, more importantly, our observables:

Amplitude
Phase
Delay
Delay-rate

If everything has worked, we have a successful VLBI Observation and the data is available to the investigators to use for Reference Frames, Earth Orientation, Crustal Motion, imaging, etc.