S Parameter Causality: A Sampled Data Perspective

Mike Steinberger
Todd Westerhoff

1.0 It's a Continous World. Or is it?

The causality of S parameters has been getting a lot of attention because non-causal S parameters can cause SPICE simulations to become unstable. This attention gets heightened because, at least at first glance, causality seems to be a mysterious property that requires more than a casual inspection of the data.

Causality is the property whereby a system only produces a response after it has received a stimulus and not before. In other words, a causal system is in no way clairvoyant. It makes sense that a non-causal system would be difficult to simulate in the time domain because in a very real sense, the system would be at least partially trying to run time backwards.

By considering S parameter data from a sampled data perspective, this white paper offers additional insight into the causality, or perceived causality of S parameter data. In particular, it presents the following observations:

1. The frequency spacing of the S parameter data can affect the apparent causality of the data. Closer frequency spacing will in general be better. The maximum acceptable frequency spacing is determined by the delay and rise/fall time of the network being characterized.
2. The maximum frequency of the S parameter data can affect the apparent causality of the data. A greater maximum frequency will in general be better. There's no sense in going to extremes, however, because the maximum frequency effects are impossible to remove without some (relatively simple) data manipulation. It's enough to have data beyond the highest frequency that is relevant to the system's performance.
3. Any continuous time model, even one that is theoretically sound and mathematically precise, will appear to be at least a little non-causal when expressed as sampled data.

The most common approach to this subject has been through the Kramers-Kronig relation, and there are several good explanations of this point of view [1], [2], [3]. While this point of view is certainly valid and useful, it isn't very intuitive for most engineers, and therefore doesn't help engineers exercise engineering judgement as well as one might wish.

At least part of the difficulty in applying the Kramers-Kronig relation lies in the assumptions underlying that relation. While there is a sampled data version of the Kramers-Kronig relation, most treatments of the subject, including those cited above, are based on continuous time mathematics. That is, the waveforms involved are assumed to be continuous functions of time. Functions in the frequency domain are assumed to be continuous as well and, by virtue of the continuity in the time domain, are assumed to extend from minus infinity to infinity. Even for the sampled data version of the Kramers-Kronig relation, there are critical assumptions which need not be valid, as will be explained later in this white paper.

The trouble is that the S parameters used in engineering are almost always supplied as a table of values at some number of discrete frequencies and not as a set of continuous frequency domain equations. That's how the data comes from a Vector Network Analyzer (VNA). That's how the data comes from a circuit solver or field solver computer program.

One way to address the sampled nature of the original S parameter data is to produce continuous frequency domain equations which approximate the original data, thus once more making the continuous time mathematics directly applicable. For example, one can produce a Rational Compact Model (RCM) [1] in which the original data is approximated as a set of rational transfer functions. While this can be a useful approach, one must understand clearly how an RCM will be used in subsequent analysis. For example, computer simulations are all sampled time simulations, so there must be some conversion from the RCM's continuous time domain back to the simulation's discrete time domain. There are several known methods for this conversion, each with its own advantages and disadvantages. There will necessarily be some numerical artifacts associated with that conversion, so it's important to understand which method was used, what its artifacts are, and how they affect the results.

Another approach is to address the S parameter data directly using sampled data techniques. That is, the data is supplied as a set of discrete samples in the frequency domain, so use the mathematics that was developed to handle such a data set directly.

Sampled data techniques generally assume that the samples are uniformly spaced in the frequency domain (as well as the time domain). In most cases, the original data is supplied with uniform frequency spacing, making this assumption immediately valid and the sampled data mathematics completely rigorous. However, even when the frequency spacing of the original data is non-uniform, some insight can be gained by approximating it with a uniform frequency spacing.

2.0 Sampled Data Facts of Life

2.1 A Simple Example

The demonstration case we will use is the simplest of rational transfer functions. In the continuous time domain,

$$H(s) = \frac{1}{1 + j\frac{\omega}{s_0}} \qquad \text{(EQ 1)}$$

Pole is at $-s_0$.

$$h_c(t) = s_0 e^{-s_0 t} \qquad \text{(EQ 2)}$$

This transfer function could, for example, be the S21 or S12 entry in the S parameter matrix for a shunt capacitor. It is well known to be 100% causal in the continuous time and frequency domains, so if there is any non-causality in the sampled time/frequency domain, it must somehow be related to the difference between the continuous and sampled domains.

Suppose that in a spreadsheet we choose $-s_0 = 3x10^9$ (C=13pF) and compute 1024 samples of this transfer function in approximately 100MHz steps (corresponding time domain sample interval 10pS). The gain vs. frequency for the result is plotted in Figure 1.

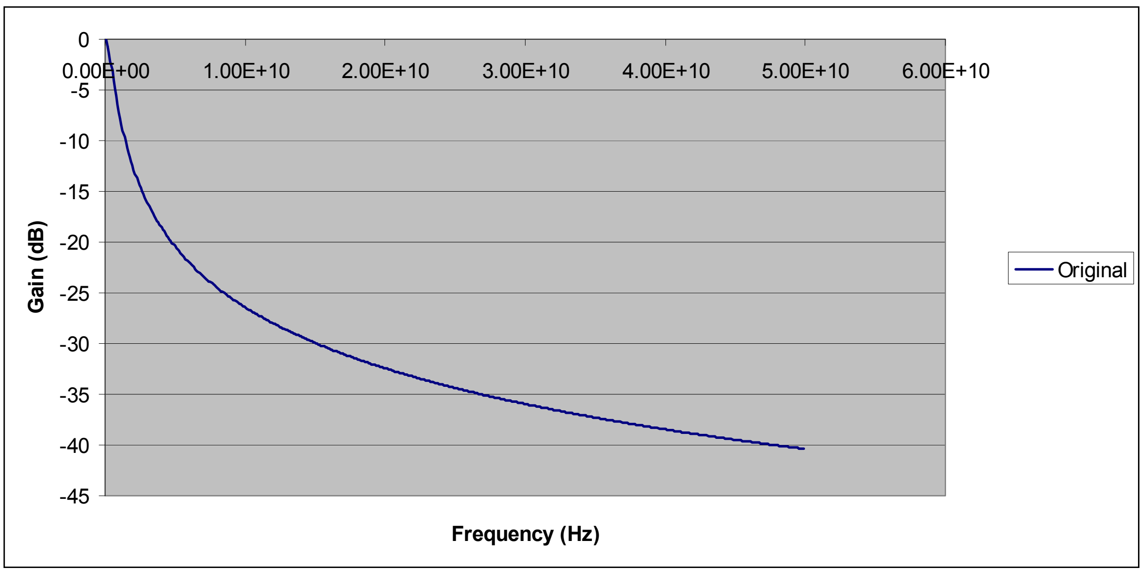

FIGURE 1. Single pole gain vs. frequency

The frequency has been taken far enough out to get to 40dB loss, so that should be more than adequate for most communications analyses.

If in this spreadsheet we now apply the inverse discrete Fourier transform (Tools->Data Analysis->Fourier Analysis), the sampled time domain result is as shown in Figure 2.

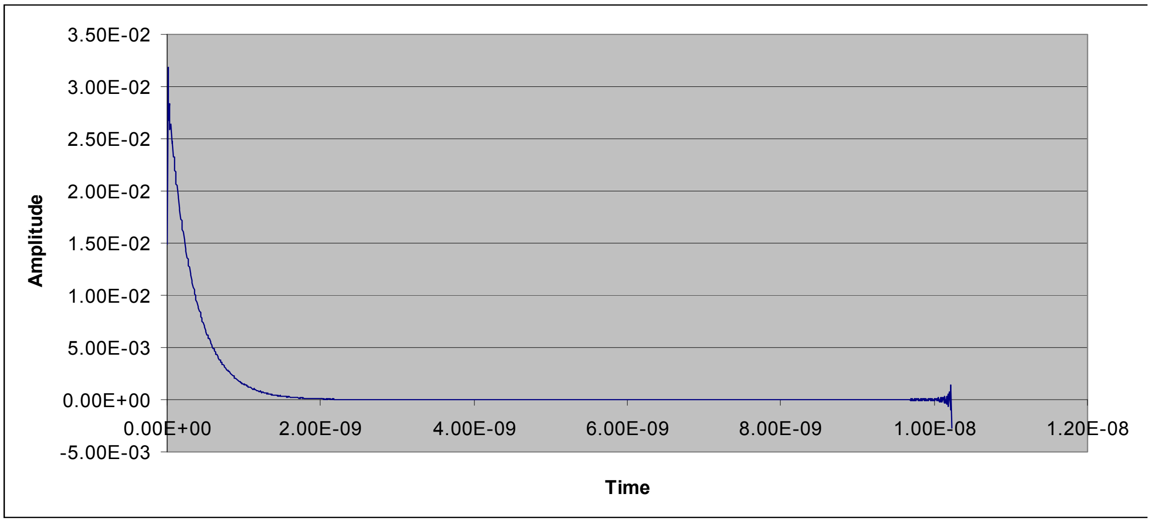

FIGURE 2. Single pole sampled time domain respone

Note that in Figure 2, the amplitude of each sample in the sampled data response is the impulse response integrated over the sample interval. In this case, the sample interval is 10pS.

In general, the response in Figure 2 is more or less what one would expect; however, there's a segment at the end of the response, around 10nS, that doesn't seem to make any sense at all. This segment is an aliased replica of the non-causal portion of the sampled data response. That is:

1. The inverse discrete Fourier transform calculation was consistent with the classic definition [4]

$$f(lt) = \frac{1}{N}\sum_{k=0}^{N-1} F(k\Omega)e^{jTkl\Omega} \quad \text{(EQ 3)}$$

   where N is the number of samples, Ω is the sample spacing in the frequency domain, and $T = \frac{2\pi}{N\Omega}$ is the sample spacing in the time domain.

2. The time domain samples were calculated for $0 \le l < N$, which is the range over which this calculation is usually performed.
3. Equally valid and complete results would be obtained if the samples were calculated over the range $-\frac{N}{2} < l \le \frac{N}{2}$. Such results are illustrated in Figure 3.

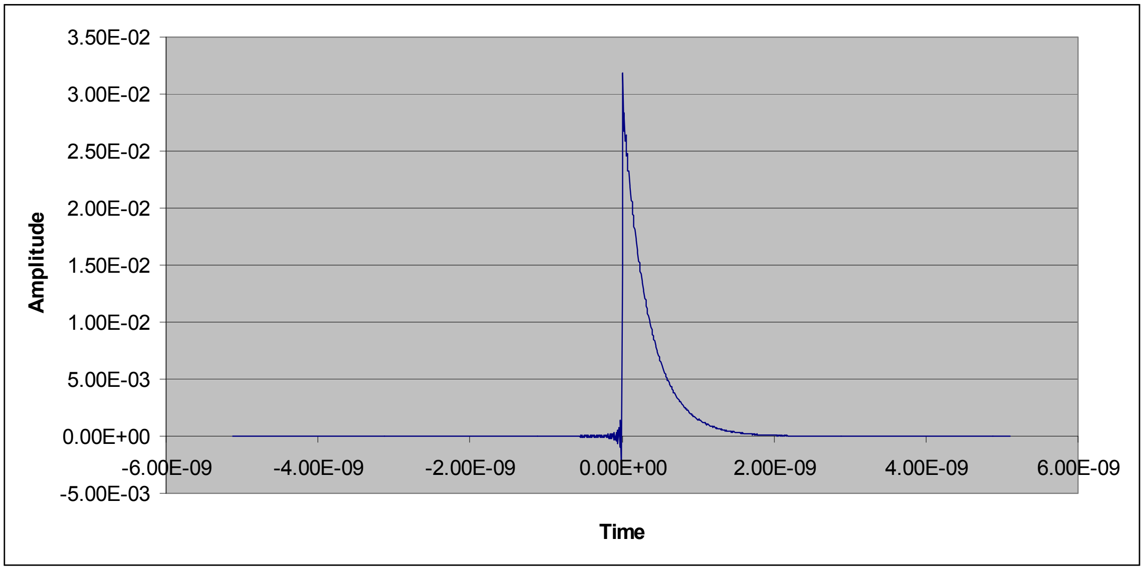

FIGURE 3. Aliased sampled time domain response

What appears to be happening in Figure 3 is that treating samples of a continuous frequency domain response as though it were sampled frequency domain data produces some unexpected numerical artifacts. These artifacts are generically referred to as Gibb's phenomenon, and are actually widely known, mathematically sophisticated, and thoroughly studied [5]. They are most pronounced whenever there is a discontinuity in a function or. to a lesser extent, in one of its derivatives.

This time domain response can be made completely causal by setting all samples before time = 0 to zero. Figure 4 shows the time domain response with and without the noncausal part.

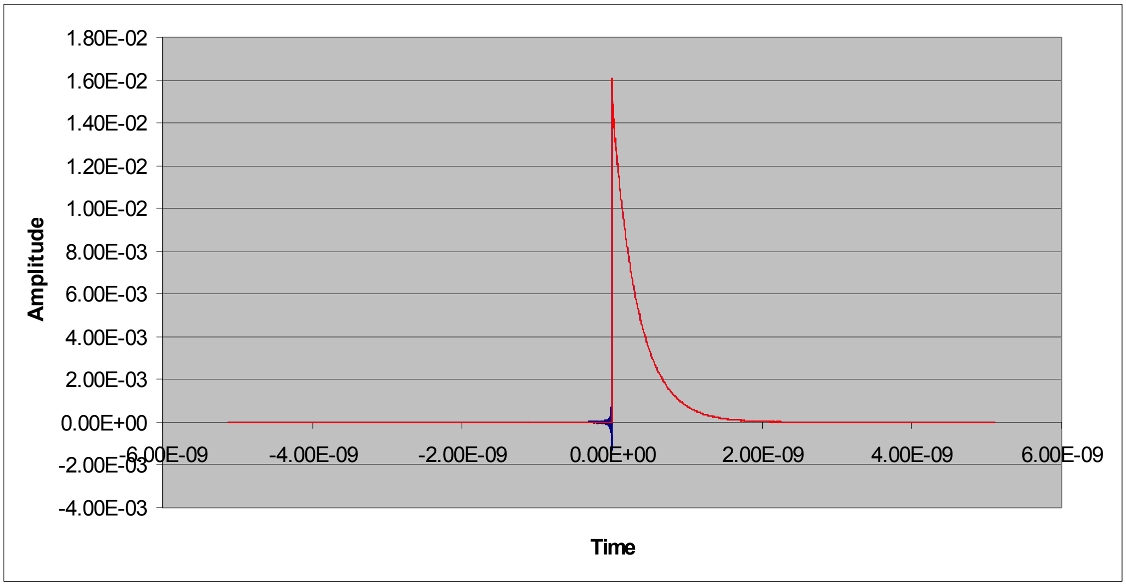

FIGURE 4. Aliased sampled time domain response with and without noncausal part

For later reference, the energy in the non-causal part of this response is $1.28x10^{-5}$. Interestingly enough, this energy is dependent almost entirely on the frequency spacing and insensitive to the maximum frequency.

From Figure 3, one way to make a sampled transfer function causal is readily apparent: Use an inverse Fast Fourier Transform (FFT) implementation of the inverse discrete Fourier transform to obtain the sampled time domain response, truncate the noncausal part, and then use the FFT to get the corresponding frequency domain samples. For the current example, the results are shown in Figure 5.

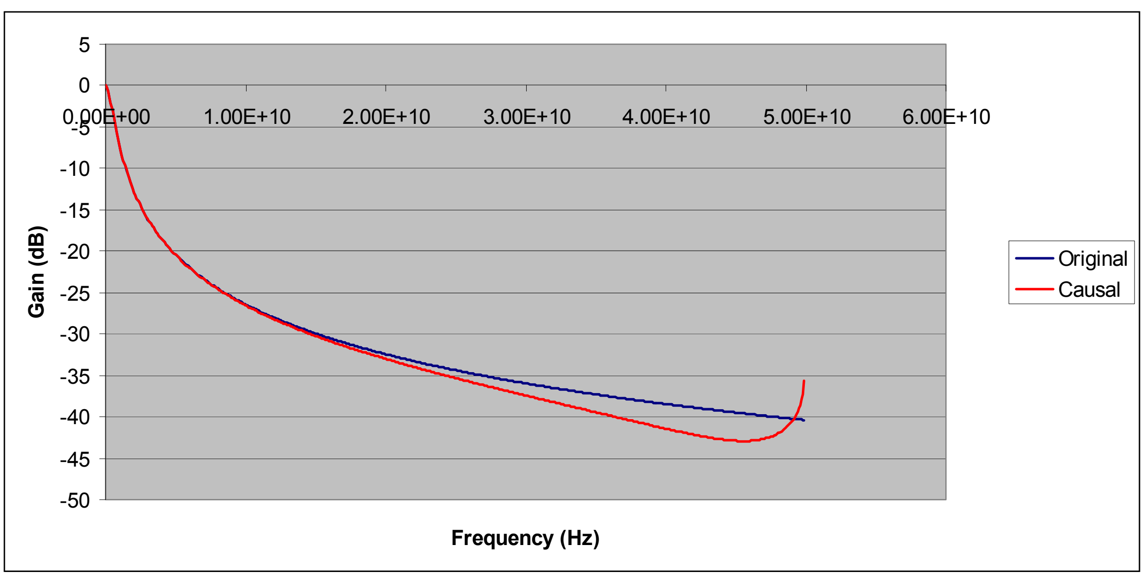

FIGURE 5. Sampled transfer function before and after causality correction

From Figure 5, it is apparent that the fidelity remains quite good at low frequencies, but that at high frequencies the loss has to be increased in order to attain causality in the sampled data domain.

2.2 Improving the Result

From Figure 5, one would suppose taking the samples out to a higher frequency would improve the result. To test this, we increased the frequency domain step to approximately 200MHz (5pS time domain interval) while maintaining the same number of samples. The results, before and after causal correction, are shown in Figure 6.

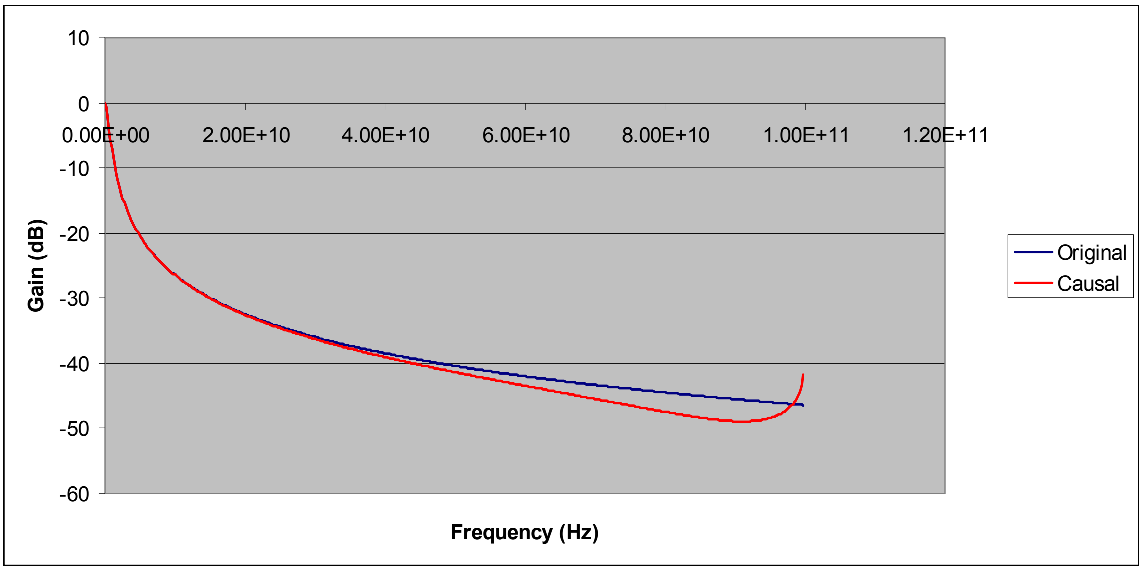

FIGURE 6. Frequency domain interval increased to ~200MHz

These results do look better, and the time domain results corroborate that impression. The non-causal energy has dropped from 1.28×10^{-5} to 3.21×10^{-6}.

For future reference, Figure 7 shows the corresponding time domain response. Note that the time span has decreased by a factor of two because the frequency interval was increased by a factor of two.

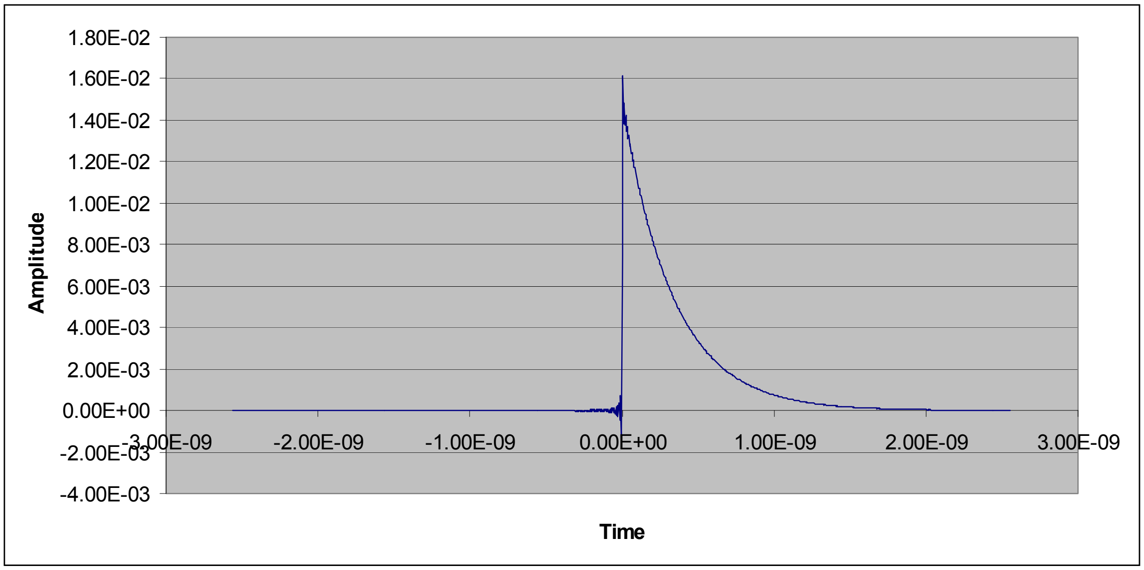

FIGURE 7. Time domain response for ~200MHz frequency spacing

2.3 Too Much of a Good Thing

On the theory that "more is better", we increased the frequency spacing again to approximately 500MHz (2pS time domain interval), again while maintaining the same number of samples. The results are shown in Figure 8.

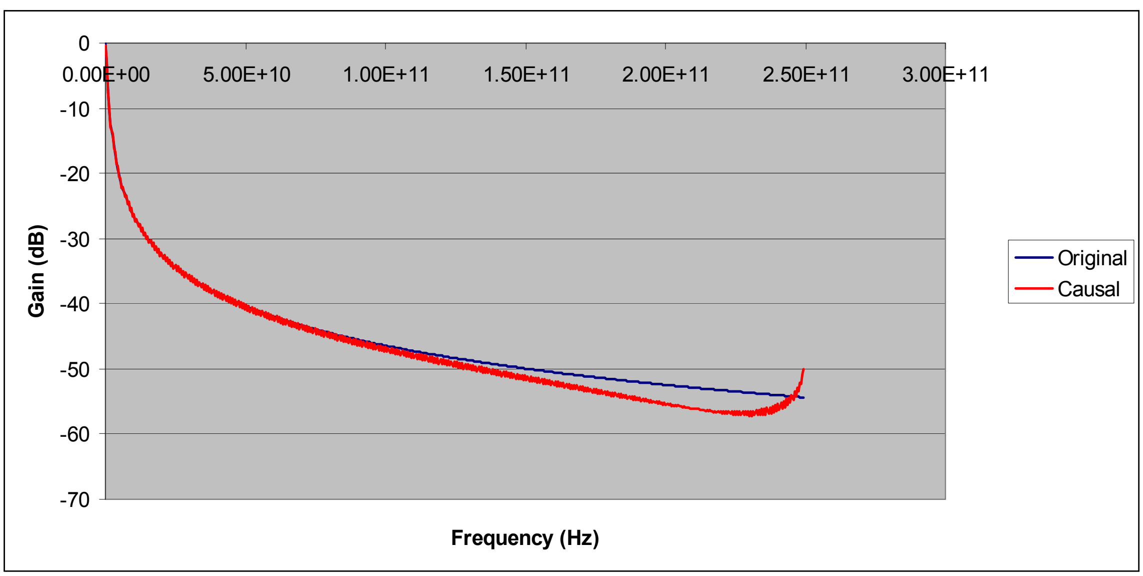

FIGURE 8. Frequency interval increased to ~500MHz

Something appears to be amiss in Figure 8 in that there appears to be a lot of ripple in the causal transfer function. This impression is confirmed by the fact that the non-causal energy has increased to 7.00×10^{-6}.

So, what went wrong? The answer comes from looking at the aliased time domain response, Figure 9.

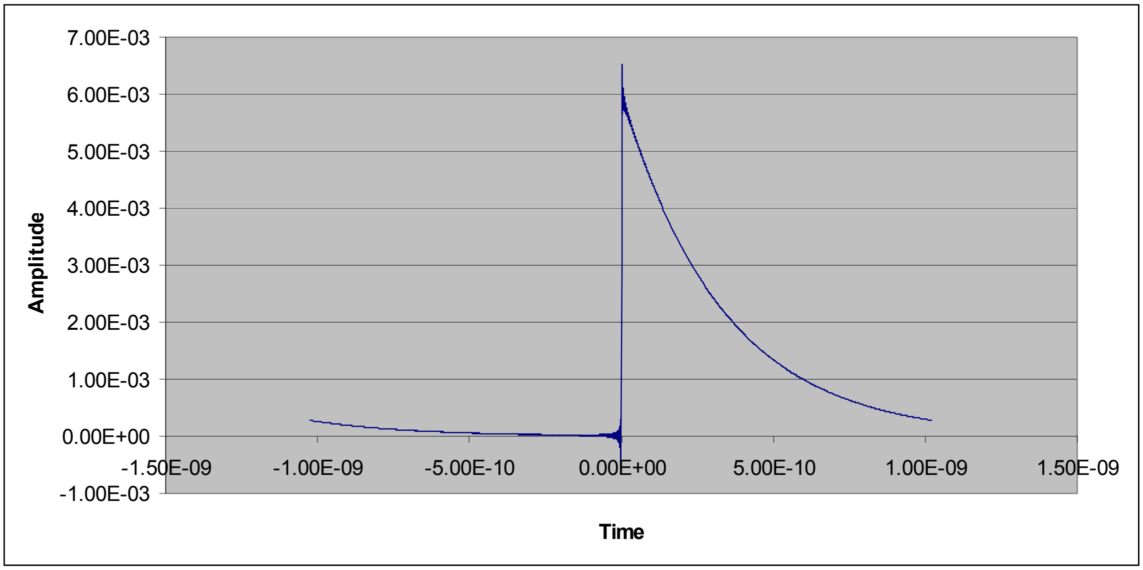

FIGURE 9. Time domain response for ~500MHz frequency spacing

What happened was that not only did some of the Gibb's phenomenon response get aliased to the non-causal part, but some of the desired response got aliased as well. The positive half of the sampled time domain didn't extend far enough to contain the entire response.

2.4 Making the Right Choices

2.4.1 Frequency spacing

The size of the positive half of the sampled time domain is determined directly from the frequency spacing by

$$T_{max} = \frac{1}{2\Omega} \qquad \text{(EQ 4)}$$

Therefore, when measuring or estimating the S parameters for a circuit, it's important to understand how big a T_{max} is needed. For a via, one can reasonably expect the time domain response, including internal reflections, to die out in a lot less than 1nS; whereas a 10 meter cable could have an average group delay of 50nS and require a little more time than that for the impulse response to decay. If the cable is very low loss, one might also need to account for a round trip reflection, or 150nS.

Applying Equation 4, one might use a frequency spacing as large as 500MHz for a via, but should expect to use a frequency spacing of less than 3MHz for a 10 meter cable.

2.4.2 Frequency range

From the demonstration above, it appears that if a causal correction is made to frequency domain samples of a continuous time process, the data should remain nearly unchanged for the first quarter of the frequency range and reasonably close over the first half of the frequency range; however, the data could change significantly over the second half of the frequency range. There are all imprecise statements; however, getting a more precise statement need not be much more than a spreadsheet exercise. (The hardest part is getting Touchstone[R] formatted S parameter data into the spreadsheet in the first place.)

For the analysis of high speed serial channels over lossy transmission paths, most of the energy is at or below a frequency equal to one half the data rate and very little energy ends up above a frequency equal to the data rate. Given the observations above, an S parameter frequency span equal to twice the data rate should be sufficient for most applications. Go to three times the data rate if you have a low loss channel or want to make a Project of it.

2.5 On the Hazards of Extrapolation

One might be tempted to use extrapolation to extend the frequency span of their data in an effort to obtain more precise time domain simulations. This may or may not be a good idea, depending on the assumptions one can make about the circuit being characterized and the extrapolation method to be used.

When there are reasonably reliable assumptions one can make about a given circuit, then there may be a reliable way to extrapolate from the available data. After all, all engineering is based on extrapolation from a set of assumptions that have proven to be valid often enough that the extrapolation, and therefore the engineering decision, has acceptably low risk. For example, if one is characterizing a transmission line such as a PC board trace or a cable, it may be perfectly reasonable to extract some telegrapher's equation parameters from the available data and use that to extrapolate to higher frequencies. The telegrapher's equations have proven to be accurate in many applications over many years.

For many sampled data simulations, it's also necessary to extrapolate the S parameter data to higher frequencies because a smaller time interval than necessary was used- for example, to simplify the numerical processing or to obtain a more pleasing rendering of the waveforms. The smaller time interval translates directly to a higher maximum frequency, and so data of some sort is needed at those higher frequencies. This is acceptable so long as reliable data has been provided for all frequencies that are significant for the application being analyzed, and the extrapolation is only to frequencies where there will be very little signal energy in the first place.

One should not, however, depend on an extrapolation to frequencies which are significant to the application unless the extrapolation is based on proven assumptions. A mathematical extrapolation may look smooth, and so one is tempted to trust it. In the absence of a proven physical assumption, however, this may not be a good idea. For example, it may be tempting to use an RCM because it "extrapolates to infinity"; however, that does not guarantee a meaningful extrapolation to finite frequencies. Unless additional constraints are placed on the RCM, all that's guaranteed is that the extrapolation will go to zero at infinity. Mathematicians have known for years that polynomials are ill behaved unless fully constrained [6], and that observation extends to the numerator of a rational transfer function.

3.0 So, now what?

1. Make the frequency spacing in your S parameter data small enough to more than cover the largest delay expected, remembering that there can be a significant decay time after the arrival of the main pulse.

   $$\Omega = \frac{1}{2T_{max}}$$

2. Individual transfer functions in a S parameter matrix can be tested by applying the inverse FFT and examining the resulting time domain samples. Depending on the definition of time domain limits you choose, the suspect time domain samples could either be at times less than zero or at times greater than $\frac{N}{2}$. This can be done in a spreadsheet if the spreadsheet program provides an FFT.
3. Fix any non-causalities you find, if you must. Whether or not you make this fix may depend on the simulation/analysis tool you're using. Some definitely need causal S parameters, some can do just fine even if the S parameters aren't quite causal.
4. Directly supply reliable data for all frequencies that are significant to the application. For high speed serial channels, two to three times the data rate should be more than good enough.

4.0 References

[1] Yuriy Shlepnev, "Quality metrics for S-parameter models", DesignCon IBIS Summit, Santa Clara, February 4, 2010.

[2] "Kramers-Kronig relation", http://en.wikipedia.org/wiki/Kramers-Kronig_relations

[3] Colin Warwick, "Understanding the Kramers-Kronig Relation Using a Pictorial Proof", white paper, Agilent Technologies, http://cp.literature.agilent.com/litweb/pdf/5990-5266EN.pdf

[4] Bernard Gold and Charles M. Rader, *Digital Processing of Signals*, Chapter 6, McGraw-Hill, copyright 1969.

[5] http://en.wikipedia.org.wiki/Gibbs_phenomenon.
This site references J. W. Gibbs, “Fourier Series”, Nature 59, 200 (1898) and 606 (1899).

[6] George H. Sherr, ed., *The Best of The Journal of Irreproducible Results*, “The Adventures of Polly Nomial”, p.147, Workman P:ublishing, 1983.
Known to have been in wide circulation by 1967 (MLS saw it at NSF summer camp), and to have been submitted to *The Journal of Irreproducibe Results* by Richard A. Gibbs some time between 1955 and 1983.