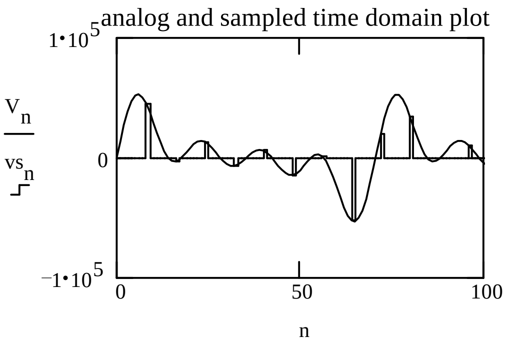
Sampling, Oversampling, Imaging and Aliasing - a basic tutorial

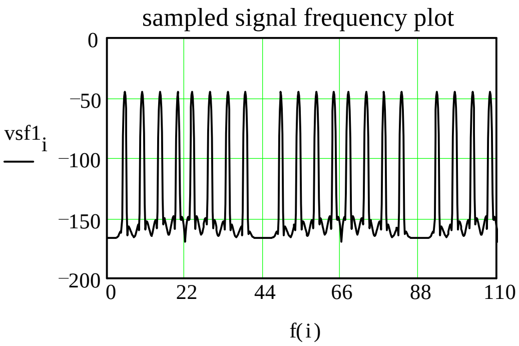
by Dan Lavry, Lavry Engineering

A motion picture presents a continues phenomena as a sequence of frames. We perceive continuity by "filling in the gaps" and "smoothing out " the missing information. While faster motion requires more frames during a given time period, there is a point where more frames per second yield no further improvement. The "sampling speed" must be fast enough to allow "appropriate averaging of dead times between frames", but need not accommodate motion too fast to be seen by a human. The above description is somewhat analogous to sampling of audio signals. The sampling rate depends on how fast the signals vary (the highest frequency we hear). Nyquist showed that sampling frequency exceeding twice the audio bandwidth is sufficient to include all the signal information, preserving the finest details up to half the sampling rate.

Sampling theory

Sampling a value, in the Nyquist sense amounts to "pinning down" signal values at equal time intervals. Let us "cover much of the audio band" by summing four equal amplitude tones (5,10,15 and 20KHz). The time domain plot shows the resulting continues (analog) waveform and it's sampled counterpart (44.1KHz sampling). The frequency domain plot shows the energy concentration of the sampled signal across 0 to 110KHz (tones exist beyond 110KHz). The sampled signal contains the four tones bellow 22KHz and undesirable energy at frequencies above 22KHz. Recovering the analog signal requires no more then removal of all energy above 22KHz.

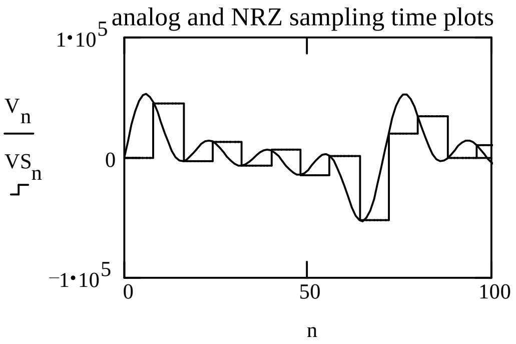


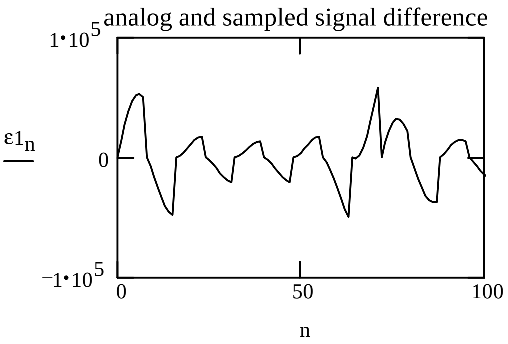


Our time domain plot shows very narrow sampling pulse width. "Sampling points" with zero duration pulses work nicely in a mathematical sense, but constitute physical impossibility. In fact, nothing ever gets done in zero time. Our real world sampling circuit can not charge a capacitor in zero time. The above plots were made with rather narrow pulses (1/8 of a sampling clock interval) causing amplitude loss. The energy peaks came to no higher then -45dB, compared to -27dB of original analog signal. Each halving the sample pulse duration results in a loss of half the signal. From an analog standpoint, small signals get buried in the inherent noise of the circuits. Digital considerations equate narrow pulses with finer time resolution, thus more storage and higher computation speeds. It is desirable to widen the sampling time duration to "span the complete sampling interval". Such a system is called NRZ (each sample time is extended to include the complete interval - Not Returned to Zero). What are the tradeoffs?

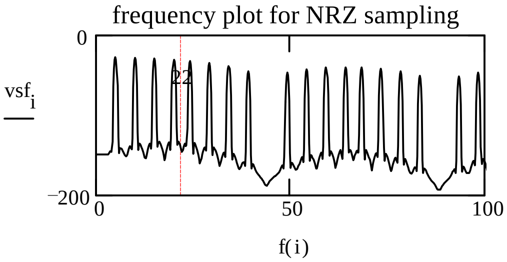
Practical sampling:

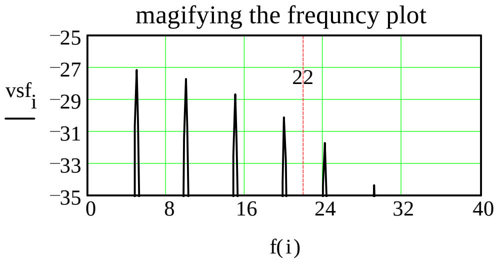
In the following NRZ sampling plots, we see each sample value held throughout the complete time interval. Subtracting the sampled signal from the analog one yields a difference (or error) signal, shown by the the right plot. Clearly, the difference is large. Did the error cause irretrievable damage?



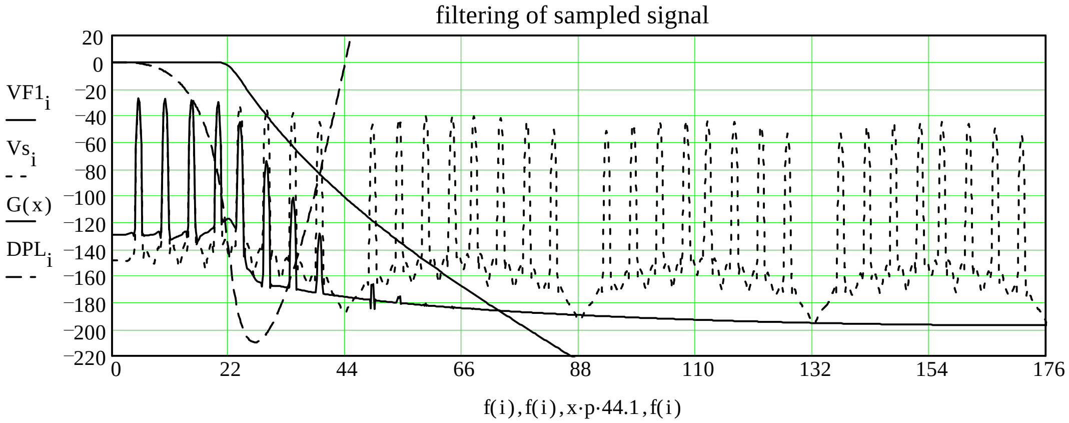


The overall appearance of the NRZ frequency plot is very similar to "theoretical sampling", but holding sample values for the complete time interval yields the advantage of higher amplitude. Zooming on the peaks of the four tones shows some undesirable side effects namely some high frequency attenuation. While such attenuation (about 4dB) is not acceptable for high fidelity, the cost for correcting the problem may be small when compared with the benefits of the NRZ system, though frequency response compensation (sinc curve) is somewhat complicated (ordinary filters can not follow the desired curve but close approximations are possible).



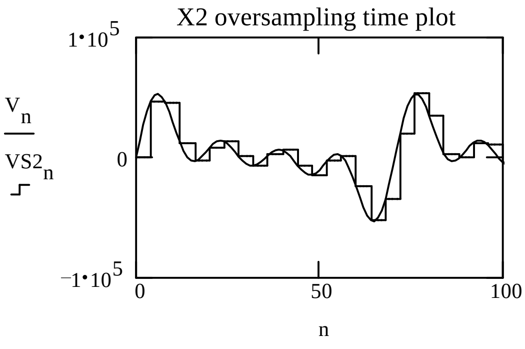


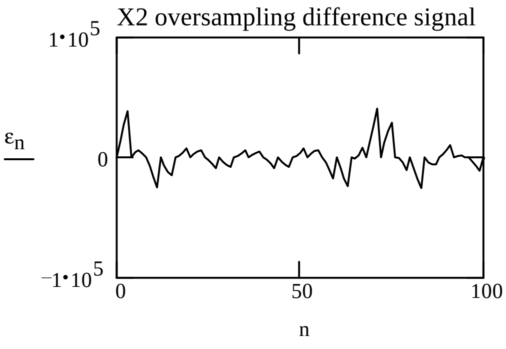
The "sinc problem" (undesirable inband high frequency attenuation) is not the only problem. We need to remove (filter) the out of band high frequency energy. The plot below shows an example of a rather serious attempt (16 pole Butterworth filter) to filter "out of band" energy. Note that the filtered energy just above 22KHz is "almost untouched" by the filter. To add insult to injury, such filters cause additional high frequency attenuation and also impact the audio path phase response at high frequencies (mostly above 10KHz), The numbers on the vertical axis indicate dB for filter gain (solid line) and degrees for deviation from linear phase (dashed line). Also showing filtered and unfiltered data.



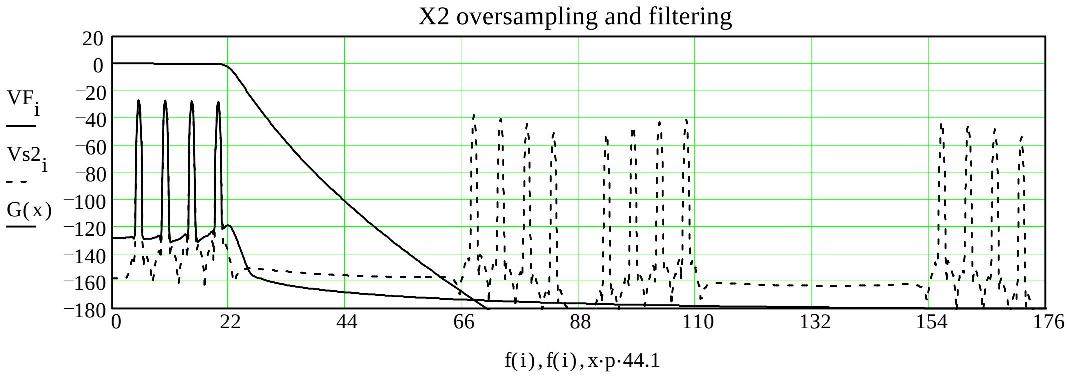
Oversampling:

Most digital audio equipment uses higher sampling rates then required by the Nyquist receipt. Oversampling offers solutions to both "sinc problem" and "filter problem". Oversampling typically takes place first during the analog to digital conversion. The signal is then converted to "standard rate", for reduced storage and computations. Such a conversion can be done without recreating much of the "sinc and filter problems". Later oversampling during the digital to analog conversion, yields freedom from from such problems as well. Let us examine our four tone signal sampled at twice the rate (X2 oversampling). The right plot shows the error signal (high frequency energy):

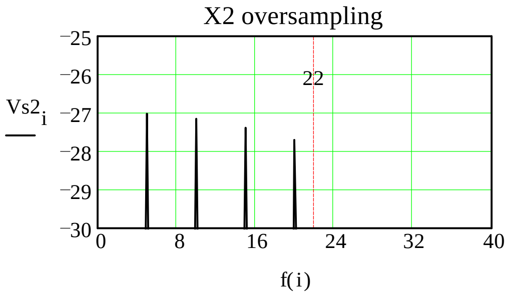


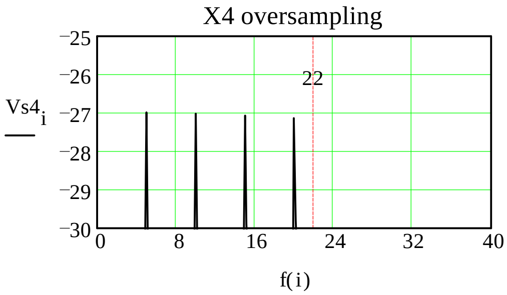


Sampling twice as fast, makes the NRZ time interval half as long, thus closer to the theortical flat response. The "sinc filter shape" is moved up by an octave, but doubling the number of samples, overcomes amplitude attenuation. Sampling at twice the speed also provides an "energy free zone" between the desirable frequency band and the undesirable out of band frequencies. Our filter is steep enough to remove all unwanted high frequencies. In fact, the cutoff can be moved higher to pass all the inband with minimal attenuation and phase distortions.

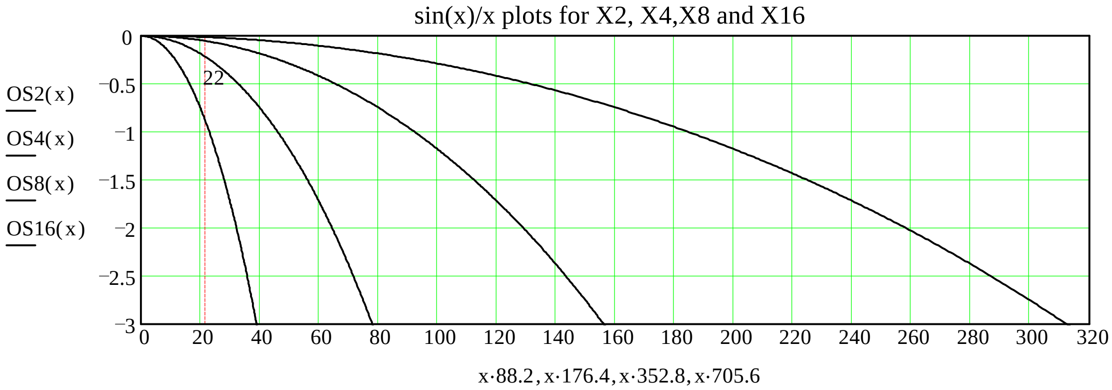


The following plots show the tone peaks for X2 and X4 oversampling. Note that sampling faster reduces the "4dB problem" to about .9dB at X2, and to .2dB at X4 oversampling.

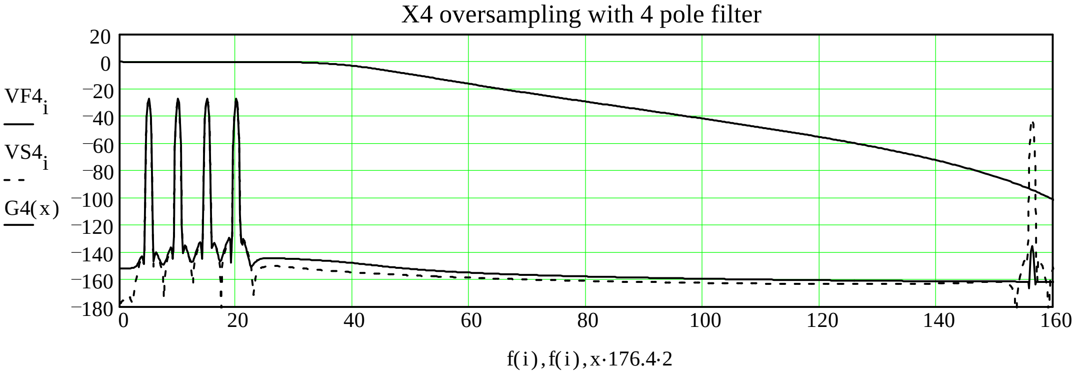




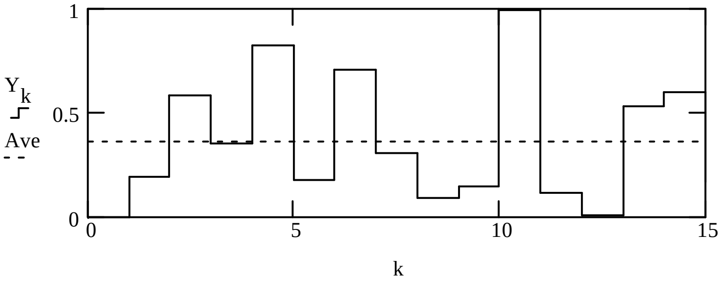
Oversampling by X4 may still require a slight amplitude compensation (an easy task). Higher rates yield so little attenuation that often no compensation is necessary. The plot bellow shows the sinc rolloff curves (sin(x)/x functions) for X2 through X16 rates. The axis are attenuation (dB) and frequency (KHz):



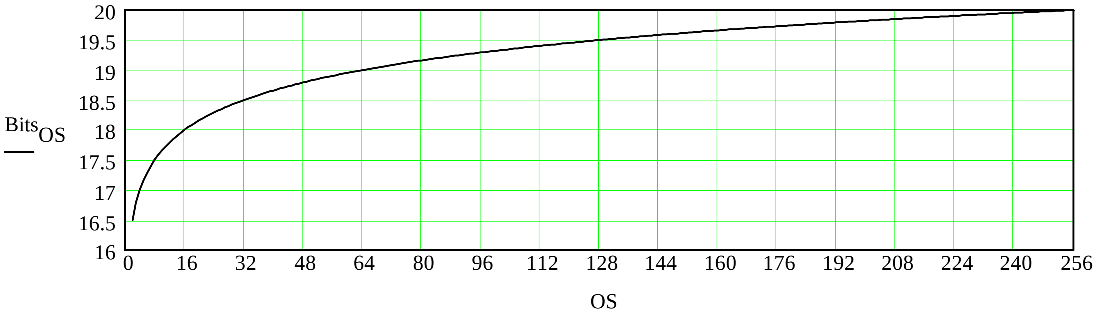
The next plot shows the removal of unwanted high frequency energy by use of a "simple" 4 pole filter for X4 oversampling. The energy at 150KHz (the first "unwanted spike") is attenuated to almost -140dB. The results are much better then before (16 pole filter shown earlier with no oversampling). The 3dB filter cutoff is moved to 40KHz, thus filter attenuation at 22KHz is minimized. Phase linearity in the passband is drastically improved for two reasons: a. Lower order filter yield better phase characteristics b. Filter cutoff moved up by an octave.



Oversampling and "more bits":
Your stereo dealer is selling a CD player with X8 oversampling 20 bits DAC. Do you hear 20 bits? Clearly, incoming samples with 16 bit accuracy can not be interpolated into 20 bits. The best of geographical surveying equipment yields errors when the reference markers are off. Oversampling interpolation is an "averaging concept" thus it yields some better "average accuracy", but each interpolated sample accuracy is limited to that of the input samples (16 bits in the case of CD players). Let us examine 16 random values (between 0 and 1) and their average value:

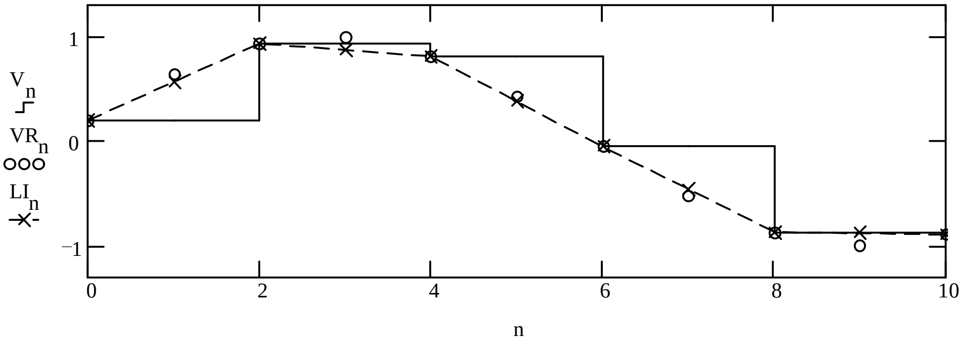


Averaging more samples make the average value converge closer towards 1/2, which is the mid point between our two 16 bit numbers (0 and 1). Indeed, expressing the average value (1/2) requires one more bit. Doubling of the sample rate may require additional bit but the extra detail is worth up to half a bit. A different point of view: roundoff errors are spread over a given bandwidth. With twice the bandwidth, only half of the roundoff energy resides within the audible range yielding 3dB less noise voltage (square root of X2).
While an X8 offers 9dB better noise (1.5 "more bits"), the improvement does not yield better sample by sample outcome.
In fact, the above mentioned X8, 20 bits DAC, yields 17.5 bits of average theoretical performance. The plot below show the improvement in bits for various oversampling ratios. Note that we are talking about linear PCM systems. Noise shaping technology offers enhancement to the oversampling improvements but the subject is not covered in this article.

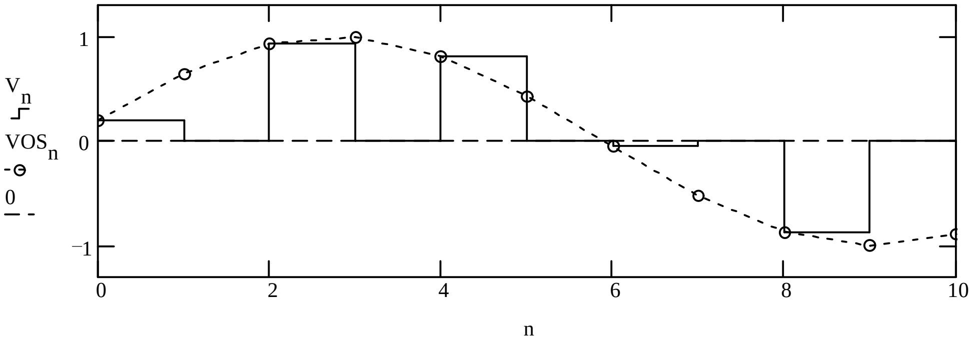


Oversampling offers great benefits in terms of amplitude flatness response and easy filtering, with much freedom from unwanted inband phase linearity problems. These concepts are beyond reach for most consumers, thus the "marketing department" decided to equate it with "more bits". While there is some truth to the story, much is being "stretched" a bit to far (and sometimes 3 bits).

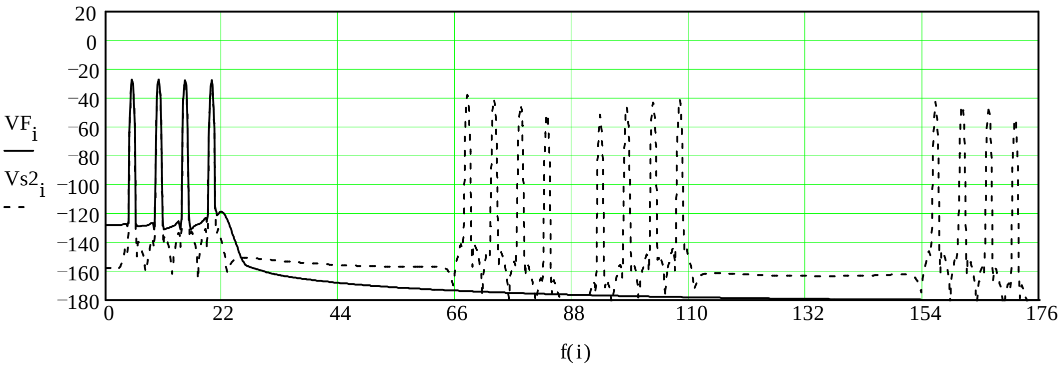
Interpolation and imaging:
Up-sampling is a mathematical process receiving incoming samples at a given rate and "creating samples" for a higher sampling rate. Let us examine an interpolation of a 7KHz tone from 44.1 to 88.2KHz (X2 oversampling ratio). The solid line represents the incoming data. The O's show the desired goal for X2 sample values. The dotted line with X's show the outcome of a simple two point straight line interpolation.



A linear straight line interpolation does not yield much precision (the X's are not centered on the O's). Increasing the order of algebraic interpolators (higher order polynomials) yield better results, yet the optimum performance is achieved by use non algebraic interpolation. If we could come up with sample values such that the all the error energy reside in out of band frequencies, we could then get the desired results by filtering. Filtering and averaging are close cosines. All we need to do is to insert zero values at the missing locations and filter out the undesirable energy at the new sample rate. The better the filter, (better attenuation of out of band undesirable energy) the better the interpolation. The next plot shows the process of zero insertion (solid line) and the filtered outcome (dotted line):



Insertion of zero values (or alternatively repeating previous sample values) does creates unwanted error energy, but all the undesirable energy reside in higher frequencies. Such error energy is distributed in a well predictable manner referred to as imaging. The unwanted frequency spikes described in the "sampling" sections are images. The X2 picture below shows the "mirror like" behavior of the images around the new sampling frequency (88.2KHz). The image of the highest original tone (20KHz) is found at 68.2KHz (88.2-20) thus can be filtered out.



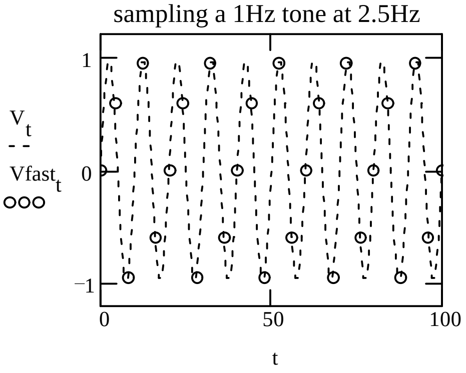
How much filtering is required? One may argue that since the ear can not hear 68.2KHz, filtering is not needed. The answer varies from case to case. Further signal processing may demand complete removal of all images. A playback DAC may need no filtering in theory, but real hardware is often led astray by the presence of high amplitude and high frequency tones (170KHz in the example above). Most power amplifiers, head phones, speaker systems and more, work better when not loaded by images. While higher oversampling generate higher frequency images, which could cause more problems, filtering becomes easier due to the larger gap between wanted signal unwanted error energy.

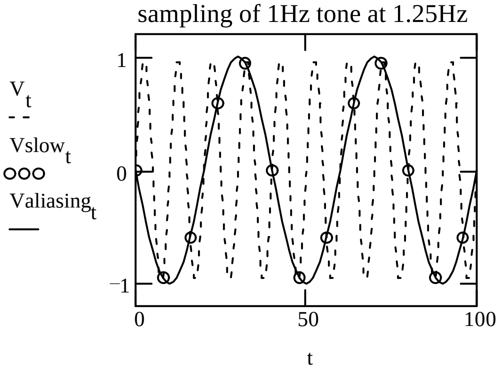
Aliasing:

Aliasing relates to having unwanted energy find it's way into the audio band. Violating the Nyquist limit (sampling at less then twice the rate of the highest frequency content) results in aliasing. Aliased energy can not be distinguished from desirable audio content, thus it can not be filtered out or later removed. One can avoid aliasing by either increasing the sample rate or by filtering the inband high frequencies prior to sampling (pre attenuating above half the sampling frequency).

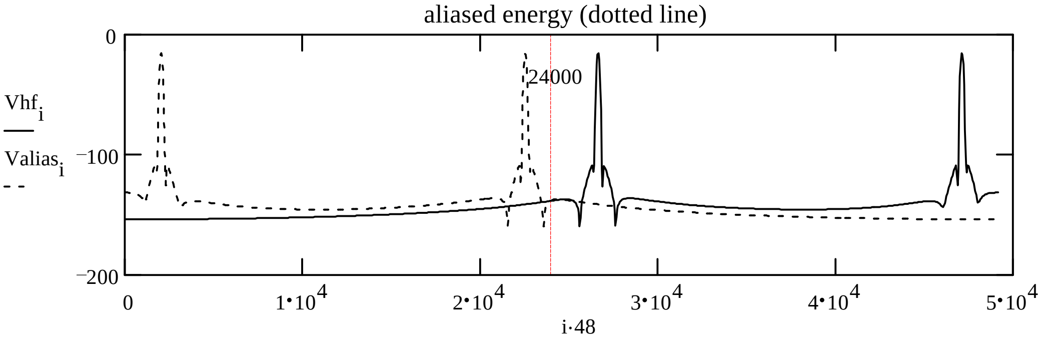
Most of us are familiar with the optical illusion created when viewing motion under a strobe light. A wheel rotating once per second will seem perfectly still with a 1Hz strobe light. Slowing the strobe to .9Hz makes for 10% more rotation between strobes thus the wheel seems to rotate at fraction of its real speed. These are examples of aliasing. Changing the strobe frequency to say, 2.5Hz will show that the wheel is not still, nor is it "moving slower".

Let us examine two waveforms: sampling at 2.5Hz provides enough sample points to "track" the signal. Sampling at 1.25Hz. shows a solid line passing through the O's shows - wrong slower "inband" frequency.





Aliasing "folds back" frequencies above Nyquist (half the sampling rate) with Nyquist being a "pivoting point". Sampling 25 and 35Khz tones at 48KHz, (24KHz Nyqiust) makes for aliases at 1 and 11KHz away from Nyquist. The fold back frequencies are 24-1=23Khz and 24-11=13Khz. Aliased tones are very irritating to the ear. They do not occur at multiples frequencies of the musical content thus are distinguishably inharmonic in nature. The plot below shows symmetry around Nyquist. Once again, oversampling would eases pre filtering with freedom from amplitude response and phase linearity problems.



Conclusions:

Theoretical sampling retain all signal information. Practical considerations take advantage of higher then theoretical sampling rates. Most analog to digital converters take advantage of higher sampling rates to overcome undesirable high frequency rolloff problems. Oversampling simplifies anti alaising filtering requirements and provides room for phase linear transfers. Storage and processing economy requires conversion from higher to lower sampling rates (downsampling). Data found in compact disk recording and similar formats is often oversampled prior to digital to analog conversion, to simplify anti imaging filtering.