NEW AUDIO FORMATS:
A Time of Change, and a Time of Opportunity

New audio formats will change the way recording, mixing, and mastering are done. There is no doubt about it. High resolution formats (96 kHz, 24-bit, DSD) and multi-channel formats (uncompressed 5.1) can be expected to permeate the audio industry in every way. The next few years promise to be a period of tremendous experimentation and invention. It is a very exciting time, indeed.

There seems to be a great deal of confusion about the meaning of these developments. It seems wise to go over the facts and see if we can separate the subjective components from the objective components. The purpose of this document is to try to put our evaluation of the new formats on a rational basis. This will not, by itself, answer all our questions: there is undoubtedly a subjective and preferential component to this discussion that we cannot address here. Each engineer and studio will have to decide how they will distinguish themselves. There is enough freedom and flexibility in the upcoming formats to support a wide variety of specialization and mastery. Let 1024 blossoms bloom.

Let us start with observations that are largely beyond question. These observations are not a subject of debate, but they beg further discussion:

- *96 kHz audio universally sounds better than 48 or 44.1 kHz audio.*
- *Many people prefer DSD to 96 kHz.*
- *Surround of just about any kind provides an enhanced experience. The difference between 5.1 and stereo is as great as the difference between stereo and mono.*
- *Nobody knows how interactivity will be used in music recordings, but we all know that it will be used.*

The topics have such wide engineering and psychoacoustic ramifications that it is hardly clear where to start. To quote the White King's reply to Alice when she expressed a similar sentiment, "Begin at the beginning. Go on until you reach then end, then stop."

Whither 96 kHz?

Raising the sampling rate to 88.2 or 96 kHz allows a pass-band of more than 40 kHz. The first question most everyone asks is whether we are making music for people or for dogs and bats. The highest measured frequency that has been found among humans is about 26 kHz - no individual has been found that can hear a steady-state tone any higher than that. Consequently, one might conclude that there is no point in having a sampling rate higher then, say, 60 kHz (to give us a bit of head-room as well).

Any student of human perception will point out to you that there is one aspect of hearing where extremely short time delays can be easily perceived, and that is in binaural (2-ear) hearing. If you put a pulse into one ear, then a pulse slightly delayed into the other ear, most people can hear a time delay of 15 microseconds or more. Under some circumstances, some people can hear time delays of 3-5 microseconds. Note that one sample at 48 kHz is 20.833 microseconds. At 96 kHz, it is 10.4167 microseconds. The minimum inter-aural (across the 2 ears) time delay that most people can hear is *less than one sample period at 48* *kHz*. This result is very hard to interpret in terms of frequency response, since we are talking about 2 channels of audio. Our conclusion from this is the following:

> *The appreciation of 96 kHz (or higher) audio, compared to 48 or 44.1 kHz audio, is a binaural (2-ear) phenomenon. If we plug one ear, it is unlikely that anyone would be able to distinguish a 96 kHz recording from a 48 kHz recording.*

The converse of this forms the basis of our first observation above: that when listening with both ears, everyone can distinguish 96 kHz recordings from 48 kHz recordings, and everyone prefers the 96 kHz recordings. We might say further that the reason they prefer the recordings is not because steady-state tones from 26 kHz to 48 kHz can be represented, but probably because some kind of time-domain resolution between the left and right ear signals is more accurately preserved at 96 kHz.

Whither DSD?

DSD® (Direct-stream digital) proceeds naturally from the philosophy that "if a little is good, a lot must be better." This technique uses a 1-bit signal sampled at 2.8224 mHz with noise-shaping. It provides even response over the audio band, and some amount of response in the very high ranges of 200 or 300 kHz - well beyond the range of direct human hearing. The time-domain resolution, then, is even more accurate than 96 kHz, or a hypothetical 192 kHz PCM recording. Most subjects report that the difference between a DSD recording and a state-of-the-art 96 kHz/24-bit system is subtle, but many professional engineers prefer the sound of DSD over 96 kHz.

Whither 5.1?

Note that what we mean by 5.1 is not compressed audio: for the purposes of this discussion, what we refer to is 5 channels of PCM audio (at any sampling rate from 44.1 to 96, and any bit width from 16 to 24 bits) and 1 channel of low-frequency data (generally taken to mean below 150 Hz).

Major motion pictures have been made in multi-channel for years. Spatial hearing seems to be such an integral part of our aural perception that it is widely believed to greatly enhance the experience. We should not be surprised that the same is true for recorded music without accompanying picture.

The commercial failure of quadraphonics (sometimes referred to as "quadra-fizzle") has been taken as a caution to optimism about consumer acceptance of multi-channel music. The situation is quite different at the present time, however, since many homes already have multi-channel (home theater) setups, and the delivery system (a DVD of some kind) is reasonably-priced and much more elegant than the quadraphonic delivery systems. Since in many homes, there is virtually no incremental cost to enjoying multi-channel releases, we can be sure that there is a guaranteed market for some number of such releases. We may differ as to the exact size of the market, but we may be assured that there is one.

Whither Interactivity?

At this time, noone knows what the impact of interactivity will be on pure music recordings. To date, all live music presentation is non-interactive. You go to a concert, the music starts, it proceeds to the end, then stops, echoing the White King's advice for a linear presentation. The only interactivity possible is that you could always leave the concert (or flip the CD to a different cut), or perhaps sufficient applause could entice the conductor into an encore presentation. Surely, classical music will continue to be presented this way, since it was written specifically to be heard in a linear fashion.

There is at least one way in which interactivity could be used in a non-threatening way that would be useful, and that is with alternate performance presentation. For instance, there are many recordings of live performances of the Benny Goodman Orchestra playing, say, "Stomping at the Savoy". How marvelous would it be to be able to flip between several different solos, taken at different times and different places. Certainly with performances that involve improvisation, there is some point to providing several different versions on the disk that are not played linearly, but would have to be selected.

It is not out of the question that there will arise a new form of music that has interactivity as an integral part of the composition. This would be music written by and for the "Nintendo Generation". Whatever we think of the generation of children that have been raised around the fast-paced electronic and computer games, and the kaleidoscopic, quick-cut editing that is used in modern media production, it is clear that they will demand their own culture and their own music. We have no idea of what form these compositions will take, but they will certainly arise. The next few years will certainly be an intense period of experimentation with the possibilities of the format, often with no more justification than "because it is there".

Technical Issues With High Sampling Rates

We will try to present some of the technical reasons why high sampling rates (96 kHz, DSD) are preferred. Note that to date, we have no hard proof that the aspects discussed here are the true reasons for the audibility of the differences. We know of no objective measurement that can consistently and clearly point to the superiority of the higher sampling rates in the audible band. Whatever is happening, it is happening in the range of frequencies from 0 to 20 kHz (more or less). The presence of higher frequencies may interact in some as yet unknown manner, but they are not directly perceived. The best we can do at this point is to go over what is known about higher sampling rates and to point out where these may interact with perception.

About Quantization Energy

A PCM word with a certain number of bits has a certain error. You could describe it as a certain number of volts at the converter output as the lsb (least-significant bit) is changed. Note that the voltage represented by the lsb is the same, regardless of how fast the lsb changes. Consequently, the *energy* in the lsb is constant regardless of the sampling rate. We can talk about the error of the lsb being a certain number of *volts per Hertz.* Given this definition, it is then simple arithmetic to note that the quantization error in volts per Hertz is smaller at higher sampling rates, such as 88.2 and 96 kHz, than at 44.1 and 48 kHz. All processing, such as equalization or dynamics, will share this lowering of the error due to quantization[1]. In general, we can make a generalization: *Keeping the sound at a high sampling rate, such as 88.2 or 96 kHz from recording to the final stage will generally produce a better product, since the effect of the quantization will be less at each stage.* One can think of the additional bandwidth as spectral "head-room", or perhaps as a "guard-band."

[1] Filters have different problem at high sampling rates, which is that the coefficient quantization becomes more significant at low frequencies. A 60-Hz notch filter at 48 kHz has about a .25 Hz precision using a

How To Build A Converter

Most A/D converters are built with a high-speed 1-bit noise-shaped ("sigma-delta") converter followed by a decimation filter which converts the 1-bit signal to a PCM word. The 1-bit front-end converters operate at some multiple of the sampling rate, such as 128x or 256x. Figure 1 contrasts three different kinds of conversion: 44.1 kHz, 88.2 kHz, and DSD. A typical 44.1 kHz might have a 1-bit converter running at 128x the sampling rate, followed by a 128:1 decimation filter. A converter running at 88.2 kHz might have the same 1-bit front-end, but would have a 64:1 decimation filter. The idea of DSD is to run the 1-bit converter at as low a rate as possible, such as 64x, but then eliminate the decimation filter entirely.

Theoretically, all of these systems could produce the same response in the frequency range from 0 to 20,000 Hz. Frequently they do not. We may get some hints as to why by examining how the decimation filters are built. Figure 2 shows a common way of building these filters. It may involve two or more stages of decimation. Breaking the process up into several, smaller steps reduces the total amount of computation involved, and consequently the amount of silicon required to build the chip. Decimation filters using 3 and 4 separate stages are not uncommon. After each stage, the signal is quantized to some amount of precision. Although this precision is generally greater than the width of the final PCM word, dither is not applied at these quantization stages, and there will be distortion products. The problem with distortion products is that they are statistical in nature: they will have phases that appear random, but are, in fact, correlated with the exact musical material. Even when numerical analysis indicates that the energy in the distortion products will be below the least-significant bit of the PCM word, the noise floor after an undithered quantization is *not* flat. It is statistically likely that from time to time, the pseudo-random phases of the distortion products will align such that a distortion product may be audible, even though the total energy in the distortion product is quite low.

Note that a high-resolution, 1-stage decimation process with dithering does not have this problem. As far as accuracy of the frequency response in the 0 to 20,000 Hz range is concerned, it is possible to build decimation filters, using either general-purpose DSP or custom silicon, that will produce substantially identical results in these three cases.

24-bit coefficient. At 96 kHz, the precision is 4 times greater (not 2 times) and is somewhat more than 1 Hz. This is different from and independent of any consideration of quantization error of the signal itself.

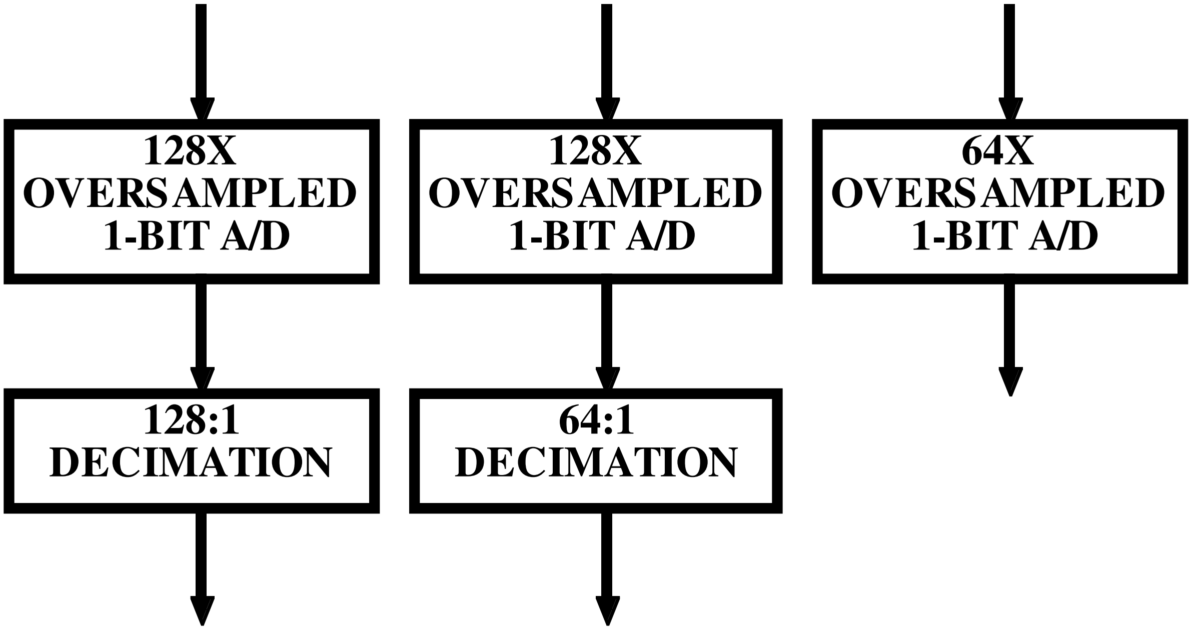

Figure 1: Comparison of the conversion chain for a 44.1 kHz converter, an 88.2 kHz converter, and a DSD converter. The key point is that the higher sampling rate converters generally involve fewer stages of decimation.

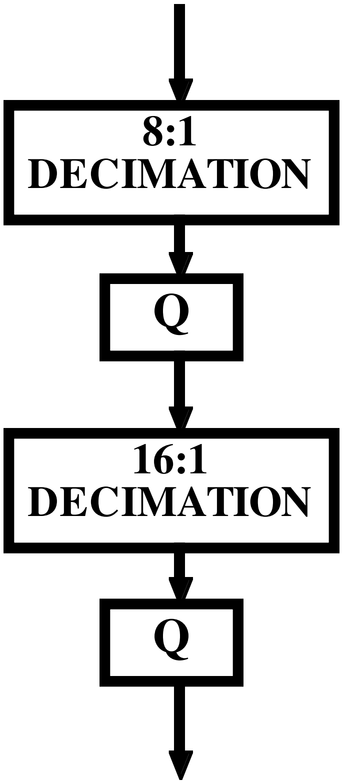

Figure 2: A diagram of a 2-stage 128:1 decimation filter. We have inserted a block labeled "Q" to make explicit the undithered quantization that necessarily follows each stage. We hypothesize that these quantization blocks contribute to the audible differences. This would suggest that the fewer decimation/quantization stages that are present, the higher the audible quality.

Given these results, we can make a statement that *on the average, it is more likely that a consumer-quality 96 kHz converter will sound better than a consumer-quality 44.1 or 48 kHz converter, simply because they might be built with one less decimation/quantization stage*.

Conversion As A Binaural Process

Human hearing, as with all our senses, depends on the firing of individual neurons. The neurons in the inner ear fire at a maximum rate of between 1000 and 2000 per second. At low frequencies, the firing of the neurons directly tracks the (positive excursion) of the sound pressure waveform[2]. At high frequencies, the neurons cannot follow the waveform exactly, so they fire as fast as they can. Note also that there is a relatively long latency between the physical stimulus and when the neuron fires - on the order of 100 microseconds or so. If you put these facts together, you might come to the conclusion that the ear has only limited time resolution. For a single ear, or for two ears that are presented with the identical signal, this is indeed the case. Short-term temporal masking in both the forward and backward directions are well-known in the literature of psychoacoustics. A single, loud sound can obliterate a quieter sound that occurs *before* the louder sound, by up to a millisecond or so.

Curiously enough, when both ears are involved, we get results that are hard to explain given what we just explained about firing latency and maximum rate. Our perception of direction of a sound is a rather complicated affair. We will not try to review the literature, but simply use one experiment to demonstrate the principal of binaural hearing. If you put a click into each ear, the time delay between the click in one ear and the click in the other ear will cause the click to appear to come from a particular direction. If the clicks are simultaneous, they appear to be directly in front of the listener. If the right ear leads the left ear, then the image moves to the right. We can now ask “what is the smallest time delay that can be perceived?” This is called the *just-noticeable difference*, or *jnd*.

As noted earlier, the jnd for time delay of clicks between the ears is about 15 *microseconds* for most people. For some people, it is less. Under certain circumstances, some people can detect delays in some sounds as low as 3-5 microseconds. It is perceived as a shift in the location of the sound. At the date of this writing, a specific structure that accomplishes this task in the brain has not been identified. It is somewhat difficult to imagine how the slow-firing neurons could accomplish this level of acuity. Clearly, for accurate binaural reproduction, more precision in the time domain is required than one would imagine, given the upper limits in our range of hearing. This confusion arises from the way these measurements are made: frequency response tests are done using steady-state sinusoids, whereas localization (binaural) tests are done with clicks or tone bursts. They measure different aspects of hearing. It is possible to have extremely fine time resolution (using two ears) without having infinitely-wide frequency response.

[2] Human hair cells along the basilar membrane of the inner ear only fire when the hair moves in the forward direction, thus they only respond to the positive (compression) wave. Guinea pigs hair cells fire in both directions. One wonders what music sounds like to them.

Decimation Filters and Time-Domain Spreading

To understand how a converter can produce time-domain distortion and still have a perfect frequency response, we have to turn to the anti-aliasing filter itself. Figure 3 shows an example of a 2:1 decimation filter that might be used, say, to reduce 96 kHz to 48 kHz. A similar filter would be used to reduce 128x 1-bit signals to 48 or 96 kHz. This filter could be designed to be virtually flat over the 0-20,000 Hz range (or even higher).

It is a fundamental principal of physics that a signal can not be simultaneously limited in frequency and in time to an arbitrary degree: the uncertainty principal shows that the product of frequency range and time range will remain roughly constant - any increase in one produces a decrease in the other. This comes into play in decimation filters, since as we attempt to limit the frequency range with a “brick-wall” filter, the response in the time domain widens. Figure 3 shows a central peak in the time-domain response of the filter, “winged” by two subsequent peaks on each side at times of 50 and 100 microseconds in each direction. These peaks form a “pre-echo” and a “post-echo”. As noted above, forward and backward masking renders these inaudible when presented to one ear. Any difference that these peaks produce between the ears is heard as some kind of spatial alteration. Since the effect the filter has on the sound depends entirely on the signal itself, it is not out of the question that the signal in one ear could excite one of these peaks, and not in the other ear. This would clearly alter the spatial imaging.

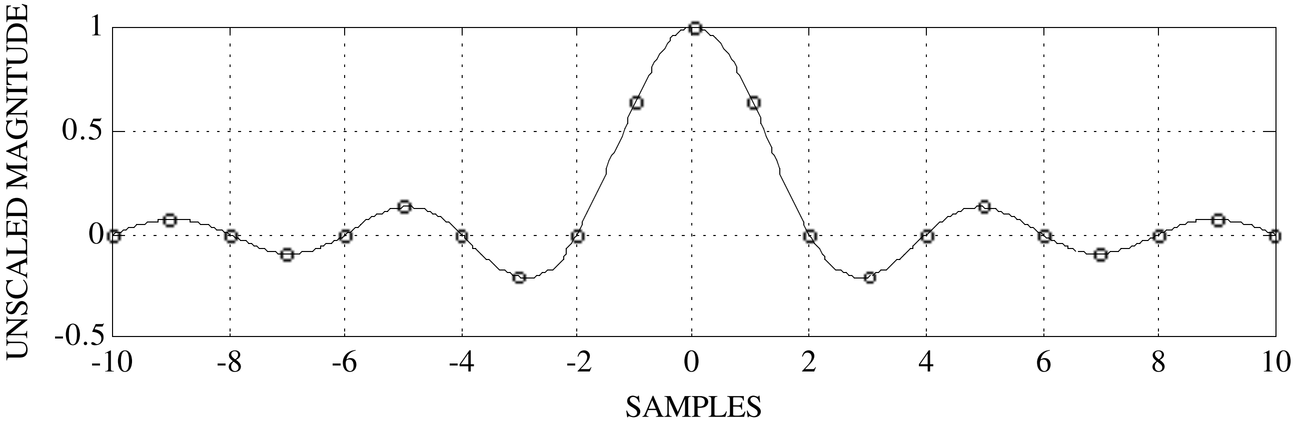

Figure 3: Time-domain representation of a possible decimation filter for use in an oversampling converter. Note that a maximum occurs at 5 samples removed from the center, which corresponds to a delay of 50 microseconds in this example.

As we increase the pass-band of the filter from 20 kHz (for a 44.1 kHz sampling rate) to, say, 40 kHz (for an 88.2 or 96 kHz sampling rate), the filter shown in Figure 3 becomes increasingly narrow. For a recording technique such as DSD, there is no time-domain spread at all. Theoretically, according to this argument, DSD should be capable of providing stereo imaging that is superior to either 48 kHz or 96 kHz.

There are numerous ways to design the anti-aliasing filters, but one cannot escape the limits of the laws of physics. There are design techniques that can smooth out the peaks on each side of the filter response, but one cannot reduce the total spread in time without increasing the bandwidth and thus increasing the sampling rate.

What About Multi-Channel?

The biggest change the consumer will experience in the new release formats is the possibility of multi-channel (surround) music recordings. Whereas the high-resolution techniques are audible and perfectly clear in the professional setting, they are somewhat subtle or even inaudible in the home. On the other hand, multi-channel is completely blatent "in-your-face" effect that is striking under any monitoring conditions.

Although we have decades of experience with multi-channel audio in the form of sound for major motion pictures, we have relatively little experience with multi-channel music recordings. Some amount of experimentation was done in the quadraphonic era, but these never achieved wide popular distribution. At least two of the reasons for the failure of quadraphonic are no longer relevant, and these are the extra speakers and amplifiers required, and the unique nature of the quadraphonic delivery medium and players. Since many homes will already be equipped with home theater surround systems, there is no extra expenditure by the consumer required to enjoy multi-channel music. The downside of this is that the audio quality of the home theater systems is not as high as one might want for pure music appreciation, and many of the current family of DVD players do not do multi-channel PCM. The DVD format itself, however, is a convenient and practical medium for the delivery of high-quality, multi-channel PCM.

There is one important fact about multi-channel releases that affect the professional world, and it is that *production of multi-channel pure-audio releases will affect every aspect of the production chain, from microphone placement, to recording techniques, to pan matrices, to monitoring systems, and ultimately to consumer systems.*

For instance, let us consider pan matrices. In film-style panning, the sound is positioned at an angle to the listener by feeding some amount of audio to the two speakers on each side of the desired position. It is relatively straightforward to show by objective calculation that this produces some of the worst spatial imaging imaginable. This produces an "image-spread" of up to 15°. The imaging is correct only when the sound comes out of one and only one speaker, or when equal amounts of sound are delivered to adjacent speakers. Without going into the technical details, we can say conclusively that *by feeding some sound into all five (or more) speakers, the imaging can be made essentially perfect by any objective measure*. The calculation of the exact gains to each of the channels is rather complex, but easily within the bounds of what is possible on modern DSP systems in the studio. Additionally, it is possible to correct in the home for speaker layouts that differ from the speaker positioning in the studio by DSP.

One important issue that we are all aware of but is seldom discussed explicitly is that *different recording engineers listen to and optimise different aspects of the sound.* This is good, since it gives the consumer a wider variety of recording and production styles. It does make it difficult to design one production chain that will cater to all kinds of music production.

For example, some studio artists seek a feeling of spaciousness in their recordings. Localization (that is, precise imaging of each individual sound) is only one part of this predilection. One objective measure of spaciousness is left-right decorrelation - that is, making the sound as different as possible from left to right, but still maintaining some coherence (such as diffuse reverberation that is equal in all channels). This immediately implies that the center channel should be set to zero, since *any* contribution from the center channel will increase the left-right correlation. This is not just armchair psychoacoustics - this can be derived objectively from first principles. Consequently, any system that forces signal in the center channel will not be used by one group of professionals.

There is a great deal of experience with microphone techniques for stereo recording and presentation. There is no way to directly generalize this to multi-channel in any rational way. It might be possible in some cases with some recordings, but there is no “universal” technique that will produce uniform results for all recordings. Looking forward, we can anticipate that future recordings will be made with multiple, simultaneous microphones - some will be used for the stereo release and some will be used for the multi-channel release. Again, we can look forward to a period of intense experimentation and creativity in this domain.

Probably the most controversial aspect of multi-channel is that there will be significant pressure to re-release mono and stereo catalog material in multi-channel format. Whatever we may think of this, it is worthwhile to consider rational methods for doing this that have some objective, perceptual basis. There are technical ways to extract spatial information from a stereo recording that can help in designing multi-channel feeds for this material that can aid in preserving the original intent in the recording without resorting to ad hoc “surroundizing” techniques. This involves some rather difficult DSP and a great deal of user direction to produce good results. There will not be an automatic technique that will consistently produce superior results.

What About Multiple-Format Releases?

There are a few things we know about the release formats at this time:

- *The 44.1/16-bit CD will be with us for some time to come.*
- *If you ever want to release with video, then the audio must be available in 48 or 96 kHz.*
- *Even if music is released in multi-channel form, a stereo release will be required.*

This gives us an interesting quandary: Given that we will be releasing in multiple formats, how should we record, edit, and mix sound to preserve the flexibility necessary for the multiple release formats? Although there are some techniques and suggestions for mixdown of 5-channel recordings to stereo, the highest-quality productions will have to mix the 5-channel differently from the stereo. There is no matrixing technique from 5 channels to 2 that would give the same results as if the program were mixed explicitly for stereo.

Another issue is high-quality sample-rate conversion from 96 kHz to 44.1 kHz. This can be done in a totally transparent fashion, but it must be done with care. The possible forms of distortion in a 2:1 conversion, such as from 88.2 to 44.1 kHz, are only those of coloration of the sound and possible aliasing of high-frequency material. The 320:147 ratio implied by the 96 to 44.1 conversion adds an additional problem, which are distortion products equivalent to multiplying the signal by a pseudo-random periodic sequence with a 147-sample period. Figure 4 shows a highly-exaggerated simulation of an improperly implemented 96 to 44.1 kHz downsampling. The original 1kHz sinusoid is accompanied by distortion products at a 300 Hz spacing in both the positive and negative direction.

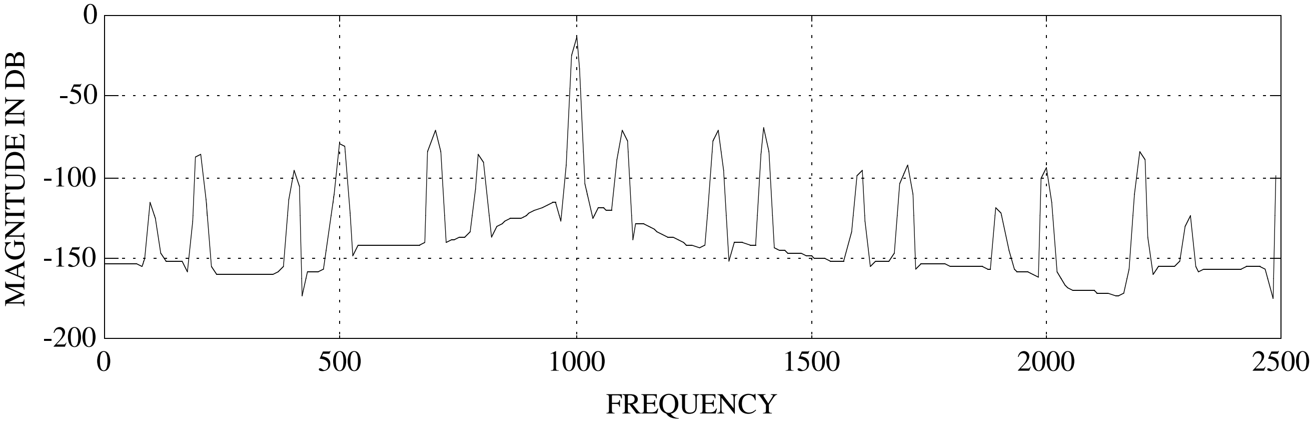

Figure 4: Exaggerated error spectrum resulting from improper downsampling from 96 kHz to 44.1 kHz

Note that with sufficient precision, this conversion process can be done without measurable error, but it requires considerably more precision than the 2:1 downsampling.

The Challenges of DSD

Whereas 96 kHz PCM audio is in some ways just “more of the same,” DSD presents a different challenge. The computation rate is so high that it cannot be accomplished by general-purpose DSP chips in real time. It requires custom silicon. At 96 kHz, all stages of the production chain are available at this time, from conversion to recording to editing, mixing, and equalization, to producing the final master, and to the consumer format and consumer players (in stereo, at least). With DSD, there exist at this time prototype recording devices and a stereo DSD editor using custom silicon. Silicon for equalization and other processing (dynamics, reverberation, *etc.*) is not available at this time. To carry DSD through the entire chain will require rebuilding each stage of the chain explicitly for DSD. Although the same chain can switch smoothly among 44.1, 48, and 96 kHz, the same can not be said of DSD. This notwithstanding, there is no technical barrier to doing just that: revolutionizing the entire production chain to move entirely to DSD. The barriers are strictly of the commercial nature and not technical.

So What Does It All Mean?

We know that there will be interest in using all the aspects of the new recording formats. There will be increased pressure to produce high-resolution, multi-channel, and perhaps even interactive releases. We are on the beginning of a revolution in how music is made. We will be the ones that explore the possibilities of the new format, and we will be the ones to discover the best ways to bring enhanced listening experiences to the consumer. It may be a difficult few years, but it will be an exciting few years as well. We may never arrive at a technological level such that we are only limited by our imagination, but the possibilities that we can explore today get us one step closer.