DISPELLING THE MYTH OF THE FLOATING-POINT

Patrick Warrington, Calrec Technical Director, 1st September 2010

There are many fallacies and urban myths that persist well beyond their shelf-life. For example, it is said that a mattress may double its weight with dust mites after ten years, and that the Twinkie possesses a shelf-life longer than that of the Universe. Another is that the floating-point number format is a prerequisite for audio quality.

It is not only the credulous who fall prey to these fictions - if a myth is repeated enough times, it accrues a veneer of credibility that can fool any of us. It is the duty of scientists, engineers, and all people of good character and common sense to stand up and fight this tide of perfidious bilge.

For my part, I intend to put to rest all three of the above apocrypha beginning with the most easily dispatched

The myth of floating-point numbers

It is widely believed that floating-point is a byword for good quality in digital audio systems, or worse, a guarantee of it. In fact, neither of these assertions is true.

Floating-point numbers are like the scientific notation on calculators; they have a mantissa – the number part, and an exponent, which is a multiplier used to scale the number part. For example, 1.414×10^3 is a floating-point number with a mantissa of 1.414 and an exponent of 3. The attraction of this form is that it can be used to express numbers over a much larger range than would be possible if the same number of digits were used in a fixed-point (integer) number. They are the exuberant swashbucklers of the number universe, one minute spanning the interstellar vastness and the next swooping down into the sub-atomic cracks and fissures of particle physics. And if floating-points are the numeric super heroes then fixed-point numbers are the Clark Kents, rooted and reliable, fastidiously representing numbers within their compass and never venturing into the infinite darkness beyond.

So what impact does the number format have on digital audio systems? We must consider two properties; resolution and dynamic range.

Resolution or numerical precision is determined by word length; as it increases, the resolution improves. Dynamic range is also determined by word length, but may also be dramatically extended in the floating-point format by the choice of exponent. Compare, for example, a 24-bit fixed-point number which has a dynamic range of about 144dB, to a 24-bit floating-point number where eight bits are designated as an exponent, which has a dynamic range of over 1500dB.

How do we choose the right number format for a digital audio system? Well, the dynamic range and resolution needs to be sufficiently large to allow faithful representation of all audio signals that may be encountered. The table in Figure 1 shows the dynamic range of sounds in the real world. SPL (sound pressure level) is a logarithmic measurement of sound levels where 0dB represents the threshold of human hearing.

While the narrative requirements of most music, sport and drama do not demand the uncompressed reproduction of a jet engine's roar at close quarters followed by gently rustling leaves, (and if it did, would render the audience physical pain and possible hearing damage), the 24-bit fixed-point format would suffice. But does this make it a suitable format for digital audio systems? The answer is an emphatic no and the reason is that processing audio can introduce errors, which are manifested as audible noise, unless additional resolution is provided. Note, it is resolution, not dynamic range that is needed. To illustrate this, let's take a look at the most important audio processes.

Gain

Applying gain to a digital audio signal simply requires a multiplication; multiply by a bigger number to make the audio louder, or a smaller one to make it quieter. Easy! The problem comes when the result is a number that doesn't fit neatly into number of digits you have to represent it. There is usually an extra bit that you have to get rid of, known as truncation. How you do this has a big impact on the quality of the result. You can choose to round up or down but this introduces unpleasant quantisation noise. A smarter solution is to add a random number to the left-over bits and then round up or down. This fiendishly counter-intuitive idea is known as

FIGURE 1 - REAL WORLD SOUND LEVELS

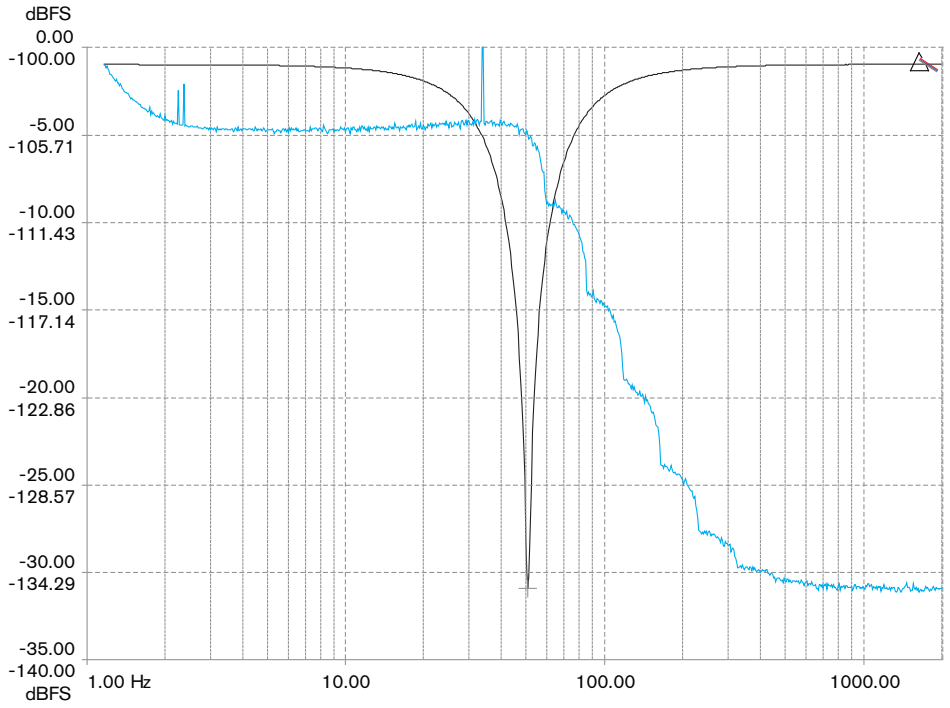
Calm breathing, or gently rustling leaves	10 dB
Normal conversation	40 – 60 dB
Passenger car at 10 metres	60 – 80 dB
Hearing damage (long term exposure)	85dB
Vuvuzela	120 dB
Level of sound that can cause physical pain	130 dB
Jet engine at 30 metres	150 dB
M1 Rifle at 1 metre	168 dB
Stun Grenades	170 – 180 dB

dithering and makes low-amplitude signals sound much better. The price – there's always a cost - is that the number format is required to carry additional resolution in the form of 'foot room' bits to dither against. Floating-point does not help as there is no requirement for an extended dynamic range. In fact, any bits given over to carrying an exponent are serving no useful purpose and would be better deployed in the mantissa, extending resolution.

Mixing

If a number of signals are to be mixed (added), then it is a good idea to provide some additional 'head room' bits which extend the dynamic range during the calculation, but this does not need to be very large. If a floating-point number were used to generate intermediate headroom, as the mantissa is scaled up, foot room bits are lost, mucking up any subsequent dither calculation and introducing noise. Hence, in mixing calculations, the floating-point format is, if anything, a liability.

GRAPH 1



Graph 1. THD + N for a 30dB notch filter at 50 Hz using 40-bit floating point. (Blue trace is THD+N)

Equalisation

EQ is a more complex case. The ubiquitous biquad calculation used in digital filters may be arranged in different forms to make it more convenient for processing. For example, the popular *direct form II* is used to reduce the number of computations in systems where DSP cycles are at a premium. The *quid pro quo* is that the intermediate calculations have large dynamic range and demand a floating-point format (at least it does in a system constrained to a fixed word length such as a DSP chip). In other words, the requirement for floating-point comes as a consequence of cost-cutting rather than the pursuit of quality.

The problem that arises from the use of a rigid floating-point format (for example, that found in ADSP SHARC chips) is that the resolution is fixed by DSP architecture, not by the requirements of the calculation. Most of the time, it is adequate, but there are certain filter configurations where it is not. Graph 1 shows a plot of THD+N for a 30dB notch filter at 50Hz executed in 40-bit floating-point (*direct form II*). It is quite clear that there is a significant elevation of the noise floor due to the resolution limit of the floating-point format.

A more quality-minded approach is to first decide what level of performance is desired and then to select the number format to achieve it. In the case of EQ, a very high level of resolution is needed in parts of the calculation in order to avoid generating the kind of noise evident in Graph 1. A flexible architecture, such as Calrec's Bluefin2, allows word length to be increased to precisely match the required performance.

Graph 2 (over the page) shows the results of the same filter in a Calrec Apollo console, with Bluefin2 DSP. The high word-length fixed-point approach has reduced the noise floor of the filter to more or less that of the test set.

There is a secondary effect resulting from the lack of resolution that impairs filter performance; you can't add very big numbers to very small ones. It simply doesn't work! Let me demonstrate this with an extreme case.

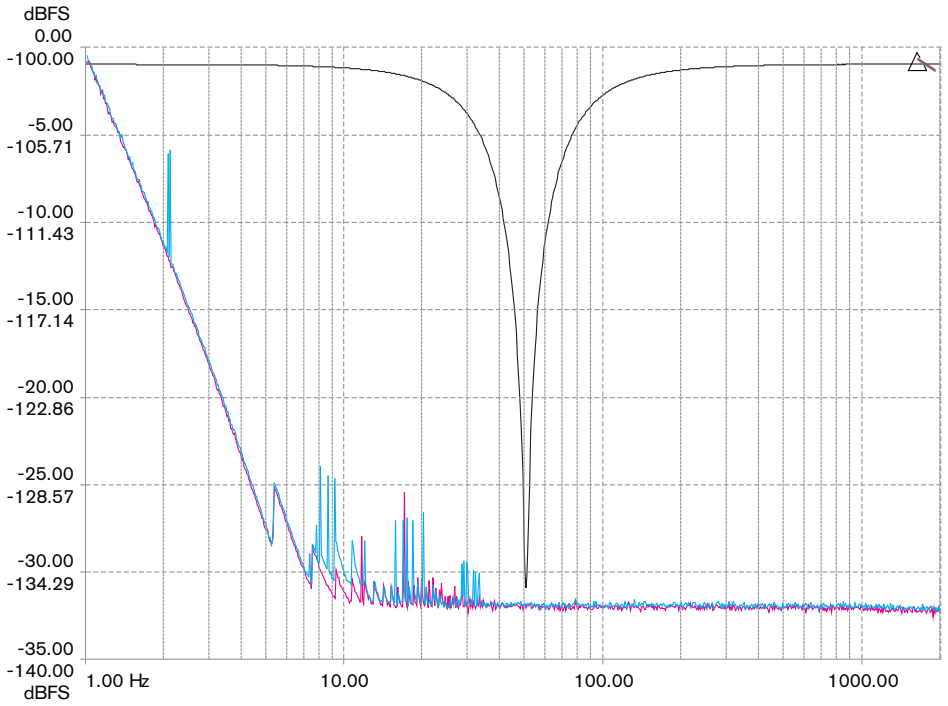
Imagine that you have adopted a 7-digit floating-point format with 4-digit mantissa and 3-digit exponent. You can represent the number one million by writing it as 1000×10^3. Now, add the number 999 to this. You should get 1,000999, but since you have only 4 digits available, the

result you end up with is 1000 x 10^3 – the number you first started with! Adding 999 has had no effect on the result. The floating point system has failed because it lacked resolution!

It is an inescapable consequence of the format that as floating-point numbers whiz around the numerical firmament, they deposit little piles of arithmetic ejectamenta in their wake. In the interest of balance, I would point out that the errors are, on the whole, quite small, and that for the majority of calculations, they are irrelevant, especially when compared to the more serious threats to quality along the route from microphone to living room. But if we choose to make numerical precision an aim (which we should) then we ought to do it properly and not let the want of a little analysis be a barrier to good science.

So that's the floating point myth dealt with, what about the mattress doubling its weight? Well, I've weighed mine and it hasn't. And the everlasting Twinkies? Actually, come to think of it, that might actually be true.

GRAPH 2



Graph 2. THD + N for a 30dB notch filter at 50 Hz (fixed point). The blue trace is THD + N of Calrec's Bluefin2 processing. The red trace is THD + N of the test set looped output to input.

Calrec Audio Ltd
Nutclough Mill
Hebden Bridge
West Yorkshire
England UK
HX7 8EZ

Tel: +44 (0)1422 842159
Fax: +44 (0)1422 845244

calrec.com