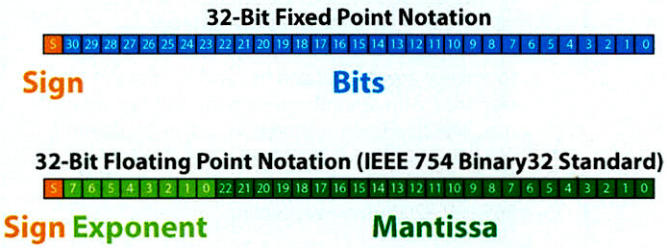
Fixed Point vs. Floating Point Illustration

By Donny Chow

Digital signal processing (DSP) can be separated into two categories: fixed point and floating point. There are a lot of discussions and information available on the two counterparts, yet the concepts are still confusing to many in understanding their differences and how those differences affect real-world products. The main purpose of this article is not to go too deep technically, but instead to explore the difference of the numeric representations in a simple yet intuitive way. Since it is only fair to compare an apple to another apple, we're going to use the same 32-bit length for both notations in this article as shown in Fig. 1.

FIG. 1



For Fixed Point Notation:

Integer Value = -1^{Sign} x Bits
Sign = 0 (Positive value) or 1 (Negative value)
Bits = 2^{31} possible values
Positive Min: 1
Positive Max: +2147483647

With simple scaling, fractional numbers can be represented as well. The represented values are equally spaced across the whole range. The gaps between adjacent values are always the same.

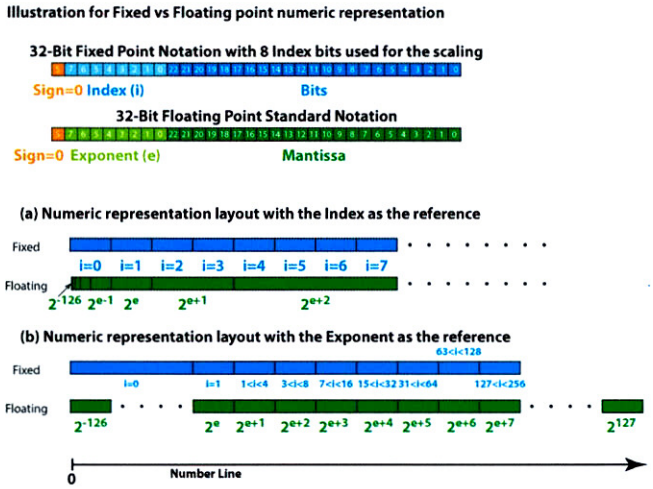
For 32-bit Floating Point (IEEE 754):

Value = -1^{Sign} x [(1+ Mantissa) x $2^{(Exponent-127)}$]
Sign = 0 (Positive value) or 1 (Negative value)
Exponent = 1 to 254 (0 and 255 are reserved for special cases). Subtracting 127 gives -126 to 127
Mantissa = 0 to 0.999999881 for all '0s' to all '1s'. Adding 1 gives 1.000000000 to 1.999999881
Positive Min: $1 \times 2^{-126} \approx 1.2 \times 10^{-38}$
Positive Max: $1.99999881 \times 2^{127} \approx 3.4 \times 10^{38}$

The represented values are unequally spaced between these two extremes, such that the gap between adjacent numbers is much smaller for small values and much larger for large numbers. This notation "steals" eight bits to become the exponent, which gives the extensive dynamic range, but to gain this benefit, it loses the "stolen" eight bits' resolution for all values.

Fig. 2 illustrates how the two notations differ in their numeric representations. In this illustration, we just focus on positive numbers (Sign Bit=0) to avoid complications. Then the eight most significant bits in the standard fixed point format are used as the scaling index to match the number of bits of the exponent in the floating point format. This leaves the remaining 23 bits to be matched for both formats. Once the bit-matching is accomplished, we can then generate meaningful results.

FIG. 2



Exponent e can be any arbitrary number as long as all blocks are within the min/max range (-126 to 127). The Index=1 is mapped to e once the desired data range is chosen for the fixed point computation. The graph shows that both notations have the same 23-bit resolution at i=1. For i=0, the resolution of floating point is better; in fact, tremendously better. There are e+126 floating point blocks, while there is only one fixed point block (i=0), both covering the same numerical range.

For 1<i<256, the resolution of fixed point is better. Although it seems floating point loses ground at higher values, keep in mind that the fixed point counterpart can only have a total of 256 (2^8) blocks with 23-bit resolution per block. Once the value goes beyond the upper limit, the developer is forced to scale up further and live with a lower resolution in order to be able to carry out the assigned mathematical task. This is why floating point is dominant when the computation values are small, and can handle large numbers when fixed point cannot. To make things more complicated for fixed point DSP developers, they have to handle overflow and truncation errors as well.

In Summary:

Advantages for Floating Point:

- Extremely large dynamic range
- Applications with intensive computations
- The smaller the number, the higher the precision and lower quantization noise
- Possible computation for large numbers, beyond fixed point's capability
- Simpler and faster development without worrying about overflow and truncation errors

Advantages For Fixed Point:

- Better resolution within a narrow range of 1<i<256 (7 out of 256 floating point exponent blocks)
- Simpler DSP silicon

Donny Chow is the Founder and Chief Engineer of Xilica. Following its motto of Passion Through Performance, Xilica manufactures class-leading networked and standalone DSPs for live and installed sound applications. www.xilica.com.