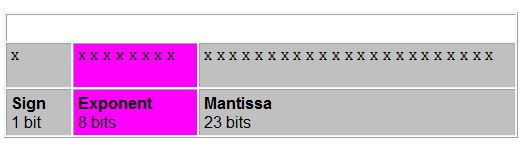
Priscilla N. Siquieros

February 14, 2013

IEEE Floating Point Number Single Precision



Source: http://www.cs.duke.edu/~raw/cps104/TWFNotes/floating.html

The first sign bit represents the sign of the number (either a 1 or a 0).

The next 8 bits determine the exponent stored in biased 127 (single precision) form, meaning it needs to represent both positive and negative exponents. That’s where the bias comes into play. The bias gets the stored exponent.

The last 23 bits will carry the mantissa normalized to be between 1 (less than or equal) and less then 2.

(Mantissa: is the number of digits. Always starts with a non-zero digit.)

With mantissa we have to limit the bits.

To convert an integer: first convert the integer part of the number generate the binary equivalent. Then convert the fractional part and generate the binary fraction, place together and normalize.

Truncation versus Rounding

Truncation: limits the no. of digits to the right of the decimal point and discard the least significant ones. Disadvantage is it undervalues the result and can cause Truncation error.

Rounding: If a lost digit is greater than a half 1 will be added to the least significant bit to round up. With rounding we have more accuracy and end up with fair error. But it requires up to two further arithmetic operations.

Otherwise we just leave as is and round down.