IEEE 754 FLOATING POINT NOTES AND EXAMPLE

POWER OF 2 TABLE

$2^{10} = 1024$
$2^{9} = 512$
$2^{8} = 256$
$2^{7} = 128$
$2^{6} = 64$
$2^{5} = 32$
$2^{4} = 16$
$2^{3} = 8$
$2^{2} = 4$
$2^{1} = 2$
$2^{0} = 1$
.
$2^{-1} = 0.5$
$2^{-2} = 0.25$
$2^{-3} = 0.125$
$2^{-4} = 0.0625$
$2^{-5} = 0.03125$
$2^{-6} = 0.015625$

The following is a step by step roadmap to go from a decimal number to its IEEE 754 32 bit floating point representation. The example will use -176.375 as an example

STEP 1: OBSERVE THE SIGN

For -176.375 the sign is negative. This means the first bit will be 1, if positive the first bit will be 0. Going forward the sign will be ignored, but then used again in the last step.

STEP 2: FROM DECIMAL TO BINARY

Transform the decimal to binary (ignoring the sign). To do this subtract the largest power of 2 relative to the decimal until you reach 0. Note that floating point is an approximation and cannot be perfectly represented – but the potential rounding error is very small.

Calculations

Use 2^{7}	Use 2^{5}	Use 2^{4}	Use 2^{-2}	Use 2^{-3}
176.375	48.375	16.375	0.375	0.125
- 128.000	- 32.000	- 16.000	- 0.250	- 0.125
48.375	16.375	0.375	0.125	0.000

As a sum : $2^{7} + 2^{5} + 2^{4} + 2^{-2} + 2^{-3} = 176.375$
As binary : 10110000.011 (Place a 1 in each position used)

STEP 3: MOVE TO SCIENTFIC NOTATION AND GET SIGNIFICANT

IEEE Floating points need to be in the format of $1.xxxxx * 2^{y}$. The significant is the xxxxx component (ignore the 1.) and has 23 bits. If your significant is shorter then 23 bits add trailing zeros.

$10110000.011 = 1.0110000011 * 2^{7}$

Significant = 01100000110000000000000 (had to add 13 trailing zeros)

STEP 4: CALCULATE EXPONENT IN BINARY

The exponent is represented by 8 bits (256 states) and is shifted by 127. In our example ($1.0110000011 * 2^{7}$) the exponent is 7. So we need to express 134 (from 7+127) in binary. Using the same technique as step 2:

As a sum : $128 + 4 + 2 = 134$
As a sum : $2^{7} + 2^{2} + 2^{1} = 134$
As binary : 10000110 (Place a 1 in each position used)

STEP 5: COMBINE SIGN, EXPONENT, AND SIGNIFICANT

The format is:

Sign (1 bit)	Exponent (8 bits)	Significant (23 bits)
1	10000110	01100000110000000000000

Or: 11000011001100000110000000000000 (Hex : 0xc3306000)