

Conversion Examples to help with Fractional Part of IEEE Floating Point

J.Wunderlich PhD

Adapted from: <http://sandbox.mc.edu/~bennet/cs110/flt/dtof.html>
(I added **bolds** and underlines for fractional parts, and *italics* and small fonts for exponents)

Convert -1313.**3125** to IEEE 32-bit floating point format.

The integral part is $1313_{10} = 10100100001_2$. The **fractional**:

| | | | | |
|-----------------------|--------------|-------|---|--|
| 0. <u>3125</u> | $\times 2 =$ | 0.625 | 0 | Generate 0 and continue. |
| 0.625 | $\times 2 =$ | 1.25 | 1 | Generate 1 and continue with the rest. |
| 0.25 | $\times 2 =$ | 0.5 | 0 | Generate 0 and continue. |
| 0.5 | $\times 2 =$ | 1.0 | 1 | Generate 1 and nothing remains. |

So $1313.\textbf{3125}_{10} = 10100100001.\textbf{0101}_2$.

Normalize: $10100100001.\textbf{0101}_2 = 1.0100100001\textbf{0101}_2 \times 2^{10}$.

Mantissa is 0100100001**0101**000000000, *exponent is* $10 + 127 = 137 = 10001001_2$, sign bit is 1.

So -1313.**3125** is $1\textbf{1000100101001000010101000000000} = \text{c4a42a00}_{16}$

Convert 0.**1015625** to IEEE 32-bit floating point format.

| | | | | |
|--------------------------|--------------|----------|---|--|
| 0. <u>1015625</u> | $\times 2 =$ | 0.203125 | 0 | Generate 0 and continue. |
| 0.203125 | $\times 2 =$ | 0.40625 | 0 | Generate 0 and continue. |
| 0.40625 | $\times 2 =$ | 0.8125 | 0 | Generate 0 and continue. |
| 0.8125 | $\times 2 =$ | 1.625 | 1 | Generate 1 and continue with the rest. |
| 0.625 | $\times 2 =$ | 1.25 | 1 | Generate 1 and continue with the rest. |
| 0.25 | $\times 2 =$ | 0.5 | 0 | Generate 0 and continue. |
| 0.5 | $\times 2 =$ | 1.0 | 1 | Generate 1 and nothing remains. |

So $0.\textbf{1015625}_{10} = 0.\textbf{0001101}_2$.

Normalize: $0.\textbf{0001101}_2 = \textbf{1.101}_2 \times 2^{-4}$.

Mantissa is **101**00000000000000000000, *exponent is* $-4 + 127 = 123 = 01111011_2$, sign bit is 0.

So 0.1015625 is $0\textbf{01111011101}00000000000000000000 = 3\text{dd}00000_{16}$

Convert 39887.**5625** to IEEE 32-bit floating point format.

The integral part is $39887_{10} = 10011011111001111_2$. The fractional:

| | | | | |
|-----------------------|--------------|-------|---|--|
| 0. <u>5625</u> | $\times 2 =$ | 1.125 | 1 | Generate 1 and continue with the rest. |
| 0.125 | $\times 2 =$ | 0.25 | 0 | Generate 0 and continue. |
| 0.25 | $\times 2 =$ | 0.5 | 0 | Generate 0 and continue. |
| 0.5 | $\times 2 =$ | 1.0 | 1 | Generate 1 and nothing remains. |

So $39887.\textbf{5625}_{10} = 10011011111001111.\textbf{1001}_2$.

Normalize: $10011011111001111.\textbf{1001}_2 = 1.001101111001111\textbf{1001}_2 \times 2^{15}$.

Mantissa is 001101111001111**1001**0000, *exponent is* $15 + 127 = 142 = 10001110_2$, sign bit is 0.

So 39887.**5625** is $0\textbf{10001110011011110011111001}0000 = 471\text{bcf90}_{16}$