

Tutorial 7

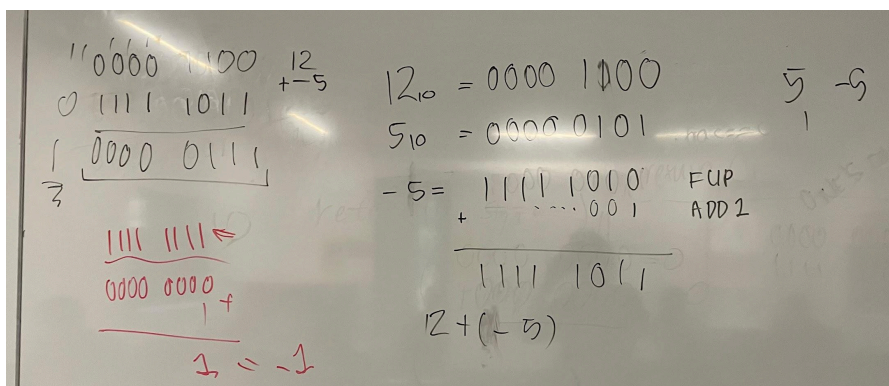
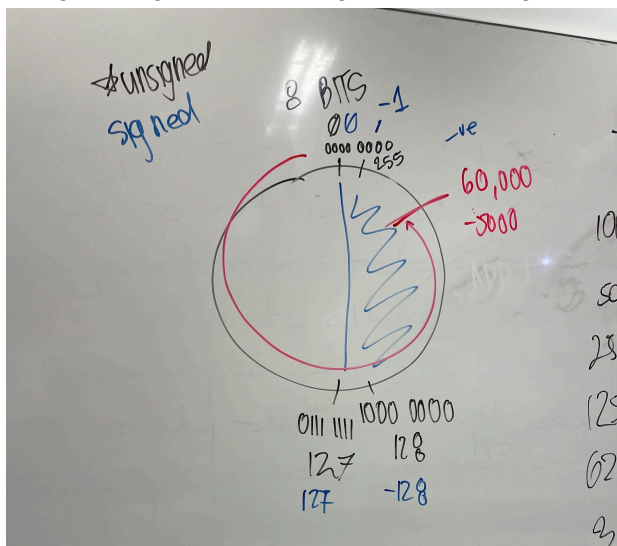
Useful links

- Two's complement explanation: <https://www.ralismark.xyz/posts/twos-complement>
- Floating point numbers wiki: https://en.wikipedia.org/wiki/Single-precision_floating-point_format
- Floating point nums converter: <https://www.h-schmidt.net/FloatConverter/IEEE754.html>

Negative Numbers - two's complement

- To convert from decimal \rightarrow two's complement: decimal \rightarrow binary IF negative flip and add 1
- To convert from two's complement \rightarrow decimal:
 - If positive, (if MSB or left most bit is 0) convert to decimal
 - If negative (MSB/left most bit is 1), FLIP AND ADD 1 to get positive number, then convert to decimal (don't forget that it's negative)

Range of signed and unsigned 8 bit integer



two's complement - flip and add one

Example: -1000 to 16 bit two's complement

-1000 → 16 bit two's complement = 0x03E8

	RES	REM	
1000 / 2	500	0	<div style="display: flex; align-items: center;"> <div style="margin-right: 10px;"> $\begin{array}{r} 0000\ 0011\ 1110\ 1000 \\ 1111\ 1100\ 0001\ 0111 \\ \hline 1111\ 1111\ 1100\ 0001\ 1000 \\ 0000\ 0101 \end{array}$ </div> <div> $\begin{array}{r} 0000\ 0011\ 1110\ 1000 \\ 1111\ 1100\ 0001\ 0111 \\ \hline 1111\ 1111\ 1100\ 0001\ 1000 \\ 0000\ 0101 \end{array}$ </div> </div>
500 / 2	250	0	
250 / 2	125	0	
125 / 2	62	1	
62 / 2	31	0	
31 / 2	15	1	
15 / 2	7	1	
7 / 2	3	1	
3 / 2	1	1	
1 / 2	0	1	*MSB

LSB ← 2⁰ 2¹ 2² 2³ 2⁴ 2⁵ 2⁶ 2⁷ 2⁸ 2⁹ 2¹⁰ 2¹¹ 2¹² 2¹³ 2¹⁴ 2¹⁵

FUP

Floats

Binary fractions

$\begin{array}{|c|c|c|c|c|c|c|} \hline 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & \dots \\ \hline 2^{-2} & 2^{-1} & 2^0 & 2^1 & 2^2 & 2^3 & 2^4 & 2^5 & 2^6 & \dots \\ \hline \end{array}$

$\begin{array}{|c|c|c|c|} \hline 16 & 3 & 0 & \\ \hline 10^3 & 10^2 & 10^1 & 10^0 \\ \hline 10^1 & 10^0 & 10^{-1} & 10^{-2} \\ \hline \end{array}$

$$0.25 \Rightarrow 0.01$$

$$0.255 \approx -\frac{127}{256}$$

$$(1 + \text{frac}) \times 2^0$$

$$\downarrow 2^{255}$$

$$\text{MIN. } (1 + \text{frac}) \times 1$$

Example decimal to float string

$a = 27$
 (110.6875)
 $27 = 1.6875 \times 2^4$

$EXP = 131$
 $FRAC = 0.6875 \times 2 = 1.375$
 $0.375 \times 2 = 0.75$
 $0.75 \times 2 = 1.5$
 $0.5 \times 2 = 1.0$

$0.6875_{10} = 0.1011_2$

$sign \times 2^{EXP-127} \times (1 + frac)$
 1 BIT 8 BITS 23 BITS

TOP MSB
 1 1 0 0 0 0 0 1 1 1 0 1 1 0 0 0 0 0 ...
 sign exp frac
 BOTTOM

• $\frac{1}{2}$ $\frac{1}{4}$ $\frac{1}{8}$ $\frac{1}{16}$ $\frac{1}{32}$...
 1 0 1 0 0 ...

Example float strings - dec

b. $= -0$
 $1 \ 00000000 \ 00000 \text{ sign} \times 2^{\text{EXP} - 127} \times (1 + \text{frac})$
 a. $0 \ 00000000 \ 0000 \dots$ $\text{EXP} = 00000000$ & $\text{Frac} = 0$
 $\rightarrow \pm 0$

$$+1 \times 2^{\text{EXP} - 127}$$

$$+1 \times 2^{-127} \times (1 + 0)$$

g. $0 \ 10010100 \ 1000 \dots$

$$+1 \times 2^{\text{EXP} - 127} \times (1 + 0.5)$$

$$= 2 \times (1.5) = 3 \text{ bits}$$

0.0 01111110

$$+1 \times 2^{\text{EXP} - 127} \times (1 + 0.999\dots)$$

$$\frac{1}{2} \times (1.99999) \approx 1$$

$$01111110 = 126$$

$$\text{nearby } 1 \approx 0.999997$$

Six_middle_bits (masking and shifting)

b. 1 $\overset{=-0}{00000000}$ 00000 sign $\times 2^{\text{EXP}-127} \times (1 + \text{frac})$ EXP=0000 0000 & FRAC=0
 a. 0 $\overset{S}{0} \overset{EXP}{0000} \overset{FRAC}{0000}$ 0000 - - - 8 BITS 23 BITS $\rightarrow \pm 0$
 $+1 \times 2^{\text{EXP}-127}$ 0000 0000 0000 0
 $0 \times 12345678 = 00010010001101000101010001010101$ SK middle
 SHIFT: $\gg 13$ put thing we want in right place
 $000...0111111 = \text{MASK} = 0x3F$
 MASK: val & mask
 \rightarrow get rid of bits we don't want
 $x_p = \text{EXP} - 127$
 $\rightarrow \text{NaN}$