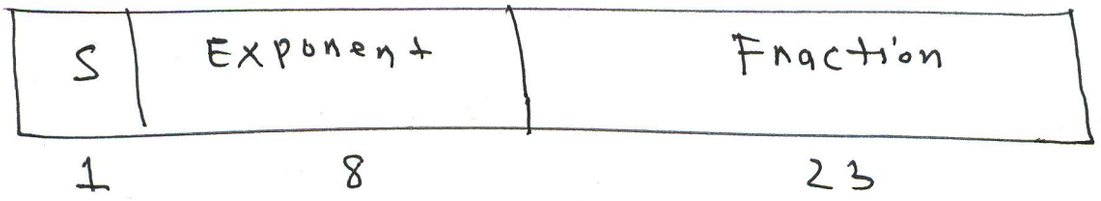
Example: Convert each of the following IEEE-754 floating point representation to decimal values.

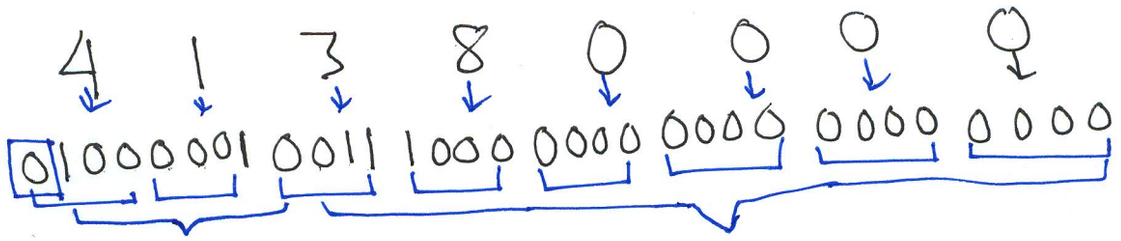
a. 0x41380000 → IEEE single precision format.

b. 0xC0E80000

$$(-1)^{S} \times (1 + \text{Fraction}) \times 2^{(\text{Exponent} - \text{Bias})}$$ Bias = 127.



0x



S = 0

Exponent = $100\,0001_2 = 130_{10}$

Fraction = $0.0111_2 = 0.4375_{10}$.

Decimal = $(-1)^{S} \times (1 + \text{Fraction}) \times 2^{\text{exponent - Bias}}$

$= (-1)^{0} \times (1 + 0.4375) \times 2^{130-127}$

$= 1 \times 1.4375 \times 8$

$= 11.5$.

IEEE 754 Double Precision Format (64 bit)

S	EXPONENT	FRACTION
1 bit	11 bits	52 bits

- S: Sign bit (0 → Non-negative, 1→ Negative)
- Exponent = Actual Exponent + Bias

  For double precision, Bias = 1023 ←
- Fraction : 52-bit fractions from normalized number

Example: Convert the decimal value 2.75 to IEEE-754 double precision format. Write your converted result in hexadecimal format.

2 1 0.5 0.25

2.75 = 10.11

Normalized value = 1.011×2^1

S = 0

Exponent = 1 + 1023 = 1024 = 100 00000000

Fraction = 0.011

Sign Exponent Fraction.

0	100 0000 0000	0110 0000 0000 0000 0000 0000 0000 0000
	4 0 0	6 0 0 0 0 0 0 0

0010 0000 0000 0000 0000

0 0 0 0 0

0X4006 0000 0000 0000 ←

The value of the floating-point number can be determined by the following expression:

$$(-1)^S \times (1 + \text{Fraction}) \times 2^{(\text{Exponent} - \text{Bias})}$$ Bias = 1023

Example: Convert the decimal value -4.25 to IEEE-754 double precision format. Write your converted result in hexadecimal format.

Example — Convert the following IEEE-754 floating point representation to decimal values.

0X4035 0000 0000 0000

Sign

0 100 0000 0011 0101 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000

Exponent

Sign = 0

Exponent = $100\ 0000\ 0011_2 = 1027_{10}$

Fraction = $0.0101_2 = 0.3125_{10}$

Decimd = $(-1)^0 \times (1 + 0.3125) \times 2^{1027-1023}$

$= 1 \times 1.3125 \times 2^4$

$= \boxed{21.}$

Floating Point Addition (Binary)

1. Align binary points:
   - Align binary point of the number with smaller exponent
2. Add significands:
3. Normalize result:
4. Round and renormalize if necessary:

NB. We assumed that significands can be only 4 bits or digits.

Problem: Perform binary floating-point addition and convert the resulting values to IEEE-754 single and IEEE-754 double precision format. Finally convert them to hexadecimal values. 0.5 + 0.4375

$0.5 = 0.1$

$= 1.0 \times 2^{-1}$

$0.4375 = 0.0111_2$

$0.4375 \times 2 = 0.875$

$0.875 \times 2 = 1.75$

$0.75 \times 2 = 1.5$

$0.5 \times 2 = 1.0$

$0.4375 = 1.11 \times 2^{-2}$

1. Align Binary Point)

$0.5 = 1.0 \times 2^{-1}$

$0.4375 = 0.111 \times 2^{-1}$

2. $1.0 + 0.111 = 1.111$

Value = 1.111×2^{-1}