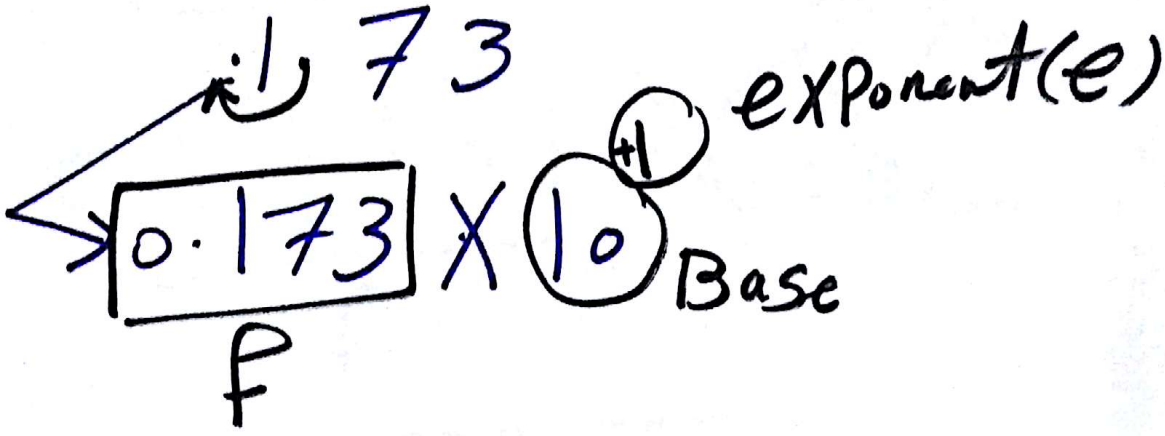
بسم الله الرحمن الرحيم

* Floating Point:
  - Signed magnitude. X
  - 2's Complement X
  - Excess Notation. X
  - IEEE ~~notation~~. ✓ < Code / Decode

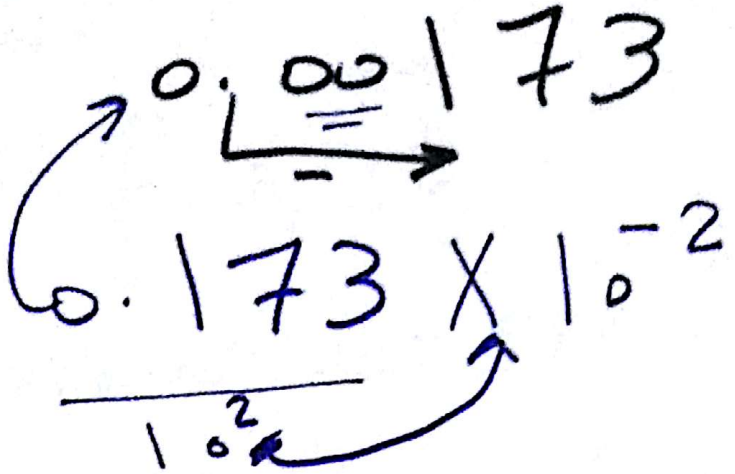
IEEE

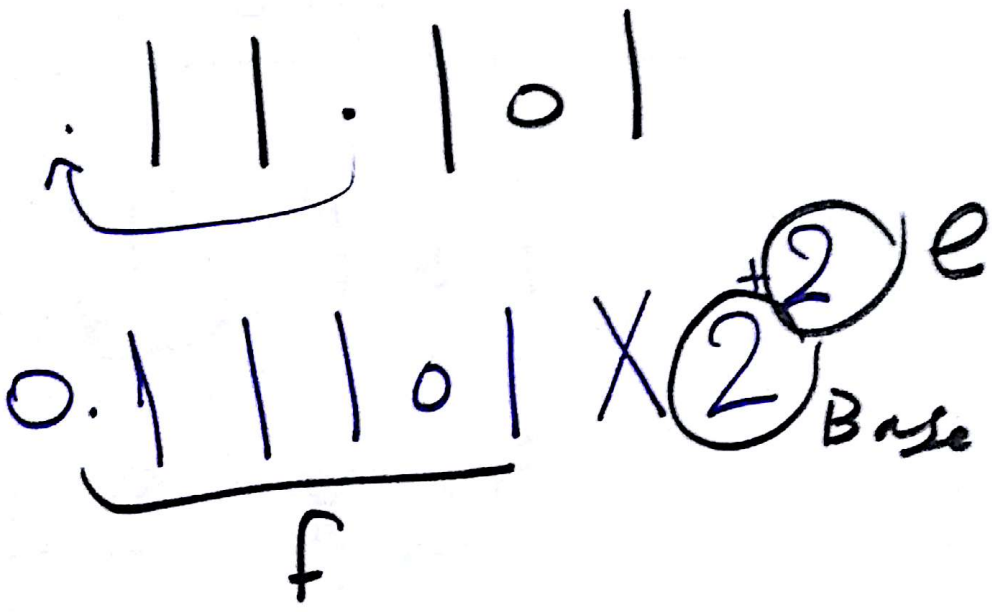


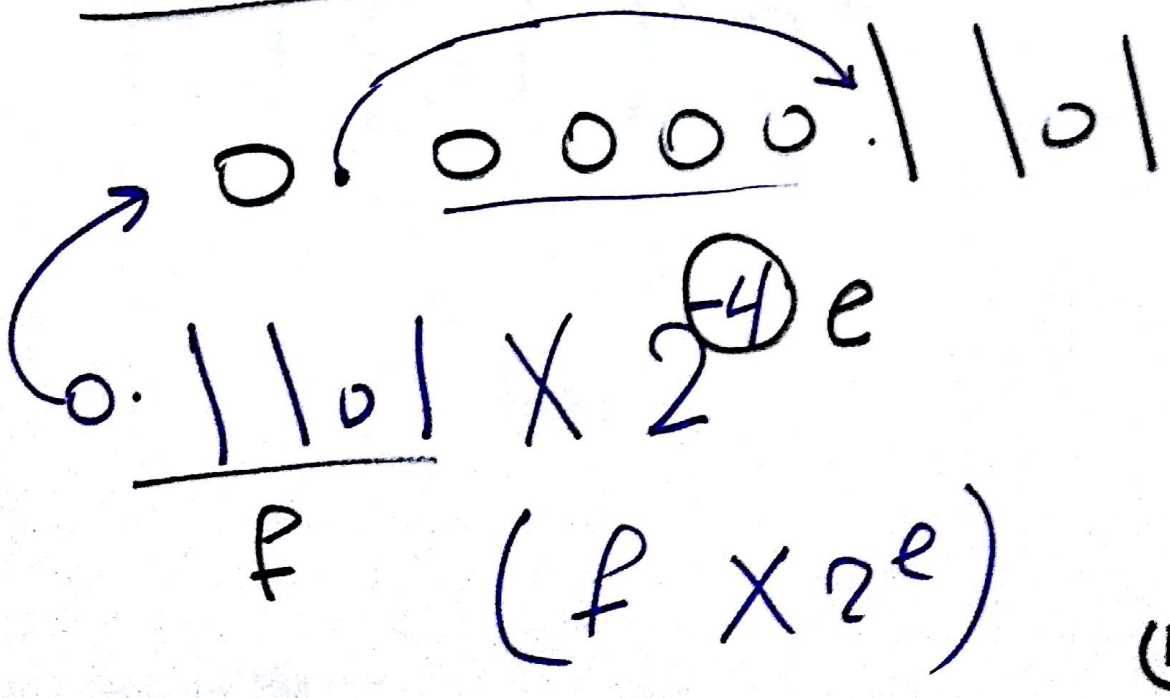
17.3

$0.173 \times 10^{+2}$

$(f) \times (Base)^{e}$


0.00173
0.173 X 10^-2
10^2


.11.101
0.11101 X 2 ±e
Base
f

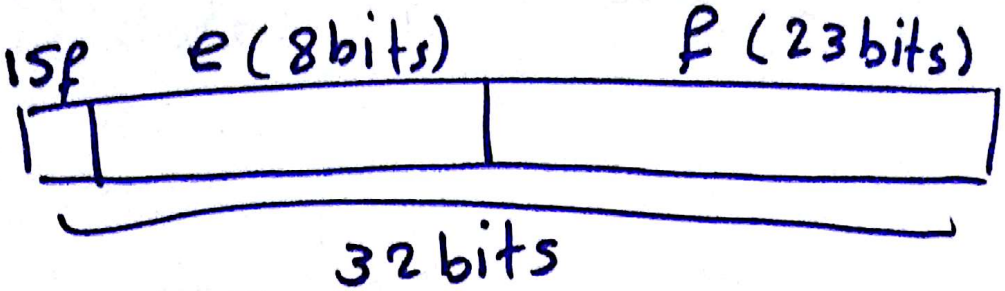

0.00001101
0.1101 X 2 -4 e
f
(f X 2^e)

* IEEE $1.f \times 2^e$

① $1.f \times 2^e$

② exponent $\Leftrightarrow$ excess 127

③



$1.1\,0\,1$

$1.f \times 2^e$

$1.1101 \times 2^{+1}$

* Code | Represent | Express

$(\_\_\_)_{10} \rightarrow (\quad\quad)_2$ in floating Point using IEEE

* Decode. $(\_\_\_\_\_)_2 \rightarrow (?)_{10}$

EX1

* Code $(+3.75)_{10}$ → as abinary Pattern of floating point format using the IEEE method.

① $(+3.75)_{10}$

⇓

$\oplus 1.1.11$

② $1.f \times 2^{e}$

$+1.111 \times 2^{+1}$

f

③ $f = +0.111$

$e = (+1)_{10}$

④ eXcess 127

$e = +1 + 127 = 128$

128	64	32	16	8	4	2	1
1	0	0	0	0	0	0	0

2 ÷ | 133

8 4 2+1

1 1

2 * | 0.75

0.25 | 1

0.5 | 1

eXcess

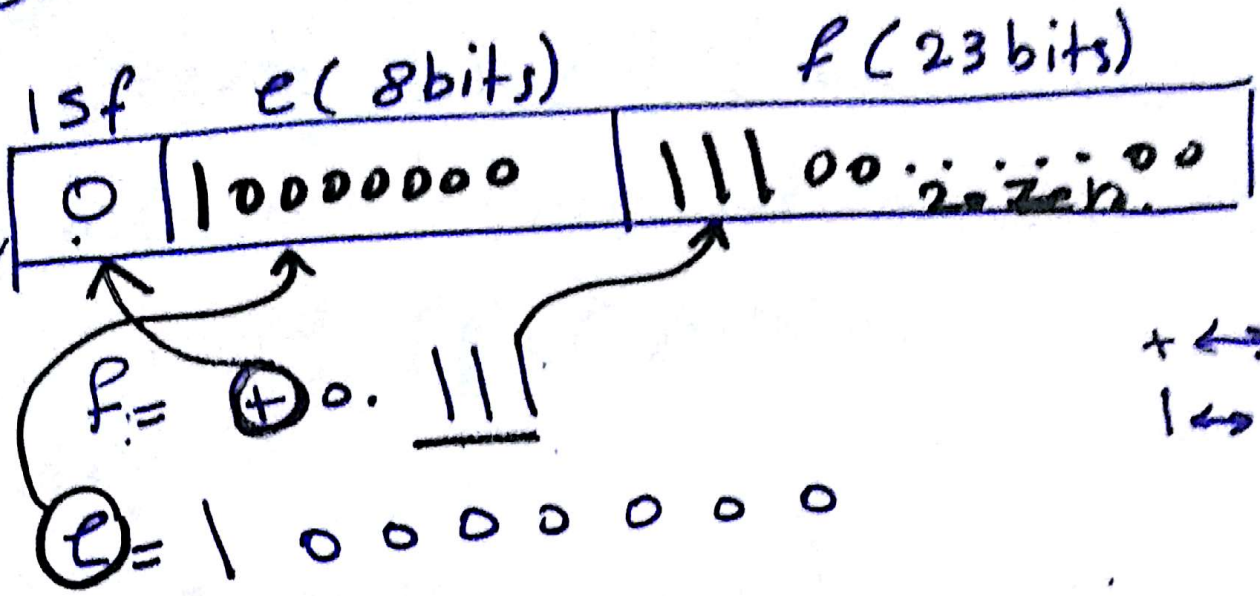
+not

-not

decimal

2 ÷ | 128

⑤



*The Floating Point RePresentation Code

$$(\underbrace{0\,|\,00}_{8\,4\,2\,1}\ \underbrace{0\,0\,0\,0}_{8\,4\,2\,1}\ \underbrace{0\,|\,1\,1\,1}_{8\,4\,2\,1}\ \underbrace{00\ldots00}_{\text{20 zero's}})_2$$

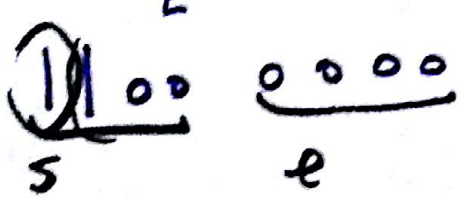
$$\rightarrow (4\ 0\ \ 7\ 0\,0\,0\,0\,0)_{16}$$

*Ex2:

Decode $(C0A4\,0000)_{16}$ into its equivalent decimal value. using IEEE method:.

Ans.

① $(C0A4\,0000)_{16}$



A 10
B 11
C 12
D 13
E 14
F 15

② 1 bit | e(8b) | f(23b)

1 | 1000000 | 01001 0....0

12

8 4 2 1

12 → 1 1 0 0

10 → 1 0 1 0

0 1 0 0

③ $f = -0.01001$

$e = 128\,64\,32\,16\,8\,4\,2\,1$ | $0\,0\,0\,0\,0\,0\,0\,0$ | -127

$= 129 - 127$

$e = (+2)_{10}$

exces $()_2$

(-not 127

④ decimal

④ $1.f \times 2^{e}$

$-1.01001 \times 2^{+2}$

⑤

2^{2}	2^{1}	2^{0}	.	2^{-1}	2^{-2}	2^{-3}
−	1	0	.	0	0	1

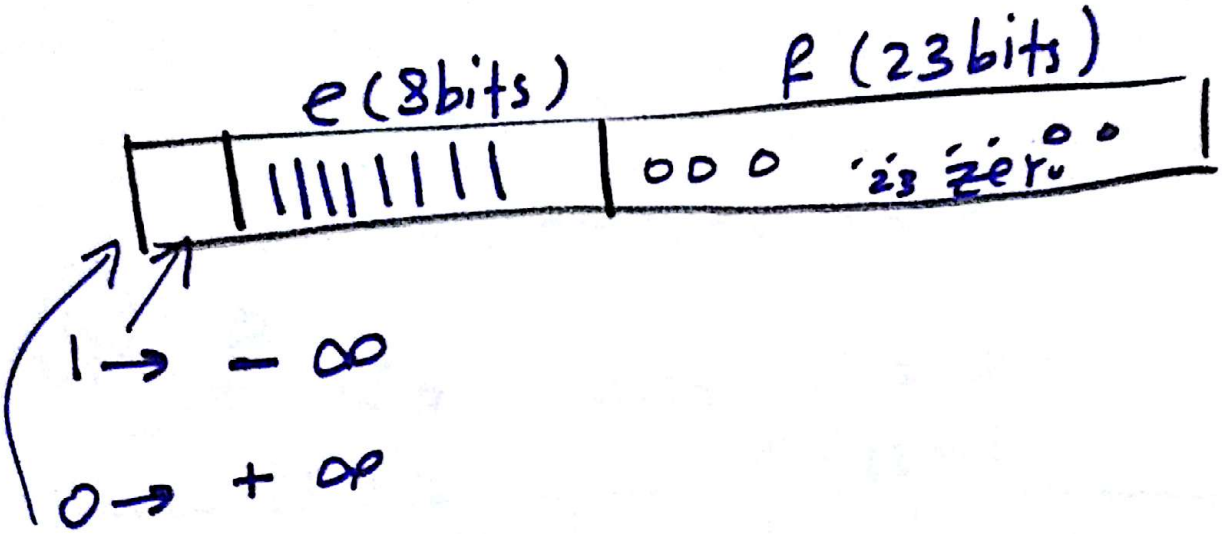
⑥ (-5.125)

* Special cases in IEEE method

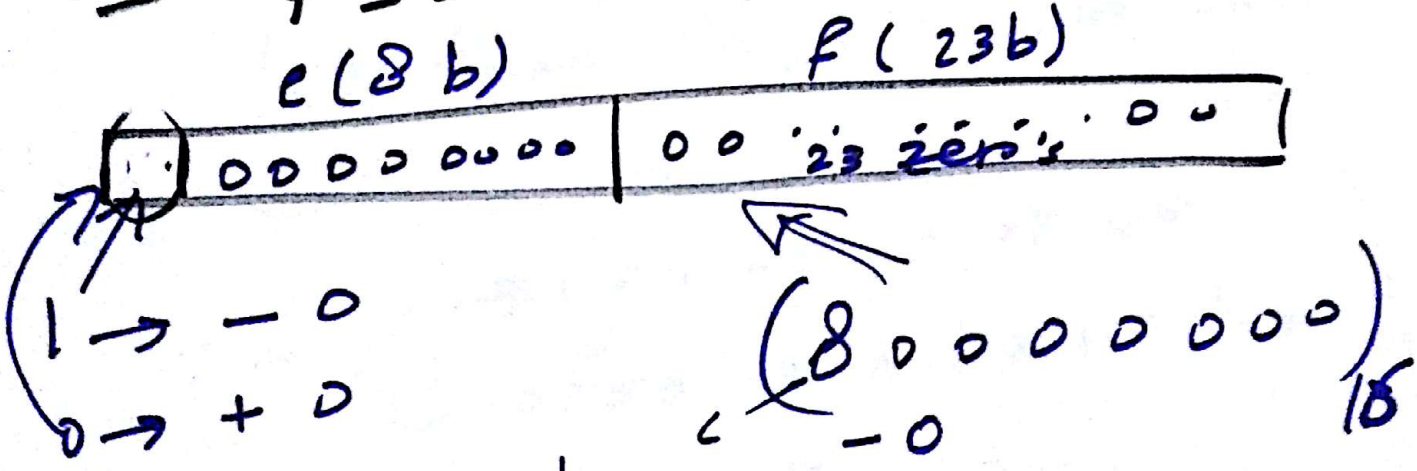
Decode

[1] $f = 0$ and $e = 1$

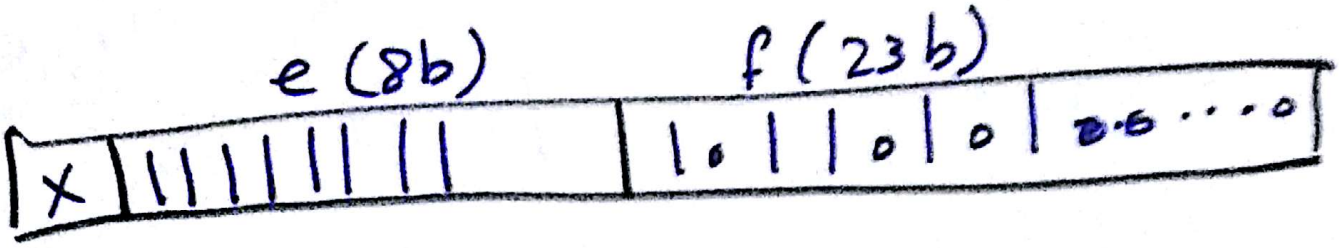
23 bits ← f ; e → 8 bits



[2] $f = 0$ and $e = 0$

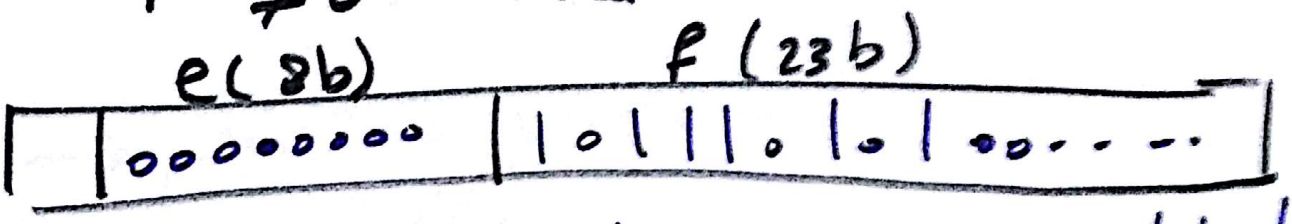


(3) $f \neq 0$ and $e = 1$



→ Not a Number (NaN)

(4) $f \neq 0$ and $e = 0$



⇒ the number has a value that is less than the allowed minimum value.

$e = 1$
- $f = 0$ → $-\infty$ or $+\infty$
- $f \neq 0$ → Not a Number

$e = 0$
- $f \neq 0$ → the number . . . . . .
- $f = 0$ → -0 or $+0$