1)

a) $18.25_{10} \longrightarrow 18_{10} + 0.25_{10}$

$18 \mid 2$
$18 \mid 9 \mid 2$
$0 \quad 8 \mid 4 \mid 2$
$\quad 1 \quad 4 \mid 2 \mid 2$
$\quad \quad 0 \quad 2 \mid 1 \mid 2$
$\quad \quad \quad 0 \quad 1 \mid 0$
$\quad \quad \quad \quad 1$

$18 = 9 \times 2 + 0$
$9 = 4 \times 2 + 1$
$4 = 2 \times 2 + 0$
$2 = 1 \times 2 + 0$
$1 = 0 \times 2 + 1$

$0.25 \times 2 = 0.5 \quad 0$
$0.5 \times 2 = 1 \quad 1$

$18.25_{10} = 10010.01_2$

b) $57.9_{10} \longrightarrow 57_{10} + 0.9_{10}$

$57 \mid 2$
$56 \mid 28 \mid 2$
$1 \quad 28 \mid 14 \mid 2$
$\quad \quad 0 \quad 14 \mid 7 \mid 2$
$\quad \quad \quad 0 \quad 6 \mid 3 \mid 2$
$\quad \quad \quad \quad 1 \quad 2 \mid 1 \mid 2$
$\quad \quad \quad \quad \quad 1 \quad 0 \mid 0$
$\quad \quad \quad \quad \quad \quad 1$

$57 = 28 \times 2 + 1$
$28 = 14 \times 2 + 0$
$14 = 7 \times 2 + 0$
$7 = 3 \times 2 + 1$
$3 = 1 \times 2 + 1$
$1 = 0 \times 2 + 1$

$0.9 \times 2 = 0.8 + 1$
$0.8 \times 2 = 0.6 + 1$
$0.6 \times 2 = 0.2 + 1$
$0.2 \times 2 = 0.4 + 0$
$0.4 \times 2 = 0.8 + 0$
$0.8 \times 2 = 0.6 + 1$
$0.6 \times 2 = 0.2 + 1$
$0.2 \times 2 = 0.4 + 0$
$0.4 \times 2 = 0.8 + 0$

$57.9_{10} = 111001.1\overline{1100}_2$

1)

c) 100101.11_2

$2^5 \times 1 = 32$

$2^4 \times 0 = 0$

$2^3 \times 0 = 0$

$2^2 \times 1 = 4$

$2^1 \times 0 = 1$

$+$

37

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

$2^{-2} \times 1 = 0.25$

$+$

0.75

$100101.11_2 = 37.75_{10}$

d) $1101.1\overline{10}_2$

$2^3 \times 1 = 8$

$2^2 \times 1 = 4$

$2^1 \times 0 = 0$

$2^0 \times 1 = 1$

$+$

13

$0.1\overline{10} = x$

$1.1\overline{10} = 2x \ (2^3 x)$

$110.1\overline{10} = 8x \ (2^3 x)$

$x(2^3 - 2) = 110.1\overline{10} - 1.1\overline{10}$

$6x = 5$

$x = \frac{5}{6}$

$x = 0.8\overline{3}$

$1101.1\overline{10}_2 = 13.8\overline{3}_{10}$

$110.1\overline{10}$

$-\ 1.1\overline{10}$

$101.000 \rightarrow 101$

$2^2 \times 1 = 4$

$2^1 \times 0 = 0$

$2^0 \times 1 = 1$

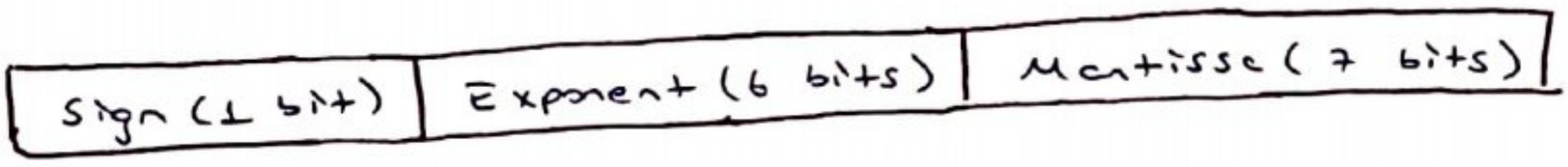
$+$

5

2)

Given

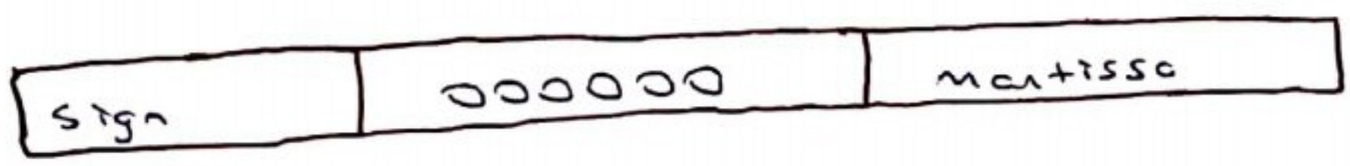
IEEE - 754 where



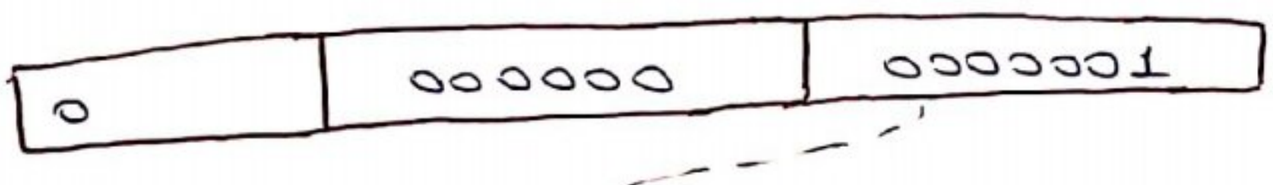
$$Bias = 2^{exp-1} - 1 = 2^{6-1} - 1 = 31$$

$$Le = 1 - Bias = 1 - 31 = -30$$

For the denormalized number, we have to use -30 as the exponent value. The encoding will be as follows.

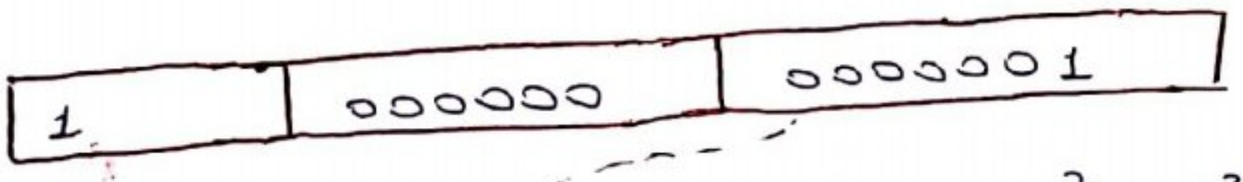


a) Smallest positive subnormal number will have following encoding



$$Value = 0.0000001 \times 2^{-30} = 2^{-7} \times 2^{-30} = \frac{1}{128} \times 2^{-30}$$

b) Largest negative subnormal number will have following encoding



$$Value = -0.0000001 \times 2^{-30} = -2^{-7} \times 2^{-30} = \frac{-1}{128} \times 2^{-30}$$

2)

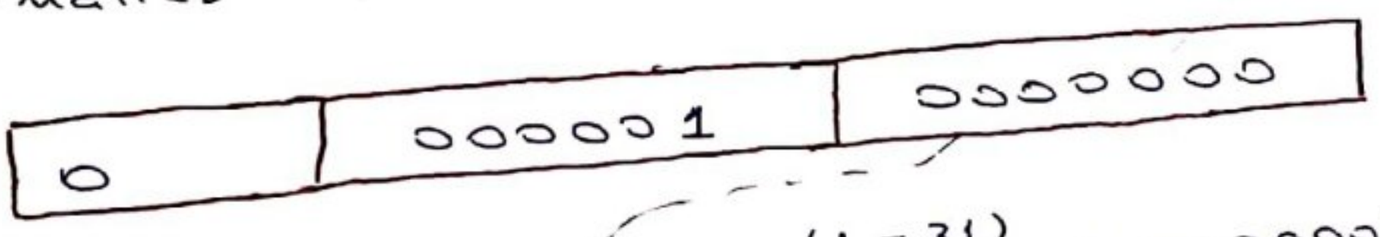
c) Largest positive subnormal number will have following encoding:

$$\text{Value} = 0.1111111 \times 2^{-30} = \left(1 - \frac{1}{2^7}\right) \times 2^{-30} = \frac{127}{128} \times 2^{-30}$$

d) Smallest negative subnormal number will have following encoding.

$$\text{Value} = -0.1111111 \times 2^{-30} = -\left(1 - \frac{1}{2^7}\right) \times 2^{-30} = -\frac{127}{128} \times 2^{-30}$$

e) Smallest positive normalized number will have following encoding:



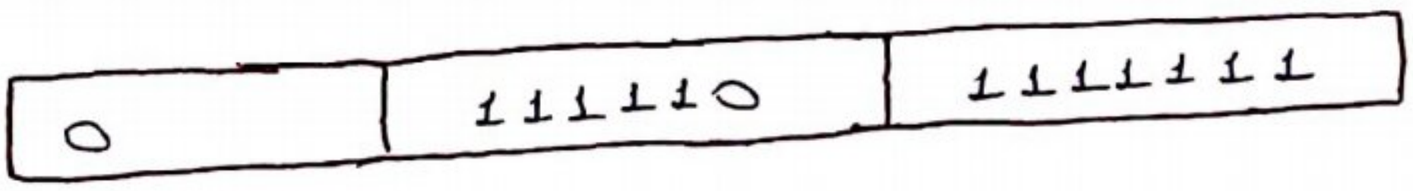
$$\text{Value} = 1.0000000 \times 2^{(1-31)} = 1.0000000 \times 2^{-30} = 2^{-30}$$

f) Largest negative normalized number will have following encoding:

$$\text{Value} = -1.0000000 \times 2^{1-31} = -1.0000000 \times 2^{-30} = -2^{-30}$$

2)

g) Largest positive normalized number will have following encoding:



Value = $1.1111111 \times 2^{62-31} = 1.1111111 \times 2^{31} = (2 - 2^{-7}) \times 2^{31}$

$= \left(2 - \frac{1}{128}\right) \times 2^{31}$

$= \frac{255}{128} \times 2^{31}$

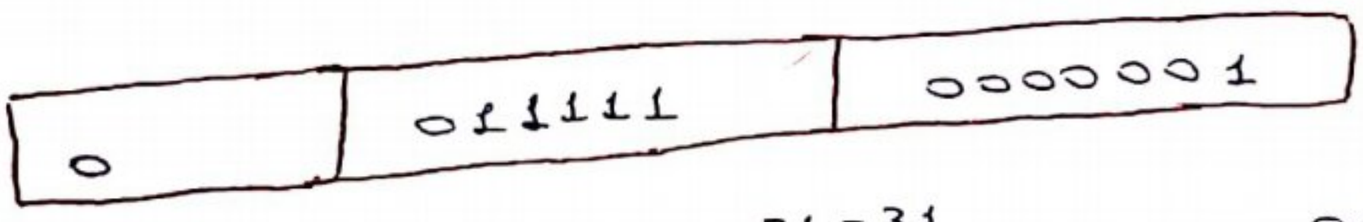
h) Smallest negative normalized number will have following encoding:

Value = $-1.1111111 \times 2^{62-31} = -1.1111111 \times 2^{31} = -(2 - 2^{-7}) \times 2^{31}$

$= -\left(2 - \frac{1}{128}\right) \times 2^{31}$

$= -\frac{255}{128} \times 2^{31}$

i) Smallest number greater than 1 will have following encoding.



Value = $1.0000001 \times 2^{31-31} = 1.0000001 \times 2^{0} = (1 + 2^{-7}) \times 2^{0}$

$= \left(1 + \frac{1}{128}\right) \times 2^{0}$

$= \frac{129}{128} \times 2^{0}$

2)

j) Largest number smaller than 1 will have following encoding

Value = $1.11111111 \times 2^{30-31} = 1.11111111 \times 2^{-1} = (2 - 2^{-7}) \times 2^{-1}$

$= \left(2 - \frac{1}{128}\right) \times 2^{-1}$

$= \frac{255}{128} \times 2^{-1}$

2)

	Sign	Exponent	Fraction	value	Comment
a)	0	000000	0000001	$\frac{1}{128} \times 2^{-30}$	Smallest positive subnormal
b)	1	000000	0000001	$\frac{-1}{128} \times 2^{-30}$	Largest negative subnormal
c)	0	000000	1111111	$\frac{127}{128} \times 2^{-30}$	Largest positive subnormal
d)	1	000000	1111111	$\frac{-127}{128} \times 2^{-30}$	Smallest negative subnormal
e)	0	000001	0000000	2^{-30}	Smallest positive normalized
f)	1	000001	0000000	-2^{-30}	Largest negative normalized
g)	0	111110	1111111	$\frac{255}{128} \times 2^{31}$	Largest positive normalized
h)	1	111110	1111111	$\frac{-255}{128} \times 2^{+31}$	Smallest negative normalized
i)	1	011111	0000001	$\frac{129}{128} \times 2^{0}$	Smallest number greater than 1
j)	0	011110	1111111	$\frac{255}{128} \times 2^{-1}$	Largest number smaller than 1

CamScanner ile tarar

3)

$7.2_{10} \longrightarrow 7_{10} + 0.2_{10}$

$7 = 3.2 + 1$
$3 = 1.2 + 1$
$1 = 0.2 + 1$

$0.2 \times 2 = 0.4 + 0$
$0.4 \times 2 = 0.8 + 0$
$0.8 \times 2 = 0.6 + 1$
$0.6 \times 2 = 0.2 + 1$
$0.2 \times 2 = 0.4 + 0$
$0.4 \times 2 = 0.8 + 0$
$0.8 \times 2 = 0.6 + 1$
$0.6 \times 2 = 0.2 + 1$

$7.2_{10} = 111.\overline{0011}_2$

Question says 1 bit for sign
2 bits for exponent
3 bits for significand parts.

$1,\underbrace{\underbrace{110}_{b_1 b_2 b_3}0110011\ldots}_{\text{Mantissa}} \times 2^2$

$fl(7,2) = 7,2 - (0,\overline{0110} \cdot 2^{-3} \cdot 2^2)$

$= 7,2 - \left(\frac{0,\overline{0110}}{2}\right)$

$= \frac{14,4 - 0,\overline{0110}}{2}$

$= \frac{1110,\overline{0110} - 0,\overline{0110}}{2}$

$= \frac{1110}{2}$

$= \frac{14}{2}$

$= 7$

$fl(7,2) = 7$

$14,4 \longrightarrow 14 + 0.4$

$14 = 7.2 + 0$
$7 = 3.2 + 1$
$3 = 1.2 + 1$
$1 = 0.2 + 1$

$0.4 \times 2 = 0.8 + 0$
$0.8 \times 2 = 0.6 + 1$
$0.6 \times 2 = 0.2 + 1$
$0.2 \times 2 = 0.4 + 0$
$0.4 \times 2 = 0.8 + 0$
$0.8 \times 2 = 0.6 + 1$
$0.6 \times 2 = 0.2 + 1$
$0.2 \times 2 = 0.4 + 0$

3)

a) Absolute Error :

The absolute error is the difference between

the number x and its finite representation.

$$\text{Absolute Error} = |x - fl(x)|$$

$$= |7,2 - 7|$$

$$= 0,2$$

b) Relative Error :

The relative error is the ratio of the absolute

error and the number x

$$\text{Relative Error} = \frac{|x - fl(x)|}{|x|}$$

$$= \frac{17,2 - 71}{17,21}$$

$$= \frac{0,2}{7,2} = 0,02\overline{7}$$