Assignment-1 [CSE330]

Name: Ishtiaq Ahmed

S_ID: 21301289

Sec: 08 [SADF]

Ans. to the Q. no 1(a)

the standard form $F = (0.d_1d_2d_3d_4d_5)_2 \times 2^e$

$\because d_1 \neq 0$

$\therefore$ largest significand $= (0.11111)_2$ $\because d_1 \neq 0$

$\therefore (0.11111)_2 = 1\times2^{-1} + 1\times2^{-2} + 1\times2^{-3} + 1\times2^{-4} + 1\times2^{-5}$

$= \frac{31}{32}$

Maximum exponent, $e = 5$

$\therefore$ maximum number $= \frac{31}{32} \times 2^5 = 31$

the IEEE Normalized form, $F = (0.1d_2d_3d_4d_5)_2 \times 2^e$

Largest significant $= (0.11111)_2$

$\therefore (0.11111)_2 = \frac{31}{32}$

maximum exponent, $e = 5$

$\therefore$ maximum number $= \frac{31}{32} \times 2^5 = 31$

the IEEE Denormalized form, $F = (0.b11111)_2 \times 2^{-2}$

Here leading bit $= 0$.

$\therefore$ largest possible mantissa $= 011111$

$\therefore (0.011111)_2 = 0\times2^{-1} + 1\times2^{-2} + 1\times2^{-3} + 1\times2^{-4} + 1\times2^{-5} + 1\times2^{-6}$

$= \frac{31}{64}$ $\therefore F = \frac{31}{64} \times 2^{-2} = 0.11718$

Ans. to the Q. no. 1(b)

For standard form:

Smallest mantissa $= (0.10000)_2 = 0.5$

smallest exponent, $e = -2$

$\therefore F = 0.5 \times 2^{-2} = 0.125$

For IEEE normalized form:

mantissa is always $= 0.10000 = 0.5$

minimum, $e = -2$

$\therefore F = 0.5 \times 2^{-2} = 0.125$

For IEEE Denormalized form:

Minimum positive mantissa $= 00001$

$\therefore (0.00001)_2 = 0.03125$

exponend fixed te $e = -2$

$\therefore F = 0.03125 \times 2^{-2}$

$= 0.03125 \times 0.25$

$= 0.0078125$

Ans. to the Q. no. 1(c)

In floating point representations, the limits of representable numbers depend on the range of the exponent and the form of the mantissa. When a number is too small to represent, it cause underflow and is considered zero. When a number is too large, it cause overflow and it treated as $\pm\infty$.

We analyze all three forms are bellows:

for IEEE Denormalized form:

smallest positive number ($min > 0$)

~~$F = 0.0078125$ ∴ from~~

~~So, any~~ Mantissa $= (0.00001)_2 = (0.03125)_{10}$

$e = -2$

$\therefore F = 0.03125 \times 2^{-2} = 0.0078125$

So, any positive number less than 0.0078125 cause underflow and is treated as zero.

Maximum value (before overflow):

$$\text{Mantissa} = (0.11111)_2 = (0.96875)_{10}$$

$$\therefore P = 0.96875 \times 2^{-2} = 0.24218$$

Any number > 0.24218 cannot be represented in denormalized form.

~~On ther~~

Again, Standard and IEEE Normalized form:

Smallest positive number.

$\therefore$ mantissa $= (0.10000)_2 = (0.5)_{10}$

$$e = -2$$

$$P = 0.5 \times 2^{-2} = 0.125$$

Any number smaller than 0.125 is not representable in normalized form and must be represend in denormalized form or else become zero.

Maximum number (before overflow):

Mantissa $= (0.11111)_2 = 0.96875$

$e = 5$

$\therefore F = 0.96875 \times 2^5 = 31.$

$\therefore$ Any number greater than 31 causes overflow and is treated as $\pm\infty$.

Combining all three representations:

$[-31.0, -0.0078125] \cup \{0\} \cup [0.0078125, 31]$.

Ans. to the Q. no. 2(a)

given that, $x = (10.3027)_{10}$

$2 \mid 10$
$2 \mid 5 - 0$
$2 \mid 2 - 1$
$1 \mid 1 - 0$
$1 - 1$ ↑

$\therefore (10)_{10} = (1010)_2$

$0.3027 \times 2 = 0.60 - 0$
$0.60 \times 2 = 1.21 - 1$
$0.21 \times 2 = 0.42 - 0$
$0.42 \times 2 = 0.84 - 0$
$0.84 \times 2 = 1.68 - 1$
$0.68 \times 2 = 1.37 - 1$

$\therefore (0.3027)_{10} \approx (0.010011)_2$

$\therefore x = (10.3027)_2 = (1010.010011)_2$

Ans. to the Q. no. 2(b)

Normalized the binary number.

$$1010.010011 = 1.01\,0010011 \times 2^3$$

∵ [We shifted the binary point only 4 significant digit in mantissa]

$$\therefore \text{Mantissa} = (1.0100)_2 \quad \because m = 4$$

So, $$fl(x) = (1.0100)_2 \times 2^3$$

Ans. to the Q. no. 2(c)

Convert mantissa to decimal:

$$(1.0100)_2 = 1 + 0 + 0.25 + 0 + 0 = 1.25$$

$$\therefore fl(x) = 1.25 \times 2^3 = 10.$$

$$\therefore \text{Rounding Error} = |10.3027 - 10| = 0.3027$$

$$\therefore \text{Maximum Scale Invariant Round Error} = \frac{|10.3027 - 10|}{10.3027} = 0.02938..$$