BRAC University (Department of Computer Science and Engineering)

CSE 330 (Numerical Methods) for Spring 2023 Semester

Quiz 1

Student ID: 20101534

Section: 10

Name: Md. Danial Islam

Full Marks: 15

Duration: 25 minutes

13

1.[CO1] If β=2, **fraction**=2 bit, **exponent**=2 bit, what will be the **non-negative smallest** and **largest** possible number that can be generated using the **normalized** form of the floating point representation? {Assume lowest and highest exponent is **reserved** for zero and Inf respectively} (3 marks)

Ans: $e = [0,3]$ as reserved then $e = [1,2]$

smallest $= (1.00)_2 \times 2^1$ highest $= (1.11)_2 \times 2^2$

2. [CO1] If β=2, m=5, $-100 \le e \le 100$, what will be the **machine epsilon(ϵ_M)** number using the **denormalized form** of the floating point representation? (1 mark)

Ans: $\epsilon_m = \frac{1}{2}\beta^{-m} = \frac{1}{2} \times 2^{-5} = \frac{1}{64} = 2^{-6}$

$= (0.000001)_2 \times 2^0 = (0.100000)_2 \times 2^{-5}$

3. [CO1] If β=2, m=4, $-3 \le e \le 3$, how many floating point numbers can be generated using the **general convention(1st convention)** of the floating point representation? (2 marks)

Ans: $\beta^{m-1} \times$ count of exponent $\Rightarrow 2^{4-1} \times 7 \Rightarrow 56$

4. [CO1] How many **significant figures** are there in the following numbers? (2 marks)

a) **000.0000200700.** **Ans:** 6

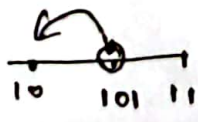
b) **0000060008000000000000** **Ans:** 5

5. [CO2] If $y = \frac{5}{16}$, then find **fl(y)** where **mantissa=2 bit**, $-4 \le e \le 4$. {Use general convention} (3 marks)

Ans: $y = \frac{5}{16} = \frac{4+1}{16} = 2^{-2} + 2^{-4}$

$= (0.0101)_2 \times 2^0 \Rightarrow (0.101)_2 \times 2^{-1}$

$\Rightarrow (0.10)_2 \times 2^{-1}$ (Ans)



6. [CO3] If x=10/16 and y=14/16, find fl(xy) where mantissa=4 bit. Also check whether xy=fl(xy). If not, find the rounding error of the product of these two numbers. (4 marks)

Ans:

$x = \frac{10}{16}$

$= \frac{2+8}{16}$

$= 2^{-3} + 2^{-1}$

$= (0.1010)_2 \times 2^0$

$y = \frac{14}{16}$

$= \frac{2+4+8}{16} = 2^{-3} + 2^{-2} + 2^{-1}$

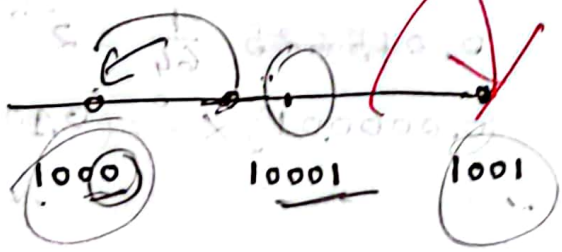
$= (0.1110)_2 \times 2^0$

$xy = \frac{10}{16} \times \frac{14}{16} = \frac{35}{64} = \frac{1+2+32}{64} = 2^{-6} + 2^{-5} + 2^{-1}$

$= (0.100011)_2 \times 2^0$

$fl(xy) = (0.1000)_2 \times 2^0$

$= 2^{-1} \Rightarrow \frac{1}{2}$



(5) $\frac{xy - fl(xy)}{xy}$

$\Rightarrow \frac{\frac{35}{64} - \frac{1}{2}}{\frac{35}{64}}$

$= \frac{3}{35}$ Ans

$= 0.0857$ (Ans