Section 3

1

A

CSE330 Fall 2022

Total time: 30 min | Quiz 1 | Set- A

Name:	ID:	Section:

For part (a), (b), (c) and (d), assume that the system has the following parameters: $\beta = 2$, $m = 3$, $e_{min} = -1$ and $e_{max} = 2$.

(a) (4 marks) Consider $x = 6/8$, and $y = 7/8$. Now convert these numbers into floating point fl(x) and fl(y).

(b) (6 marks) Compute x*y, then evaluate fl(x)*fl(y) and state whether any rounding error occurs or not.

(c) (2 mark) How many non-negative numbers are perfectly representable by this system in Lecture Notes Convention 1?

(d) (2 marks) Find the minimum positive number represented by the system in Convention 1.

For part (e), assume that the system follows the denormalized convention and has the following parameters: $\beta = 2$, 1 bit for sign, **3 bit for exponent**, and 6 bit for the mantissa.

(e) (2 + 4 marks) Find the value for e_{min} and e_{max}. Then assume e_{min} is reserved for zero and e_{max} is reserved for infinity, now calculate the highest possible and lowest possible non-negative number that can be represented by the system.

(a)

$x = \frac{6}{8} \Rightarrow fl(x) = (0.110)_2 \times 2^0 = \frac{6}{8}$

$y = \frac{7}{8} \Rightarrow fl(y) = (0.111)_2 \times 2^0 = \frac{7}{8}$

(b) $fl(xy) = fl(x) \times fl(y) = \frac{6}{8} \times \frac{7}{8} = \frac{42}{64} = \frac{32}{64} + \frac{8}{64} + \frac{2}{64} = \frac{1}{2} + \frac{1}{8} + \frac{1}{32} = (0.10101)_2 \times 2^0$

$0.101 < 0.10101 < 0.110$

↳ since $m = 3$; the 4th digit is 0.

$\therefore fl(xy) = (0.101)_2 \times 2^0 = \frac{5}{8}$ and $xy = \frac{21}{32}$

$|fl(xy) - xy| = \frac{1}{32}$

$\therefore$ rounding error occurs.

(c) $f = \pm (0.\underbrace{d_1 d_2 \cdots}_{m}) \times \beta^e$

$\therefore f = + (0.\underbrace{1\square\square}_{m=3}) \times \beta^{(-1,2)}$ } total = 16

(d) $+ (0.100)_2 \times 2^{-1}$

(e) 3 bit for e, So, $2^3 \rightarrow \square\square\square$

$e_{min} = 0$; $e_{max} = 7$

highest possible non-negative $= + (0.1111111)_2 \times 2^{(6)}$ → as 7 reserved for ∞

lowest possible non-negative $= + (0.1000000)_2 \times 2^{(1)}$ → as 0 reserved for zero

Section 3

1

B

CSE330 Fall 2022

Total time: 30 min

Quiz 1

Set- B

Name:	ID:	Section:

For part (a), (b), (c) and (d), assume that the system has the following parameters: $\beta = 2$, $m = 3$, $e_{min} = -1$ and $e_{max} = 2$.

(a) (4 marks) Consider x = 4/8, and y = 5/8. Now convert these numbers into floating points fl(x) and fl(y).

(b) (6 marks) Compute x*y, then evaluate fl(x)*fl(y) and state whether any rounding error occurs or not.

(c) (2 mark) How many non-negative numbers are perfectly representable by this system in the Denormalized form?

(d) (2 marks) Find the minimum positive number represented by the system in the Denormalized form.

For part (e), assume that the system follows the denormalized convention and has the following parameters: $\beta = 2$, 1 bit for sign, **4 bit for exponent**, and 5 bit for the mantissa.

(e) (2 + 4 marks) Find the value for e_{min} and e_{max}. Then assume e_{min} is reserved for zero and e_{max} is reserved for infinity, now calculate the highest possible and lowest possible non-negative number that can be represented by the system.

(a) $x = \frac{4}{8} \Rightarrow fl(x) = (0.100)_2 \times 2^0$; $y = \frac{5}{8} \Rightarrow fl(y) = (0.101)_2 \times 2^0$

(b) $fl(xy) = fl(x) \times fl(y) = \frac{4}{8} \times \frac{5}{8} = \frac{5}{16} = \frac{4}{16} + \frac{1}{16} = \frac{1}{4} + \frac{1}{16} = (0.0101)_2 \times 2^0$

$= (0.1010\ldots)_2 \times 2^{-1}$

$0.101 < 0.1010 < 0.1\overline{01}0$

$\hookrightarrow m = 3$; 4th digit = 0

$\therefore fl(xy) = (0.101)_2 \times 2^{-1} = \frac{5}{16}$] no rounding error occurs.

and $xy = \frac{5}{16}$

(c) $f = \pm (0.1 \underbrace{d_1 d_2 \ldots}_{m}) \times \beta^e$

$\therefore f = \pm (0.1 \square\square\square) \times \beta^{(-1,2)}$ } total = 32

(d) $+(0.1000)_2 \times 2^{-1}$

(e) 4 bits for exponent. So, $e^4 \rightarrow \square\square\square\square$

$e_{min} = 0$; $e_{max} = 15$

highest possible non-negative = $+(0.1 \underbrace{11111}_{m})_2 \times 2^{(14)}$ → as 15 reserved for ∞

lowest possible non-negative = $+(0.1 \underbrace{00001}_{m})_2 \times 2^{(1)}$ → as 0 reserved for zero.