**MCQ: Choose Only One Answer**

1. (a) In a 32-bit system, what is the highest possible number?
   **A.** $2^{31}$ **B.** $2^{31}-1$ **C.** $2^{32}-1$ **D.** $2^{32}$


So, $2^{31}-1$

(a) B

(b) Given, $\beta = 2, m = 4$ and $e \in \{-1, 2\}$. Using the Normalized form, how many non-negative numbers can you represent?
   **A.** 16 **B.** 32 **C.** 64 **D.** 128

Normalized form, $(0.1d_1d_2d_3d_4)$
$e \to -1, 0, 1, 2$

So, $2^4 \times 4 = 64$

(b) C

(c) How many significant digits does the floating point number 0.020250 have?
   **A.** 3 **B.** 4 **C.** 5 **D.** 6

$0.020250$
5sf

(c) C

(d) Given, $\beta = 2, m = 3$ and $e \in \{-1, 1\}$. Using the Normalized form, what is the value of *unit roundoff*?
   **A.** $\frac{1}{16}$ **B.** $\frac{1}{8}$ **C.** $\frac{1}{4}$ **D.** $\frac{1}{2}$

$e_m = \frac{1}{2}\beta^{-m}$
$= \frac{1}{2} \times 2^{-3} = \frac{1}{16}$

(d) A

(e) Which of the following statements is/are true?
   i. In case of machine epsilon, we consider the max value of $|x|$. → $|x|_{min}$
   ii. Machine Epsilon is the maximum scale invariant error. ✓
   iii. Loss of significance occurs when subtracting two values which are very close. ✓
   **A.** (i, ii) only. **B.** (i, iii) only. **C.** (ii, iii) only. **D.** All of these.

(e) C

2. (a) In a 64-bit system, what is the highest possible number?
   **A.** $2^{63}$ **B.** $2^{63}-1$ **C.** $2^{64}-1$ **D.** $2^{64}$

(a) B

(b) Given, $\beta = 2, m = 4$ and $e \in \{-1, 1\}$. Using the Denormalized form, how many non-negative numbers can you represent?
   **A.** 16 **B.** 32 **C.** 48 **D.** 64

Denormalized form, $(1.d_1d_2d_3d_4)$
$e \to -1, 0, 1$

So, $2^4 \times 3 = 48$

(b) C

(c) How many significant digits does the floating point number 0.020256 have?
   **A.** 3 **B.** 4 **C.** 5 **D.** 6

(c) C

(d) Given, $\beta = 2, m = 3$ and $e \in \{-1, 1\}$. Using the Denormalized form, what is the value of *unit roundoff*?
   **A.** $\frac{1}{16}$ **B.** $\frac{1}{8}$ **C.** $\frac{1}{4}$ **D.** $\frac{1}{2}$

$e_m = \frac{1}{2}\beta^{-m}$
$= \frac{1}{2} \times 2^{-3} = \frac{1}{16}$

(d) A

(e) Which of the following statements is/are true?

i. In case of machine epsilon, we consider the min value of $|x|$. ✓

ii. Machine Epsilon is the maximum scale invariant error. ✓

iii. Loss of significance occurs when subtracting two values which are not very close. should be very close

**A.** (i, ii) only. **B.** (i, iii) only. **C.** (ii, iii) only. **D.** All of these.

(e) A

3. (a) In a 16-bit system, what is the highest possible number?

**A.** $2^{15}$ **B.** $2^{15}-1$ **C.** $2^{16}-1$ **D.** $2^{16}$

(a) B

(b) Given, $\beta = 2, m = 3$ and $e \in \{-1, 2\}$. Using the Denormalized form, how many non-negative numbers can you represent?

**A.** 16 **B.** 32 **C.** 64 **D.** 128

Denormalized form, $(1.d_1 d_2 d_3)$

$e \to -1, 0, 1, 2$

So $2^3 \times 4 = 32$

(b) B

(c) How many significant digits does the floating point number 0.30370 have?

**A.** 3 **B.** 4 **C.** 5 **D.** 6

(c) C

(d) Given, $\beta = 2, m = 3$ and $e \in \{-1, 1\}$. Using the Standard form, what is the value of *unit roundoff*?

**A.** $\frac{1}{16}$ **B.** $\frac{1}{8}$ **C.** $\frac{1}{4}$ **D.** $\frac{1}{2}$

$e_m = \frac{1}{2}\beta^{1-m}$

$= \frac{1}{2} \times 2^{1-3} = \frac{1}{8}$

(d) B

(e) Which of the following statements is/are true?

i. In case of machine epsilon, we consider the min value of $|x|$. ✓

ii. Machine Epsilon is the minimum scale invariant error. should be maximum

iii. Loss of significance occurs when subtracting two values which are very close. ✓

**A.** (i, ii) only. **B.** (i, iii) only. **C.** (ii, iii) only. **D.** All of these.

(e) B

**Problems: Marks are as indicated**

4. (4 marks) Given a system for Standard form with $\beta = 2, m = 4$ and $e \in \{-2, 1\}$. Evaluate the rounding error if you store the product of x $= \frac{7}{8}$ and y $= \frac{5}{16}$. Express your answer in decimal format.

$x = \frac{7}{8}$

$= \frac{4}{8} + \frac{2}{8} + \frac{1}{8}$

$= \frac{1}{2} + \frac{1}{4} + \frac{1}{8}$

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

$= (0.111)_2 \times 2^0$

$y = \frac{5}{16}$

$= \frac{4}{16} + \frac{1}{16}$

$= \frac{1}{4} + \frac{1}{16}$

$= 2^{-2} + 2^{-4}$

$= (0.0101)_2 \times 2^0$

$x \cdot y = \frac{7}{8} \times \frac{5}{16} = \frac{35}{128}$

$fl(x.y) = \frac{35}{128}$

$= \frac{32}{128} + \frac{2}{128} + \frac{1}{128}$

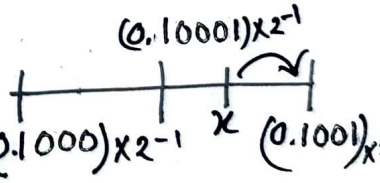
$= \frac{1}{4} + \frac{1}{64} + \frac{1}{128}$

$= 2^{-2} + 2^{-6} + 2^{-7}$

$= (0.0100011)_2 \times 2^0$

$= (0.100011)_2 \times 2^{-1}$

Since, $(m=4)$ $fl(xy) = (0.1001)_2 \times 2^{-1} = \frac{9}{32}$




Rounding error,

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

$$= \frac{\left|\frac{9}{32} - \frac{35}{128}\right|}{\left|\frac{35}{128}\right|}$$

$$= \left(\frac{1}{35}\right)_{10}$$