Tutorial - 6

Floating point numbers & Operations

Question-1

The n-bit fixed point representation of an unsigned real number X uses f bits for the fraction part .Let i = n-f. The range of decimal values of X in this representation is…

Question-1

Q. The n-bit fixed point representation of an unsigned real number X uses f bits for the fraction part .Let i = n-f. The range of decimal values of X in this representation is.

Sol contd……

So, it makes a GP series, with f bits maximum number possible is sum of GP series.

Maximum value with f bits possible

$= \frac{1}{2} + 1/4 + 1/8 + 1/16 + 1/32 + \ldots\ldots = 1/2\ [(1-(1/2)^f / (1-(1/2))] = 1 - 2^{-f}$

So, maximum fractional value possible

= max value with i bits + max value with f bits

$= 2^i - 1 + 1 - 2^{-f} = 2^i - 2^{-f}$ = so, required range will be 0 to $2^i - 2^{-f}$.

Question-2

Q. In the standard IEEE754 single precision floating point representation, there is 1 bit for sign, 23 bits for fraction and 8 bits for exponent.
What is the precision in terms of the number of decimal digits?

Question-2

Q. In the standard IEEE754 single precision floating point representation, there is 1 bit for sign, 23 bits for fraction and 8 bits for exponent.
What is the precision in terms of the number of decimal digits?

Sol :

The floating point representation has three parts:

1. Sign of Mantissa - represent the sign of number {0- negative, 1- positive},

2. Exponent – represent both positive and negative exponent
   bias is added to the actual exponent in order to get stored, exponent.

3. Normalised Mantissa- represent the digits with only one "1" to the left of the decimal. 4. The value of the Normalized number is $(-1)^s \times 1.M^{E-127}$

Question-2

Q. In the standard IEEE754 single precision floating point representation, there is 1 bit for sign, 23 bits for fraction and 8 bits for exponent.
What is the precision in terms of the number of decimal digits?

Sol :

The floating point representation has three parts:

1.Sign of Mantissa - represent the sign of number {0- negative,1- positive},

2.Exponent – represent both positive and negative exponent bias is added to the actual exponent in order to get stored, exponent.

3. Normalised Mantissa- represent the digits with only one "1" to the left of the decimal. 4.The value of the Normalized number is $(-1)^s \times 1.M^{E-127}$

Question-2

Q. In the standard IEEE754 single precision floating point representation, there is 1 bit for sign, 23 bits for fraction and 8 bits for exponent.
What is the precision in terms of the number of decimal digits?

Sol :

The floating point representation has three parts:

1. Sign of Mantissa - represent the sign of number {0- negative, 1- positive},

2. Exponent – represent both positive and negative exponent
   bias is added to the actual exponent in order to get stored, exponent.

3. Normalised Mantissa- represent the digits with only one "1" to the left of the decimal. 4. The value of the Normalized number is $(-1)^s \times 1.M^{E-127}$

Question-2

Q. In the standard IEEE754 single precision floating point representation, there is 1 bit for sign, 23 bits for fraction and 8 bits for exponent.
What is the precision in terms of the number of decimal digits?

Sol. Contd...

Precision can be represented by 1.M, So M+1 is used to represent the precise number 23 bits + 1 =24 bits

According to the formula :

$$\text{Base } X^{\text{no of digits}} = \text{Base } Y^{\text{no of digits}}$$

$$\Rightarrow 2^{24}=10^{y} \Rightarrow y=24 \log_{10}2 = 7.22.$$

Hence, the precision in terms of the decimal digits is 7

Question-3

Q.

a) What number is represented by the single-precision float
   11000000101000…00

b) Represent -0.75 in single and double precision format.

Question-3

Q.

a) What number is represented by the single-precision float

11000000101000...00

b) Represent -0.75 in single and double precision format.

Sol:

a) S = 1

Fraction = $01000...00_2$ Exponent =

$10000001_2 = 129$

$x = (-1)^1 \times (1_{10} + 01_2) \times 2^{(129 - 127)}$

$= (-1) \times 1.25 \times 2^2 = -5.0$

Question-3

Q.

a) What number is represented by the single-precision float
11000000101000…00

b) Represent -0.75 in single and double precision format.

Sol Contd…..

$-0.75 = (-1)^1 \times 0.11_2 = (-1)^1 \times 1.1_2 \times 2^{-1}$ □□ in binary

S = 1

Mantissa(Fraction) = $1000\ldots00_2$ Exponent = –1 + Bias

Single: –1 + 127 = 126 = 01111110_2 (e=E-127 => E=e + 127)

Double: –1 + 1023 = 1022 = 01111111110_2 Single: 1011111101000…00

Double: 1011111111101000…00

Question-4

Perform the following operations and check the result for errors due to rounding.

a)Add 9.78×10^{25} and 8.79×10^{24} assuming 3 digit mantissa .

b)Add 9.76×10^{25} and 2.59×10^{24} assuming 3 digit mantissa .

Question-4

Perform the following operations and check the result for errors due to rounding.

a)Add 9.78×10^{25} and 8.79×10^{24} assuming 4 digit mantissa .

Sol.

Add 9.78×10^{25} and 8.79×10^{24} (without rounding)

- Shift mantissa of the smaller number to the right:
  0.879×10^{25}
- Add mantissas: 10.65×10^{25}
- Normalize mantissa if necessary:
  1.065×10^{26}.

- 0 11010101 0001000011011110110100

-Add 9.78×10^{25} and 8.79×10^{24} (with rounding)

- Shift mantissa of the smaller number to the right: 0.879×10^{25}
- Add mantissas (note extra digit on the left): 10.6590×10^{25}
- Check and normalize mantissa if necessary: 1.065×10^{26}
  0 11010101
- Round the result: 1.07×10^{26}.

- 0 11010101 00001000010101010010101

Difference due to rounding – 1.07 – 1.065
= 0.005.

Question-4

Perform the following operations and check the result for errors due to rounding.

b)Add 9.76×10^{25} and 2.59×10^{24} assuming 3 digit mantissa .

Sol :

– Add 9.76×10^{25} and 2.59×10^{24} (without rounding)
- Shift mantissa of the smaller number to the right:
  0.259×10^{25}
- Add mantissas: 10.01×10^{25}
- Check and normalize mantissa if necessary:
  1.00×10^{26} .
- 0 11010101 00000000000000000000000
– Add 9.76×10^{25} and 2.59×10^{24} (with rounding)
- Internal registers have extra two digits:
  9.7600×10^{25} and 2.5900×10^{24}

- Shift mantissa of the smaller number to the right: 0.2590×10^{25}
- Add mantissas: 10.0190×10^{25}
- Check and normalize mantissa if necessary: 1.00×10^{26}
- Round the result: 1.00×10^{26} .

- 0 11010101 00000000000000000000000

Difference due to rounding : 0

Question-4

Perform the following operations and check the result for errors due to rounding.

b)Add 9.76×10^{25} and 2.59×10^{24} assuming 3 digit mantissa .

Sol :

– Add 9.76×10^{25} and 2.59×10^{24} (without rounding)
- Shift mantissa of the smaller number to the right:
  0.259×10^{25}
- Add mantissas: 10.01×10^{25}
- Check and normalize mantissa if necessary:
  1.00×10^{26} .
- 0 11010101 00000000000000000000000
– Add 9.76×10^{25} and 2.59×10^{24} (with rounding)
- Internal registers have extra two digits:
  9.7600×10^{25} and 2.5900×10^{24}

- Shift mantissa of the smaller number to the right: 0.2590×10^{25}
- Add mantissas: 10.0190×10^{25}
- Check and normalize mantissa if necessary: 1.00×10^{26}
- Round the result: 1.00×10^{26} .

- 0 11010101 00000000000000000000000

Difference due to rounding : 0

Question-5

As per the IEEE 754 floating point representation ,
the standard default rounding mode is rounding-to-nearest.

Here, values are rounded to the closest representable number and results that lie exactly halfway between two representable numbers are rounded such that the least significant digit of their result is even.

Given the binary numbers in the table, round these numbers to nearest ¼ and fill the rest of the entries as well.

Binary	Rounded Binary	Action(rounded up or down)	Rounded Value (decimal)
10.00011			
10.00110			
10.11100			
10.10100			

Question-5

As per the IEEE 754 floating point representation ,
the standard default rounding mode is rounding-to-nearest.

Here, values are rounded to the closest representable number and results that lie exactly halfway between two representable numbers are rounded such that the least significant digit of their result is even.

Given the binary numbers in the table, round these numbers to nearest ¼ and fill the rest of the entries as well.

Binary	Rounded Binary	Action (rounded up or down)	Rounded Value (Decimal)
10.00011	10.00	Rounded down(1/2)	2
10.00110	10.01	Rounded up (1/2)	2 1/4
10.11100	11.00	Rounded up (1/2)	3
10.10100	10.10	Rounded down (1/2)	2 1/2