Floating point Representation

-by Komal Sindhi

Floating point representation

- IEEE 754 floating point representation
- Standard format for representing decimal floating point number into binary in 32-bits
- The distribution of 32 bits is follows:

- S = Sign of the number *(1 for negative and 0 for positive)*
- E= 8- bit biased exponent
- M= 23-bit mantissa

IEEE 754 floating point representation

- Given Number:
- 85. 125

STEP 1: Convert the number into binary.

- Binary for 85 = 1010101
- Binary for 0.125= 0.001
- Final number: 1010101. 001

STEP 2: Normalize the given number such that there is a “1” before decimal point.

- Normalized number: 1.010101×2^6

IEEE 754 floating point representation

- STEP 3: Find S

Here the number is positive. So the value of S (first bit) = 0.

STEP 4: Find biased Exponent

Exponent:+6

Biased Exponent= +6 +127= 133

Find Binary for 133= 10000101 = 8-bit Exponent E

IEEE 754 floating point representation

STEP 5 : Find 23 –bit Mantissa

From the normalized number (1.010101×2^{6}) , 1 before the decimal point is implicit one. Ignore that and consider remaining bits.

Remaining bits are: 010101 - 6 –bits

23 –bit Mantissa M = 01010100000000000000000.

Final Answer : SEM = 01000101010101000000000000000000.

1.11001 X 2 ^ 3

More examples:

- Question : Hexadecimal notation for -14.25
- Binary: 1110.01
- Normalized Binary Number : 1.11001×2^3

1) S=1 because the number is negative

2) Exponent = 3 +127 =130= 10000010 (8-bit binary)

3) Mantissa: After ignoring implicit 1, 11001 is left

23- bit Mantissa= 11001000000000000000000

Final Answer: 11000001011001000000000000000000

Example:

Answer in Hexadecimal

1100	0001	0110	0100	0000	0000	0000	0000
C	1	6	4	0	0	0	0

Examples…

- Convert (-0.75) into binary
- (0.11)
- 1.1×2^{-1}

1) S= 1 (negative number)

2) Biased exponent= -1 +127 = 126 (01111110 in binary)

3) Mantissa: After ignoring implicit 1 , 1 is left

So M= 10000000000000000000000

- Answer: 10111111010000000000000000000000

Examples..

- Given: 11000000101000000000000000000000
- 1 10000001 01000000000000000000000

1) Sign = Negative (consider -1)

2) Biased exponent in Binary = 10000001 = 129

3) Exponent= 129-127 =2

4) Mantissa or fractional part: 0. 01000000000000000000000
= 0.25 in decimal

5) Add implicit 1 to 0.25, So **M =1 + 0.25= 1.25**

- **Answer= (S) (E) (M) = $(-1) \times (2^2) \times (1.25) = (-5)$**

Examples..

- Given :
- 0 10000011 10100000000000000000000
- Find the corresponding floating point number