1. Convert the decimal number 8.625 into floating point IEEE Standard 754 single precision: ???

ieee 754 format is : X             XXXXXXX  XXXXXXXXXXXXXXXXXXXXXXXX

                          1 sign bit     8 bits exp              23 bits  significant (fraction- mantissa)

8**.**625 ????

First, we convert 8 to binary (left of fixed decimal point or integral part) then .625 (right of fixed decimal point)

 8 to binary: series of division by two

 result        new base                  remainder

8                   /  2                          0  (least significant)

4                    / 2                          0

2                     /2                          0

1                    /2                           1 (most significant)

0  --> we stop here (at 0)

8 in binary is 1000

verification:

 1000 in binary means : 1 x (2 power 3) +  0 x (2 power 2) +  0  x (2 power 1) + 0  x (2 power 0)

1 X 8               + 0 x  4                  + 0 x    2               + 0 x  1

0.625 to binary : series of multiplication by two

|  |  |  |
| --- | --- | --- |
| number | X 2 | result |
| **.625** | X 2 | **1.25** most significant |
| **.25** | X 2 | **0.5** |
| **.5** | X 2 | **1.0** least significant |
| **0 (Stop here)** | X 2 |  |

**0.625 in binary is . 1 0 1**

Verification: . 1 x(1/2) + 0 x(1/4) + 1 x (1/8) = 0.5 + 0 + 0.125 = 0.625

Now the binary of 8.625 is 1000.101

We need to express this number as sign exp and fraction

Example in decimal ( 8 is 0.8 x 10 power 1

or 0.08 x 10 power 2 …)

Our binary:

1000.101 can be expressed as 1.000101 x 2 power 3

Our number is positive; it means our bit sign will be 0 [1: SIGN]

Our exponent is 3 will be represented in 8 bit with excess 127

Our fraction .000101 will be represented in 23 bits

My IEEE 754 will look like 0 xxxxxxxx xxxxxxxxxxxxxxxxxxxxxxxxx

Xxxxxxxx = 3+ 127 in binary

3+ 127 = (130) in decimal = 10000010 in binary [ II: EXPONENT]

Our fraction .000101 will be .00010100000000000000000 [III: MANTISSA or SIGNIFICANT]

8.625 in IEEE 754 is 0 10000010 00010100000000000000000

Let me know if it’s still not clear!!

Addition HW:

Please convert following numbers:

1. (17.25) 10 to (IEEE 754 single precision)
2. ( 8.625) 10 to (IEEE 754 double precision)