CS 3320 – IEEE Standard Homework

1. What is the logical bit layout of the number -12.5 in IEEE single-precision format (float)? Separate the 3 parts by a space for readability.

   $12 = 1100_2$
   $0.5 = 0.1_2$
   $12.5 = 1100.1_2 = 1.1001_2*2^3$
   Exponent = 011 + 01111111 = 10000010
   Normalized mantissa = 10010000…0
   Sign bit=1

   1 10000010 10010000000000000000000

2. What is **ulps(20,30)** in IEEE single-precision?

   $30 = 11110_2 = 1.111_2*2^4$ = 01000001111100000000000000000000
   $20 = 10100_2 = 1.01_2*2^4$ = 01000001101000000000000000000000

   ulps(20, 30) = $(111\text{-}010)_2*2^{20}$= 101_2*2^{20} =$5*2^{20}$ = 5242880
   or $10*2^{19}$

   Note: $(111\text{-}010)_2*2^{20}$ = 11100000000000000000000 – 01000000000000000000000)

3. What number is represented by the following IEEE single-precision value?

   1 10000100 10110000000000000000000

   Sign bit =1 -> Negative
   Mantissa 1.1011_2
   Exponent 10000100 – 01111111 = 00000101 = 5
   $-1.1011*2^5$ = -110110 = - (32 + 16 + 4 + 2) = –54

   Answer -54

4. The number 20 can be expressed in binary as 1.01×2^4, and 11 as 1.011×2^3. Assuming 4 bits of precision:

   a. Do the binary arithmetic to compute 20 – 11. Give the answer in decimal.

   $$\begin{array}{r} 1.010 \times 2^4 \\ \underline{-\ \ 0.101 \times 2^4} \\ 0.101 \times 2^4 \end{array} = 1010 = 10_{10}$$

b. Repeat part a) using 1 guard digit.

$$\begin{array}{l} 1.0100 \times 2^4 \\ -\ 0.1011 \times 2^4 \\ \hline 0.1001 \times 2^4 \quad =1001 = 9_{10} \end{array}$$

5. What is meant by the measure, “the number of ulps between floating-point numbers x and y?”

   The number of floating-point intervals (numbers) between x and y

6. Describe the logical bit layout of an IEEE infinity.

   Exponent is all ones; mantissa is all zeros; sign either 1 or 0

7. Describe the logical bit layout of an IEEE NaN.

   Exponent is all ones; mantissa is nonzero; sign either 1 or 0

8. Describe the logical bit layout of an IEEE zero.

   Exponent and mantissa are all zeros; sign either 1 or 0

9. Describe the logical bit layout of an IEEE subnormal number.

   Exponent is all zero’s

10. How do subnormal numbers differ from normalized numbers with respect to:

    a. Spacing
       The actual spacing between all subnormals is uniform.

    b. Relative roundoff error
       Since spacing doesn’t decrease the relative roundoff error increases to 100% as you approach zero

11. What are *guard digits*, and why are they useful?

    Extra digits preserved in the operands of floating-point operations (in the arithmetic unit). This minimizes roundoff error.