Homework Assignment 6

CS/ECE 3810: Computer Organization
October 13, 2020

IEEE-754 floating-point representation

Due Date: October 19, 2020.
(100 points)

Important Notes:

- Solutions turned in must be your own. Please, mention references (if any) at the end of each question. *Please refrain from cheating.*
- All solutions must be accompanied by the equations used/logic/intermediate steps. Writing only the final answer will receive **zero** credits.
- Partial score of every question is dedicated to each correct final answer provided by you. Please ensure both your equation/logic and final answer are correct. Moreover, you are expected to provide explanation for your solutions.
- All units must be mentioned wherever required.
- Late submissions **(after 11:59PM on 10/19/2020)** will not be accepted.
- We encourage all solutions to be typed in for which you could use software programs like LaTeX, Microsoft Word etc. If you submit handwritten solutions, they must be readable by the TAs to receive credits.
- All submitted solutions must be in the PDF format unless otherwise mentioned.

IEEE 754 Representation Format. The IEEE Standard for Floating-Point Arithmetic (IEEE 754) is a technical standard for floating-point arithmetic established in 1985 by the Institute of Electrical and Electronics Engineers (IEEE). Many hardware floating-point units use the IEEE 754 representation format. The latest version (IEEE 754-2019) was published in 2019.

Floating-point representation is also a significant area of research for architects, owing to a number of weaknesses for the IEEE-754 representation format. Alternative representations such as Posit (Click here to learn more about Posit and drawbacks of IEEE-754) have been proposed recently to alleviate some of these drawbacks

Question 1. Lecture 12 introduces single precision and double precision IEEE-754 floating-point representation formats. Using the technique explained in the lecture, do the following conversions. Show all steps involved in the conversion: **(40 points)**

1. The decimal number -97.32768 into a single-precision floating-point number
2. The decimal number -1331.65568 into a double-precision floating-point number
3. The single-precision floating-point number 01000101101000000000000000000000 into a decimal number
4. The double-precision floating-point number
   0 10000000110 0100100000000000000000000000000000000000000000000000000000
   into a decimal number

1. First, split up the the whole part and decimal part of the number to get 1100001 and 1000000000000000 respectively. This gives 1100001.1000000000000000. Now convert to scientific notation in base 2 to get $1.10000110000000000000000 * 2^6$. The normalized mantissa is 10000110000000000000000. Thus, using the formula, 97.32768 converted to single-precision floating-point is:

   1 00000110 10000110000000000000000

2. First, split up the whole part and decimal part of the number to get 10100110011 and 10000000000100000 respectively. This gives 10100110011.10000000000100000. Now convert to base 2 scientific notation to get $1.01001100111000000000010000 * 2^{10}$. The normalized mantissa is 01001100111000000000010000. Thus, using the formula, 1331.65568 converted to double-precision floating point is:

   1 00000001010 0000000000000000000000000001001100111000000000010000

3. Using the mantissa, take the sum of the inverse of the positions that have a 1. There's only one position with a 1 in this case, so the decimal value for 01000000000000000000000 is $2^{-2} = \frac{1}{4} = 0.25$
4. Similar to part 3, using the mantissa, take the sum of the inverse of the positions that have a 1. Like so:
   $2^{-2} + 2^{-5} = \frac{1}{4} + \frac{1}{32} = \frac{9}{32} = 0.2812$

IEEE-754 arithmetic. Lecture 13 introduces techniques to perform basic arithmetic (such as addition or multiplication) between two single-precision or double-precision floating point numbers.

Question 2. Perform the following operations. Show the steps involved in each addition/-multiplication operation. Represent the final answer in IEEE - 754 single-precision format **(60 points)**

1. Addition of two single-precision floating-point numbers A and B:
   A: 0 10000111 01011000000000000000000
   B: 0 10000100 10010010000000000000000

2. Multiplication of two single-precision floating-point numbers X and Y:
   X: 0 10000011 10110000000000000000000
   Y: 1 10000110 00011000000000000000000

1. $E_A = 135 \qquad M_A = 1.0101100000$
   $E_B = 135 \qquad M_B = 0.0011001001$
   $M_A + M_B = 1.1000101001$
   $A + B = 0\ 10000111\ 10001010010000000000000$

2. $E_X = 131 \qquad M_X = 1.01011$
   $E_Y = 134 \qquad M_Y = 1.00011$
   $E_{X*Y} = 265 - 127 = 138 \qquad M_{X*Y} = 10.111100001$
   $E_{X*Y} = 139 \qquad M_{X*Y} = 1.0111100001$
   $X * Y = 1\ 10001010\ 01111000010000000000000$