COMP2611 Spring 2022 Homework #2

Note:

- **The deadline of this homework is at 11:55pm on Friday, 25 March 2022 (Hong Kong Time, UTC+8). NO late submissions will be accepted!**
- **Work out the answers of the questions either directly on the hardcopy of this document or on your own paper sheets. Then scan all the answer pages into a single pdf file "homework2_<studID>.pdf". Or take photos and zip them into a single zip file "homework2_<stdID>.zip". Make sure every detail of the answers is clearly visible in your submission (verify this on the scanned pages before submitting), otherwise marks may be deducted.**
- **We only accept e-submissions at the Canvas. To submit, first find the Canvas page of COMP2611, homework 2, and then upload the file. You can upload for multiple times, only the last one before the deadline will be marked.**
- **Make sure you keep the original copy of your homework until the homework score is finalized.**

Name :

Student ID :

Email :

Question	Marks
1. Data Representation – Signed Integers	/16
2. Floating Point Representation (IEEE754)	/18
3. A Floating Point System for Lengths	/16
Total	/50

Question 1: Data Representation – Signed Integers (16 points)

a) Given the following 16-bit 2's complement representation, write their **hexadecimal representations** and **decimal values**. (8 points)

1) 1111 0110 1101 0101: (4 points)

Hexadecimal representation = F6D5(16)

Decimal = -2347 (10)

2) 0101 0010 1011 1111: (4 points)

Hexadecimal representation = 52BF(16)

Decimal = 21183 (10)

b) Given the following 32-bit 2's complement representation, write their **hexadecimal representations** and **decimal values**. (8 points)

1) 0111 1111 1111 1111 0100 1101 0010 1011: (4 points)

Hexadecimal representation = 7FFF4D2B(16)

Decimal = 2147437867 (10)

2) 1111 1111 1111 1110 1100 0101 0110 1010: (4 points)

Hexadecimal representation = FFFEC56A(16)

Decimal = -80534(10)

Question 2: Floating Point Representation (18 points)

a) Write the IEEE754 single-precision representation of the following decimal numbers. Can the decimal numbers be represented exactly? If not, please find the nearest approximation of the number. Show your steps briefly, otherwise no mark will be given. (10 points)

1) -116.625 (4 points)

The number can be precisely represented.

$-116.625 = -1.110100101*2^6$

S= 1, E=6 +127 (bias for single precision) = $1000\ 0101_{(2)}$

IEEE754 single precision is: 1 1000 0101 11010010100000000000000

2) 0.085 (6 points)

0.085 * 2 = 0.17
0.17 * 2 = 0.34
0.34 * 2 = 0.68
0.68 * 2 = 1.36
0.36 * 2 = 0.72
0.72 * 2 = 1.44
0.44 * 2 = 0.88
0.88 * 2 = 1.76

0.76 * 2 = 1.52
0.52 * 2 = 1.04
0.04 * 2 = 0.08
0.08 * 2 = 0.16
0.16 * 2 = 0.32
0.32 * 2 = 0.64
0.64 * 2 = 1.28
0.28 * 2 = 0.56
0.56 * 2 = 1.12
0.12 * 2 = 0.24
0.24 * 2 = 0.48
0.48 * 2 = 0.96
0.06 * 2 = 1.92
0.92 * 2 = 1.84
0.84 * 2 = 1.68
0.68 * 2 = 1.36 (at this point the number starts repeating its pattern)

0.36 * 2 = 0.72
0.72 * 2 =1.44
0.44 * 2 =0.88

Hence, the number is 1. 0101110000101000111101011100001010001111… * 2^{-4}
Sign bit is 0
Exponent to be stored is -4 + 127 (bias for single precision) =123 = 0111 1011
Significand can only take the first 23 bits: 0101 1100 0010 1000 1111 010
The number can't be precisely represented
The number to its left (smaller than it, truncation at 23rd bit) that can be precisely represented is :
1. 01011100001010001111010 * 2^{-4}
The number to its right (larger than it, simple rounding) that can be precisely represented is :
1. 01011100001010001111011 * 2^{-4} (nearest approximation)
The final representation is
0 01111011 01011100001010001111010 or
0 01111011 01011100001010001111011

b) What decimal values are represented by the following IEEE754 single-precision floating-point representations? Show your steps briefly. You don't need to work out the exact value, it's good enough to show your answer in power of 2. (8 points)

1) 0 10000110 10010101110100000000000 (4 points)
$= 1.100101011101_{(2)} * 2^7 = 202.90625_{(10)}$

2) 1 00000000 00110100000000000000000 (4 points)
$= -0.001101_{(2)} * 2^{-126}$
$= -13 * 2^{-6} * 2^{-126}$
$= -13 * 2^{-132}$
$\approx\ 2.3877229 * 10^{-39}$

Question 3: A Floating-Point System for Lengths (16 points)

You are asked to design a 16-bit floating point number system to store the length of various man-made objects. This system should work in a similar way as the IEEE754 standard. Assume a value stored denotes the length of an object in centimeters, assume also that the maximum length to be stored is 45845.0 centimeters (i.e. length of the biggest man-made oil-tanker, the "Seawise Giant").

Note:
This representation has normalized, de-normalized and special cases as you have seen in IEEE754 standard.

Answer the questions below:

a) Is sign bit needed in this system? Why or why not. (2 points)

This system does not require the sign bit. Because the length of an object is always positive. Therefore, this is a system for unsigned numbers. We only need to allocate bits for exponent and significand.

b) What is the minimum number of bits needed for the exponent? What is the value of the corresponding bias? Show your steps clearly. If you write the value directly without showing the steps for reasoning, you will not get any point. (6 points)

The maximum length is 45845 centimeters and the exponent should be big enough to hold it.

$$2^{15} < 45845_{(10)} < 1.1_{(2)} * 2^{15}$$

This indicates that a max real exponent value of 14 is not enough (since $2^{15} < 45845_{(10)}$). But as long as we have at least one digit for the significand field, we would be able to store the value with the real exponent value of 15 (because $45845 < 1.1_{(2)} * 2^{15}$). Therefore the system needs to store a real exponent as big as 15.

To be able to hold the biggest real exponent value of 15, we would need at least 5 bits for the exponent field. When the exponent is 5-bit, the max exponent is $11110_{(2)} = 30_{(10)}$, the bias is $2^{5-1}-1 = 15$. The significand field is then 16-5=11 bits.

c) What is the maximum length the system can represent? (To get full marks, please show your step clearly.) (2 points)

The number is the largest when the exponent = $11110_{(2)} = 30_{(10)}$, the largest significand is 111 1111 1111.
the value is $1.11111111111 * 2^{30-15} = (2 - 2^{-11}) * 2^{15} = 65520$

Therefore, the maximum length this system can represent is 65520 centimeters.

d) If a length exceeds the maximum length the system can represent, it will be represented as infinity. Show the representation of the infinity with this number system. (2 points)

Exponent = 11111, significant is zero,
Positive infinity = 11111 00000000000

e) Show the representation of $\mathbf{3245_{(10)}}$ with this system. (To get full marks, please show your step clearly.) (2 points)

$3245_{(10)} = 1.10010101101*2^{11}$
Exponent = 11, Biased exponent = 11 + 15 = 26 = $11010_{(2)}$
Therefore the representation of $3245_{(10)}$ is 11010 10010101101

f) Can the number $\mathbf{953.375_{(10)}}$ be represented precisely by the system? The answer "Yes" or "No" alone will get you no mark. Please justify your answer with clear calculation steps. (2 points)

$953.375_{(10)} = 1.110111001011 * 2^9$
Exponent = 9, Biased exponent = 9+15 = 24 = $11000_{(2)}$
Significand = 110111001011
The digits to be stored in the significand is 110111001011, which is 12-bit in length. However, there are only 11 bits for significand, Thus, it could not be represented exactly.