CARLETON UNIVERSITY

Department of Systems and Computer Engineering

SYSC 5704 **Elements of Computer Systems** **Assignment 2**

Due date: Thursday, October 23^{rd}, 18:00.

3.10 [10] <§3.2> Assume 151 and 214 are *signed* 8-bit decimal integers stored in 2's complement format. Calculate $151 - 214$ using saturation arithmetic. The result should be in decimal.[1]

3.17 [20] <§3.3> As discussed in the text, one possible performance enhancement is to do a shift and add instead of an actual multiplication. Since 9×6, for example, can be written $(2 \times 2 \times 2 + 1) \times 6$, we can calculate 9×6 by shifting 6 to the left 3 times and then adding 6 to that result. Show the best way to calculate `033` × `055` using shifts and adds/subtracts. Assume both inputs are 8-bit unsigned integers.

3.20 [5] <§3.5> What decimal number does the bit pattern `0x0C000000` represent if it is a two's complement integer? An unsigned integer?

3.22 [10] <§3.5> What decimal number does the bit pattern `0x0C000000` represent if it is a floating point number? Use the IEEE 754 standard.

3.23 [10] <§3.5> Write down the binary representation of the decimal number **63.25** assuming the IEEE 754 single precision format.

3.27 [20] <§3.5> IEEE 754-2008 contains a half precision that is only 16 bits wide. The leftmost bit is still the sign bit, the exponent is 5 bits wide and has a bias of 15, and the mantissa is 10 bits long. A hidden 1 is assumed. Write down the bit pattern to represent -1.5625×10^{-1} assuming a version of this format, which uses an excess-16 format to store the exponent. Comment on how the range and accuracy of this 16-bit floating point format compares to the single precision IEEE 754 standard.

3.32 [20] <§3.9> Calculate $(3.984375 \times 10^{-1} + 3.4375 \times 10^{-1}) + 1.771 \times 10^{3}$ by hand, assuming each of the values are stored in the 16-bit half precision format described in exercise 3.27 (and also described in the text). assume 1 guard, 1 round bit, and 1 sticky bit, and round to the nearest even. show all the steps, and write your answer in both the 16-bit floating point format and in decimal.

3.33 [20] <§3.9> Calculate $3.984375 \times 10^{-1} + (3.4375 \times 10^{-1} + 1.771 \times 10^{3})$ by hand, assuming each of the values are stored in the 16-bit half precision format described in Exercise 3.27 (and also described in the text). Assume 1 guard, 1 round bit, and 1 sticky bit, and round to the nearest even. Show all the steps, and write your answer in both the 16-bit floating point format and in decimal.

3.34 [10] <§3.9> Based on your answers to 3.32 and 3.33, does $(3.984375 \times 10^{-1} + 3.4375 \times 10^{-1}) + 1.771 \times 10^{3} = 3.984375 \times 10^{-1} + (3.4375 \times 10^{-1} + 1.771 \times 10^{3})$?

3.41 [10] <§3.5> Using the IEEE 754 floating point format, write down the bit pattern that would represent $-1/4$. Can you represent $-1/4$ exactly?

3.42 [10] <§3.5> What do you get if you `add` $-1/4$ to itself 4 times? What is $-1/4 \times 4$? Are they the same? What should they be?

[1] Hint: Convert 151 and 214 to binary before doing the subtraction.