SCHOOL OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

G51CSA Homework/Tutorial Problems #4

1. Represent following decimal numbers in 8-bit twos complement numbers

   (a) 123
   (b) 64
   (c) 127
   (d) 43
   (e) -123
   (f) -64
   (g) -127
   (h) -43

2. Find the following differences using 8-bit twos complement arithmetic and verify the results.

   (a) 123 - 127
   (b) -123 - 127
   (c) 64 - 43
   (d) 43 - 43
   (e) 127 - 123

3. Convert the following numbers into IEEE single-precision format. Give the result as eight hexadecimal digits.

   (a) 9
   (b) 5/32
   (c) -5/32
   (d) 6.125

4. Convert the following IEEE single-precision floating point numbers from hex to decimal:

   (a) 42E48000H
   (b) 3F880000H
   (c) 00800000H
   (d) C7F00000H

5. To add two floating-point numbers, you must adjust the exponents (by shifting the fraction) to make them the same. Then you can add the fractions and normalise the result, if need be. Add the single-precision IEEE numbers 3EE00000H and 3D800000H and express the normalised result in hexadecimal.

6. Calculate the following decimal number subtraction using IEEE single precision floating point arithmetic and express the normalised result in hexadecimal. Describe each step of the calculation procedure clearly

   (a) 125.25 - 75.5 (b) 123.125 - 43.5

7. Any floating point representation used in a computer represents only certain real numbers exactly; all others can only be approximated. If X is the stored value approximate the real value Y, then the relative error, R is expressed as

   $R = (Y - X)/Y$

   Represent the decimal quantity + 0.4 in the following floating-point format: base =2; exponent: biased, 4 bits, significant 7 bits. What is the relative error?