COMP 350 Numerical Computing

Assignment #1: Floating Point Computing

Date Given: Monday, September 9. Date Due: Monday, September 23, 2013

Place your clearly written or printed answers, firmly bound or stapled together, with your Name and Student Number at the front, in a marked COMP 350 box, located on the 2nd floor of Trottier.

1. (2 point) Is there a real number which has finite binary representation but infinite (or non-terminating) decimal representation? Give reasons.

2. (2 points) Using a 32-bit word, how many different integers can be represented by (a) sign and modulus; (b) 2's complement? Express the answer using powers of 2.

3. Suppose in IEEE single precision, the width of the exponent field is 4, not 8, and the width of the fraction field is 5, not 23.

   (a) (1 point) What should the exponent bias be?

   (b) (2 points) What are the largest and smallest nonnegative normalized floating point numbers in this system?

   (c) (2 points) What are the largest and smallest nonnegative subnormal floating point numbers in this system?

   (d) (1 point) What is the machine epsilon of this system.

   (e) (2 points) What are the two floating point numbers which are closest to, but not equal to 10?

   (f) (2 points) Given number $-(11.01011)_2$. Round it using the four rounding modes. Give the answers as normalized floating point numbers, in the form **binary-significand**$\times 2^E$, where E is decimal.

4. (4 points) In exact arithmetic, the addition operation is also associative, i.e.

$$(a + b) + c = a + (b + c)$$

   for any three numbers a, b, c. Is this also true of the floating point addition operator $\oplus$?

   In the above questions, we assume that a, b and c are already floating point numbers and no overflow or underflow occurs in the floating point operations. Give reasons if your answer to any question is "Yes", or give a counterexample if your answer is "No".

5. (2 point) What are the values of the expressions $\infty/0$, $-\infty/\infty$, $0/\infty$, $1/0-2/0$ and sign(-NaN)?

6. (Extra question: 3 points) In the course of solving $ax^2 + 2bx + c = 0$ for x, the expression $\sqrt{b^2 - ac}$ must be computed. Suppose a, b and c are floating point numbers. Can the true value of $b^2 - ac$ be nonnegative and yet its computed value be negative? Give a proof or a counterexample. If you only say "Yes" or "No", you will not get any credit.