1994 Paper 4 Question 9

Numerical Analysis I

With reference to a decimal floating-point implementation with 4-digit precision ($\beta = 10$, $p = 4$), describe the two most common methods of rounding. (Use 1.2345 and 1.2375 as examples.) Which method is unbiased? [3 marks]

What do you understand by the terms *machine epsilon*, and *guard digit*? [4 marks]

Suppose the largest representable floating-point number is about 10^{50}, and consider evaluation of $\sqrt{x^2 - y^2}$. How would you compute the result? (Use $x \simeq 5.10^{40}$, $y \simeq 3.10^{40}$ as an example.) How could your method also improve accuracy on some machines? [3 marks]

A programmer writes $(x+y)+z$ but a compiler evaluates the right-hand side in the form $x + (y + z)$. Explain how this could be harmful in floating-point arithmetic (a) when x, y and z are large, and (b) when x, y and z are numbers of moderate size. Which of these two problems would be more likely to occur in practice: (a) or (b)? [3 marks]

Explain the term NaN as used in IEEE arithmetic. Roughly, how many NaN values are there in IEEE single precision? Consider an *operation* to be any one of $+ - * /$. Give examples of (a) an operation that yields a NaN value when neither of its arguments is a NaN, (b) an operation with finite arguments that yields $+\infty$, (c) an operation with an argument $+\infty$ that yields a finite result. [5 marks]

What two rules govern operations where at least one argument is a NaN value? [2 marks]