2004 Paper 3 Question 6

Numerical Analysis I

(a) For Single Precision in the *IEEE* binary floating-point standard (*IEEE* 754) the *precision* is defined as 24, and the *exponent* requires 8 bits of storage. With reference to *IEEE* Single Precision, explain the terms *exponent*, *significand*, *precision*, *sign bit*, *normalised number*, *denormal number*. [6 marks]

(b) Explain the term *hidden bit*. What are the values of the hidden bit for normalised and denormal numbers? How is the exponent stored and why? How are the exponent, significand and sign bit arranged in memory? [4 marks]

(c) Let x^* denote the floating-point representation of a number x. Define the terms *absolute error* (ε_x) and *relative error* (δ_x) in representing x. How are ε_x and δ_x related? Define *machine epsilon* (ε_m). [3 marks]

(d) Assume $\delta_x = \delta_y = \delta_z = \varepsilon_m$. Using worst-case analysis, estimate δ_{xy}, ε_{xy}. Find an expression for δ_w where $w = z - xy$. [4 marks]

(e) Working to 4 significant decimal digits only, compute w^* when $x^* = 2.018$, $y^* = 2.008$, $z^* = 4.058$. Given $\varepsilon_m \simeq 0.5 \times 10^{-3}$, how many significant decimal digits of w^* can be relied on? [3 marks]