2011 Paper 1 Question 8

Floating-Point Computation

(a) Suppose a floating-point representation for unsigned numbers has a three-bit mantissa and a three-bit exponent. Suggest a useful encoding range by stating the minimum and maximum values representable without a hidden bit. [4 marks]

(b) Modify your answer to part (a) for when a hidden bit is used. [3 marks]

(c) Give roughly the smallest positive IEEE single-precision floating-point number, x, for which $\sin(x)$ is not meaningless. [3 marks]

(d) Describe **four** different rules for rounding a floating-point number and say which is generally used and why. [2 marks]

(e) Give **two** techniques for determining the number of steps used in an iteration. (Do not describe iterating until no change.) Say when one technique is preferred to the other. [4 marks]

(f) A scientific library uses the formula $(a+b+c)/(d+e)$ where $a \ldots e$ are floating-point. Aside from using an `IF` statement to check for division by zero, what further `IF` statement(s) should be included to ensure good precision? [4 marks]