1996 Paper 3 Question 10

Numerical Analysis I

Explain how the 32 bits are arranged to store the *sign*, *exponent*, and *significand* in a single precision number under the IEEE binary floating-point standard (IEEE 754). How is the exponent stored? [2 marks]

Explain the terms e_{max}, e_{min}, *normalized number*, *denormal number*, *hidden bit*, *NaN*. [5 marks]

In terms of the stored bit-pattern, how can each of the following be recognized: ± 0, $\pm \infty$, *normalized number*, *denormal number*, *NaN*? [5 marks]

Suppose the principles of IEEE arithmetic are applied to a floating-point representation with only 6 stored bits. If the precision p is 3 (including the hidden bit) and e_{min} is -2, what is e_{max}? [2 marks]

List the 16 bit-patterns which do not represent normalized numbers, and identify what each represents. [6 marks]