p3q10
p10q12
MROD

Numerical Analysis I

Acceptable Answer A

A floating point number may be represented in the form

$$\pm\, d_0 . d_1 d_2 \ldots d_{p-1} \times \beta^e$$

where the p significant digits of base β, $d_0 . d_1 d_2 \ldots d_{p-1}$, form the <u>significand</u>. The <u>sign bit</u> represents the sign of the significand as 0 (positive) or 1 (negative). The <u>exponent</u> is e, characterised by its minimum and maximum values e_{min}, e_{max}. A <u>normalised number</u> is in the unique form such that $d_0 \neq 0$. A <u>denormal number</u> is a number that can be represented with exponent e_{min} but which is too small to be normalised. Denormal numbers are conventionally stored with exponent $e_{min} - 1$. In IEEE binary arithmetic the value of d_0 can be deduced from the exponent, so is not stored. This is called the <u>hidden bit</u>. The <u>precision</u> is the number p, the number of digits of the significand including the hidden bit. The hidden bit has the value 1 for normalised numbers, or 0 for denormal numbers. [8 marks]

The parts of each number are stored in left-to-right order as: sign bit, exponent, significand. The exponent e is stored as the bit pattern with value $e + e_{max}$.

As $e_{min} = -1022$, $e_{max} = 1023$ there are 2048 different exponents, including the reserved values $e_{min} - 1$ and $e_{max} + 1$. This requires 11 bits of storage. The total number of bits required is therefore given by

1	+	11	+	52	=	64
sign bit		exponent		stored significand $= p - 1$		total

[5 marks]

PTO.

	Sign	$e =$ stored exponent $- e_{max}$	significand	value
(a)	+	$e_{max}+1$	0	$+\infty$
(b)	−	1	1	-2
(c)	+	$e_{max}+1$	$\neq 0$	Not a Number
(d)	−	$e_{max}+1$	$\neq 0$	Not a Number
(e)	−	$e_{min}-1$	0	-0
(f)	+	$e_{min}-1$	$\frac{1}{2}$	$+2^{-1023}$
(g)	+	0	1	$+1$

[7 marks]