Problem Set #1
Multiple Choice Test
Chapter 01.05 Floating Point Representation
COMPLETE SOLUTION SET

1. A hypothetical computer stores real numbers in floating point format in 8-bit words. The first bit is used for the sign of the number, the second bit for the sign of the exponent, the next two bits for the magnitude of the exponent, and the next four bits for the magnitude of the mantissa. Represent $e \approx 2.718$ in the 8-bit format.

(A) 00010101
(B) 00011010
(C) 00010011
(D) 00101010

Solution
The correct answer is (A).

Finding $(2)_{10} = (?)_2$

	Quotient	Remainder
2/2	1	0
1/2	0	1

Hence $(2)_{10} = (10)_2$

Finding $(0.718)_{10} = (?)_2$

	Number	Number after decimal	Number before decimal
0.718×2	1.436	0.436	1
0.436×2	0.872	0.872	0
0.872×2	1.744	0.744	1
0.744×2	1.488	0.488	1

$$(0.718)_{10} \approx (0.101)_2$$
$$(2.718)_{10} \approx (10.101)_2$$
$$= (1.0101)_2 \times 2^{(1)_{10}}$$
$$= (1.0101)_2 \times 2^{(01)_2}$$

Bits in mantissa $= 0101$
Bits in exponent $= 01$
Sign number bit $= 0$
Sign exponent bit $= 0$
The final floating point representation is 00010101.

2. A hypothetical computer stores real numbers in floating point format in 8-bit words. The first bit is used for the sign of the number, the second bit for the sign of the exponent, the next two bits for the magnitude of the exponent, and the next four bits for the magnitude of the mantissa. The base-10 number $(10100111)_2$ represents in the above given 8-bit format is

(A) -5.75
(B) -2.875
(C) -1.75
(D) -0.359375

Solution
The correct answer is (A).

The number is 10100111
Sign number bit $= 1$
Sign of number is negative
Sign exponent bit $= 0$
Sign of exponent is positive
Bits in exponent $= 10$

$$(10)_2 = 1\times 2^1 + 0\times 2^0$$
$$= (2)_{10}$$

Bits in mantissa $= 0111$

$$(1.0111)_2 = 1\times 2^0 + 0\times 2^{-1} + 1\times 2^{-2} + 1\times 2^{-3} + 1\times 2^{-4}$$
$$= (1.4375)_{10}$$

Hence, the number in base-10 is

$$= -1.4375\times 2^2$$
$$= -5.75$$

3. A hypothetical computer stores floating point numbers in 8-bit words. The first bit is used for the sign of the number, the second bit for the sign of the exponent, the next two bits for the magnitude of the exponent, and the next four bits for the magnitude of the mantissa. The machine epsilon is most nearly

(A) 2^{-8}
(B) 2^{-4}
(C) 2^{-3}
(D) 2^{-2}

Solution
The correct answer is (B).

The machine epsilon is

$$\varepsilon_{mach} = 2^{-(\text{number of bits for mantissa})}$$
$$= 2^{-4}$$

4. A machine stores floating point numbers in 7-bit word. The first bit is used for the sign of the number, the next three for the biased exponent and the next three for the magnitude of the mantissa. The number $(0010110)_2$ represented in base-10 is

(A) 0.375
(B) 0.875
(C) 1.5
(D) 3.5

Solution
The correct answer is (B).

Sign bit of number $= 0$
Hence, the number is positive
Bits in biased exponent $= 010$

Thus, it is biased by 3 as the maximum value in the biased exponent is $(111)_2 = 7$. So the exponent is $(010)_2 - 3 = 2 - 3 = -1$.
Bits in mantissa $= 110$

$$(1.110)_2 = 1\times 2^0 + 1\times 2^{-1} + 1\times 2^{-2} + 0\times 2^{-3}$$
$$= (1.75)_{10}$$

Hence, the number in base-10 is

$$= (1.75)\times 2^{-1}$$
$$= (0.875)_{10}$$

5. A machine stores floating point numbers in 7-bit word. The first bit is stored for the sign of the number, the next three for the biased exponent and the next three for the magnitude of the mantissa. You are asked to represent 33.35 in the above word. The error you will get in this case would be

(A) underflow
(B) overflow
(C) NaN
(D) No error will be registered

Solution

The correct answer is (B).

$(33)_{10} = (?????)_2$

	Quotient	Remainder
33/2	16	1
16/2	8	0
8/2	4	0
4/2	2	0
2/2	1	0
1/2	0	1

Therefore,

$(33)_{10} = (100001)_2$

	Number	Number after decimal	Number before decimal
0.35×2	0.70	0.70	0
0.70×2	1.40	0.40	1
0.40×2	0.80	0.80	0

$(0.35)_{10} \approx (0.01\ldots)_2$

$(33.35)_{10} \approx (100001.01)_2$

$= (1.0000101)_2 \times 2^5$

$= (1.0000101)_2 \times 2^{(101)_2}$

The biased exponent has 3 bits, so the biased exponent varies from 0 to $(111)_2 = (7)_{10}$. So the exponent can vary from -3 to 4. Hence the number $(33.35)_{10}$ which has an exponent of 5 would overflow.

6. A hypothetical computer stores floating point numbers in 9-bit words. The first bit is used for the sign of the number, the second bit for the sign of the exponent, the next three bits for the magnitude of the exponent, and the next four bits for the magnitude of the mantissa. Every second, the error between 0.1 and its binary representation in the 9-bit word is accumulated. The accumulated error after one day most nearly is

(A) 0.002344
(B) 20.25
(C) 202.5
(D) 8640

Solution

The correct answer is C.

	Number	Number after decimal	Number before decimal
0.1×2	0.2	0.2	0
0.2×2	0.4	0.4	0
0.4×2	0.8	0.8	0
0.8×2	1.6	0.6	1
0.6×2	1.2	0.2	1
0.2×2	0.4	0.4	0
0.4×2	0.8	0.8	0
0.8×2	1.6	0.6	1
0.6×2	1.2	0.2	1

$$\begin{aligned}(0.1)_{10} &= (0.000110011\ldots)_2 \\ &\approx (1.1001)_2 \times 2^{-4} \\ &= (1.1001)_2 \times 2^{-(100)_2}\end{aligned}$$

Thus, $(0.1)_{10}$ is represented approximately by

$$\begin{aligned}(0.1)_{10} &\approx \left(1\times 2^0 + 1\times 2^{-1} + 0\times 2^{-2} + 0\times 2^{-3} + 1\times 2^{-4}\right)\times 2^{-4} \\ &= 0.09765625\end{aligned}$$

The difference is $= 0.1 - 0.09765625$
$= 0.00234375$

This difference is accumulated every second for one day, giving the accumulated error as

$= 0.00234375\times 60\times 60\times 24$
$= 202.5$