PROBLEMAS DE ARITMÉTICA EN PUNTO FLOTANTE

1.- Demostrar con un ejemplo que la suma en punto flotante no es siempre asociativa.

2.- Al sumar dos números en punto flotante en el formato IEEE 754, A y B, el resultado es A. ¿Implica esto que B=0?.

3.- Redondear, por los cuatro métodos, los siguientes números expresados en IEEE 754, suponiendo que los bits de guarda, redondeo y sticky son los dados en la tabla (grs):

S	E	M	grs
0	00011111	11111111111111111111111	100
0	11111110	11111111111111111111111	100
1	11111110	11111111111111111111111	100

4.- Se desea transformar cualquier número en punto flotante de 16 bits ("media precisión") a punto flotante de 32 bits ("simple precisión") utilizando una ROM.

Se pide:

(a) Determinar organización y tamaño, especificando el significado de sus entradas y salidas, de la ROM necesaria para realizar la transformación descrita.
(b) Indicar el contenido de la ROM anterior en las direcciones 08C3h y 803Bh.
(c) Suponiendo que se cuenta con compuertas lógicas y sumadores de n bits, implementar un circuito combinatorio que transforme números en punto flotante de 16 bits normalizados al correspondiente en punto flotante de 32 bits.

5.- El formato IEEE 754 define una precisión doble de 64 bits utilizando 53 bits para la mantisa (incluyendo el 1 implícito) y un exponente de 11 bits. IA-32 ofrece una precisión extendida con 64 bits para la mantisa y 16 bits para el exponente.

1. Asumiendo que una precisión extendida es similar a las precisiones simple y doble, ¿cuál es el sesgo en el exponente?
2. ¿Cuál es el rango de números que puede representarse utilizando la precisión extendida?
3. ¿Cuánto mayor (en porcentaje) es la precisión extendida comparada con la doble precisión?

6.- Supongamos un formato IEEE-754 reducido, con 11 bits, de los cuales 4 son de exponente, determinar:
a) ¿Cuál es el mayor número positivo representable?
b) ¿Cuál es el menor número positivo representable?
c) ¿Qué valor representan los siguientes números?
- 10111000001
- 00000000000
- 00000010000
- 01111011000
- 01100111100

d) Sumar: A+B
A= 01001000000
B= 00101111110
e) Sumar A+B
A= 01001000000
B= 00101110110
f) Sumar A + B
A= 01001111111
B= 00101101000

7.- Sumar, siguiendo los pasos dados en clase, los siguientes números representados en el estándar IEEE-754, redondear por los cuatro métodos:
A=01010111001
B=00111100110

8.- Realizar las siguientes operaciones:
A+B
A-B
A*B
A/B
A**2
Siendo:
1. A= 10011010101
   B=00101110000
2. A=10000111111
   B=10001110000
3. A= 11111000000
   B=00000000000