Plancha 0

Representación Computacional de datos

2022 - Arquitectura del Computador

Licenciatura en Ciencias de la Computación

1

Introducción

El objetivo de esta plancha de ejercicios es familiarizarse con la representación de datos orientada a la computación pero independizándose de la implementación computacional propiamente dicha. Son ejercicios para resolver con papel y lápiz aunque algunos ejercicios luego pueden realizarse utilizando computadora para chequear resultados.

2

Procedimiento

Resuelva cada ejercicio en papel. Luego, se puede subir la resolución escaneándola (ser claros con la letra) o utilizando algún editor de texto.

3

Ejercicios

1. Utilizando el sistema de numeración posicional $(-1)^{\sigma}\left(a_n a_{n-1} \ldots a_1 a_0, a_{-1} a_{-2} \ldots\right)_{\beta}$ con $\beta = 2$, determinar la representación binaria de los siguientes números:

   a) 29
   b) 0,625
   c) 0,1
   d) 5,75
   e) -138
   f) -15,125

   Analizar en cada caso cuántos dígitos son necesarios para poder representar cada uno de los números.

   ### Soluciones

   a) Son necesarios 5 bits para el número y uno extra para el signo:
   - a_0: $29 = 2 \cdot 14 + \boxed{1}$.
   - a_1: $14 = 2 \cdot 7 + \boxed{0}$.
   - a_2: $7 = 2 \cdot 3 + \boxed{1}$.
   - a_3: $3 = 2 \cdot 1 + \boxed{1}$.
   - a_4: $1 = 2 \cdot 0 + \boxed{1}$.
   - $(-1)^0\,(11101)_2$.

   b) Son necesarios 4 bits para el número y uno extra para el signo:
   - a_{-1}: $0,625 \cdot 2 = \boxed{1},25$.
   - a_{-2}: $0,25 \cdot 2 = \boxed{0},5$.
   - a_{-3}: $0,5 \cdot 2 = \boxed{1}$.
   - $(-1)^0\,(0,101)_2$.

c) Son necesarios infinitos bits:

- a_{-1}: $0,1 \cdot 2 = \boxed{0},2$.
- a_{-2}: $0,2 \cdot 2 = \boxed{0},4$.
- a_{-3}: $0,4 \cdot 2 = \boxed{0},8$.
- a_{-4}: $0,8 \cdot 2 = \boxed{1},6$.
- a_{-5}: $0,6 \cdot 2 = \boxed{1},2$.
- $(-1)^0 \left(0,0\overline{0011}\right)_2$.

d) Son necesarios 5 bits para el número y uno extra para el signo:

- a_{-1}: $0,75 \cdot 2 = \boxed{1},5$.
- a_{-2}: $0,5 \cdot 2 = \boxed{1}$.
- a_0: $5 = 2 \cdot 2 + \boxed{1}$.
- a_1: $2 = 2 \cdot 1 + \boxed{0}$.
- a_2: $1 = 2 \cdot 0 + \boxed{1}$.
- $(-1)^0 (101,11)_2$.

e) Son necesarios 8 bits para el número y uno extra para el signo:

- a_0: $138 = 2 \cdot 69 + \boxed{0}$.
- a_1: $69 = 2 \cdot 34 + \boxed{1}$.
- a_2: $34 = 2 \cdot 17 + \boxed{0}$.
- a_3: $17 = 2 \cdot 8 + \boxed{1}$.
- a_4: $8 = 2 \cdot 4 + \boxed{0}$.
- a_5: $4 = 2 \cdot 2 + \boxed{0}$.
- a_6: $2 = 2 \cdot 1 + \boxed{0}$.
- a_7: $1 = 2 \cdot 0 + \boxed{1}$.
- $(-1)^1 (10001010)_2$.

f) Son necesarios 7 bits para el número y uno extra para el signo:

- a_{-1}: $0,125 \cdot 2 = \boxed{0},25$.
- a_{-2}: $0,25 \cdot 2 = \boxed{0},5$.
- a_{-3}: $0,5 \cdot 2 = \boxed{1}$.
- a_0: $15 = 2 \cdot 7 + \boxed{1}$.
- a_1: $7 = 2 \cdot 3 + \boxed{1}$.
- a_2: $3 = 2 \cdot 1 + \boxed{1}$.
- a_3: $1 = 2 \cdot 0 + \boxed{1}$.
- $(-1)^1 (1111,001)_2$.

2. Convertir los siguientes números decimales a binario utilizando la representación en complemento a dos con seis bits:

   a) -16
   b) 13
   c) -1
   d) -10
   e) 16
   f) -31

   ¿Qué tienen en común todos los números negativos y todos los números positivos al utilizar esta representación?

   Soluciones

   a) $(-16)_{10} = (-010000)_2 = (110000)_{C_2^6}$.
   b) $(13)_{10} = (001101)_2 = (001101)_{C_2^6}$.
   c) $(-1)_{10} = (-000001)_2 = (111111)_{C_2^6}$.
   d) $(-10)_{10} = (-001010)_2 = (110110)_{C_2^6}$.
   e) $(16)_{10} = (010000)_2 = (010000)_{C_2^6}$.
   f) $(-31)_{10} = (-011111)_2 = (100001)_{C_2^6}$.

   Los números negativos en complemento a 2, tienen el bit mas significativo en 1 mientras que los positivos lo tienen en 0.

3. Repetir el ejercicio anterior pero ahora utilizando 8 bits. ¿Qué conclusión se puede sacar comparando los resultados con los del ejercicio anterior?

   Soluciones

   a) $(-16)_{10} = (-00010000)_2 = (11110000)_{C_2^8}$.
   b) $(13)_{10} = (00001101)_2 = (00001101)_{C_2^8}$.
   c) $(-1)_{10} = (-00000001)_2 = (11111111)_{C_2^8}$.
   d) $(-10)_{10} = (-00001010)_2 = (11110110)_{C_2^8}$.
   e) $(16)_{10} = (00010000)_2 = (00010000)_{C_2^8}$.
   f) $(-31)_{10} = (-00011111)_2 = (11100001)_{C_2^8}$.

   Basta con agregar «unos» a la izquierda para los números negativos y «ceros» para los positivos.

4. Dadas las siguientes secuencias de bits, indicar a qué números corresponden en sistema decimal utilizando la representación en complemento a dos:

   a) $(00001101)_{C_2^8}$
   b) $(01001101)_{C_2^8}$
   c) $(11100001)_{C_2^8}$
   d) $(11111001)_{C_2^8}$
   e) $(11111111)_{C_2^8}$
   f) $(00000000)_{C_2^8}$

   Soluciones

   a) $(00001101)_{C_2^8} = 2^0 + 2^2 + 2^3 = 1 + 4 + 8 = (13)_{10}$.
   b) $(01001101)_{C_2^8} = 2^0 + 2^2 + 2^3 + 2^6 = 1 + 4 + 8 + 64 = (77)_{10}$.
   c) $(11100001)_{C_2^8} = 2^0 + 2^5 + 2^6 - 2^7 = 1 + 32 + 64 - 128 = (-31)_{10}$.
   d) $(11111001)_{C_2^8} = 2^0 + 2^3 + 2^4 + 2^5 + 2^6 - 2^7 = 1 + 8 + 16 + 32 + 64 - 128 = (-7)_{10}$.
   e) $(11111111)_{C_2^8} = 2^0 + 2^1 + 2^2 + 2^3 + 2^4 + 2^5 + 2^6 - 2^7 = 1 + 2 + 4 + 8 + 16 + 32 + 64 - 128 = (-1)_{10}$.
   f) $(00000000)_{C_2^8} = (0)_{10}$.

5. Mostrar que $(13,25)_{10} = (1101,01)_2 = (15,2)_8 = (D,4)_{16}$.

   Solución

   - $(13,25)_{10} = (1101,01)_2$:
     - a_{-1}: $0,25 \cdot 2 = \boxed{0},5$.
     - a_{-2}: $0,5 \cdot 2 = \boxed{1}$.
     - a_0: $13 = 2 \cdot 6 + \boxed{1}$.
     - a_1: $6 = 2 \cdot 3 + \boxed{0}$.
     - a_2: $3 = 2 \cdot 1 + \boxed{1}$.
     - a_3: $1 = 2 \cdot 0 + \boxed{1}$.

- $(13,25)_{10} = (15,2)_8$:
  - a_{-1}: $0,25 \cdot 8 = \boxed{2}$.
  - a_0: $13 = 8 \cdot 1 + \boxed{5}$.
  - a_1: $1 = 8 \cdot 0 + \boxed{1}$.
- $(13,25)_{10} = (D,4)_{16}$:
  - a_{-1}: $0,25 \cdot 16 = \boxed{4}$.
  - a_0: $13 = 16 \cdot 0 + 13 = \left(\boxed{D}\right)_{16}$.

6. Rellenar la siguiente tabla:

Binario	Octal	Hexadecimal	Decimal
1101100,110			
	362,2		
		A1,8	
			74,125

En cada fila verán un valor numérico expresado en la base que indica la casilla superior de la columna donde se encuentra. Completar en el resto de casillas con la representación correspondiente según la base indicada en la parte superior de manera que se mantengan las equivalencias en cada fila. Asumir que los números son sin signo.

Solución

Binario	Octal	Hexadecimal	Decimal
1101100,110	154,6	6C,C	108,75
11110010,01	362,2	F2,4	242,5
10100001,1	241,4	A1,8	161,5
1001010,001	112,1	4A,2	74,125

7. Determinar el formato hexadecimal que use el mínimo número de dígitos y que permita representar el número $(16,25)_{10}$ de manera exacta. ¿Cuál es el rango y la precisión del formato? Asumir números sin signo.

Solución Observemos que $(16,25)_{10} = (10,4)_{16}$, luego nos bastará un formato hexadecimal de punto fijo con dos dígitos para la parte entera y uno para la parte fraccionaria.

Este formato permite expresar numeros entre $0 \leq N \leq (FF,F)_{16} = (255,9375)_{10}$ y tiene una presición de $16^{-1} = 0,0625$.

8. Realizar cada una las siguientes operaciones, como lo realizaría una computadora, usando registros de 8 bits:

   a) 10 - 3
   b) -39 + 92
   c) -19 - 7
   d) 44 + 45
   e) 104 + 45
   f) -75 + 59
   g) -103 - 69
   h) 127 + 1
   i) -1 + 1
   j) -1 - 1

   - En cada una de la operaciones anteriores, analizar si los resultados son correctos/incorrectos y por qué. Ubicar los resultados obtenidos en la recta real, incluyendo los valores mínimos y máximos para esta representación.
   - Indicar el estado de las banderas *Carry* y *Overflow* luego de realizar cada una de las operaciones anteriores.

Soluciones

a) Observemos que $(10)_{10} = (00001010)_{C_2^8}$ y $(3)_{10} = (00000011)_2 = (11111101)_{C_2^8}$, luego:

$$\begin{array}{cccccccccc} & \mathbf{0} & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ + & \mathbf{1} & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ \hline 1 & \mathbf{0} & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{array}$$

- El resultado es correcto pues $(00000111)_{C_2^8} = (7)_{10} = 10 - 3$.
- La bandera *Carry Flag* se enciende y la *Overflow Flag* se **apaga**.

b) Observemos que $(-39)_{10} = (-00100111)_2 = (11011001)_{C_2^8}$ y $(92)_{10} = (01011100)_{C_2^8}$, luego:

$$\begin{array}{cccccccccc} & \mathbf{1} & 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ + & \mathbf{0} & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ \hline 1 & \mathbf{0} & 0 & 1 & 1 & 0 & 1 & 0 & 1 \end{array}$$

- El resultado es correcto pues $(00110101)_{C_2^8} = (53)_{10} = -39 + 92$.
- La bandera *Carry Flag* se enciende y la *Overflow Flag* se **apaga**.

c) Observemos que $(-19)_{10} = (-00010011)_2 = (11101101)_{C_2^8}$ y $(-7)_{10} = (-00000111)_2 = (11111001)_{C_2^8}$, luego:

$$\begin{array}{ccccccccc} & \mathbf{1} & 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ + & \mathbf{1} & 1 & 1 & 1 & 1 & 0 & 0 & 1 \\ \hline 1 & \mathbf{1} & 1 & 1 & 0 & 0 & 1 & 1 & 0 \end{array}$$

- El resultado es correcto pues $(11100110)_{C_2^8} = (-26)_{10} = -19 - 7$.
- La bandera *Carry Flag* se enciende y la *Overflow Flag* se **apaga**.

d) Observemos que $(44)_{10} = (00101100)_2$ y $(45)_{10} = (00101101)_2$, luego:

$$\begin{array}{ccccccccc} & \mathbf{0} & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ + & \mathbf{0} & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ \hline 0 & \mathbf{0} & 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{array}$$

- El resultado es correcto pues $(01011001)_{C_2^8} = (89)_{10} = 44 + 45$.
- La bandera *Carry Flag* se apaga y la *Overflow Flag* se **apaga**.

e) Observemos que $(104)_{10} = (01101000)_{C_2^8}$ y $(45)_{10} = (00101101)_{C_2^8}$, luego:

$$\begin{array}{ccccccccc} & \mathbf{0} & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ + & \mathbf{0} & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ \hline 0 & \mathbf{1} & 0 & 0 & 1 & 0 & 1 & 0 & 1 \end{array}$$

- El resultado es correcto pues $(10010101)_2 = (-107)_{10} \neq (149)_{10} = 104 + 45$.
- La bandera *Carry Flag* se apaga y la *Overflow Flag* se ***enciende***.

f) Observemos que $(-75)_{10} = (-01001011)_2 = (10110101)_{C_2^8}$ y $(59)_{10} = (00111011)_{C_2^8}$, luego:

$$\begin{array}{ccccccccc} & \mathbf{1} & 0 & 1 & 1 & 0 & 1 & 0 & 1 \\ + & \mathbf{0} & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ \hline 0 & \mathbf{1} & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{array}$$

- El resultado es correcto pues $(11110000)_{C_2^8} = (-16)_{10} = -75 + 59$.
- La bandera *Carry Flag* se apaga y la *Overflow Flag* se **apaga**.

g) Observemos que $(-103)_{10} = (-01100111)_2 = (10011001)_{C_2^8}$ y $(-69)_{10} = (-01000101)_2 = (10111011)_{C_2^8}$, luego:

$$\begin{array}{ccccccccc} & \mathbf{1} & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ + & \mathbf{1} & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ \hline 1 & \mathbf{0} & 1 & 0 & 1 & 0 & 1 & 0 & 0 \end{array}$$

- El resultado es incorrecto pues $(01010100)_{C_2^8} = (84)_{10} \neq -172 = -19 - 7$.
- La bandera *Carry Flag* se enciende y la *Overflow Flag* se **enciende**.

h) Observemos que $(127)_{10} = (01111111)_{C_2^8}$ y $(1)_{10} = (00000001)_{C_2^8}$, luego:

$$\begin{array}{ccccccccc} & \mathbf{0} & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ + & \mathbf{0} & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}$$

- El resultado es incorrecto pues $(10000000)_{C_2^8} = (-1)_{10} \neq 128 = 127 + 1$.
- La bandera *Carry Flag* se apaga y la *Overflow Flag* se **enciende**.

i) Observemos que $(-1)_{10} = (11111111)_{C_2^8}$ y $(1)_{10} = (00000001)_{C_2^8}$, luego:

$$\begin{array}{ccccccccc} & \mathbf{1} & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ + & \mathbf{0} & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 1 & \mathbf{0} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}$$

- El resultado es correcto pues $(00000000)_{C_2^8} = (0)_{10} = -1 + 1$.
- La bandera *Carry Flag* se enciende y la *Overflow Flag* se **apaga**.

j) Observemos que $(-1)_{10} = (11111111)_{C_2^8}$, luego:

$$\begin{array}{ccccccccc} & \mathbf{1} & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ + & \mathbf{1} & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \hline 1 & \mathbf{1} & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{array}$$

- El resultado es correcto pues $(11111110)_{C_2^8} = (-2)_{10} = -1 - 1$.
- La bandera *Carry Flag* se enciende y la *Overflow Flag* se **apaga**.

9. Repetir el ejercicio 1 pero ahora utilizando el sistema de numeración binario en complemento a dos para los números enteros y la norma IEEE 754 para los números con parte fraccionaria. Comparando con los resultados anteriores, ¿qué conclusiones se pueden sacar?

SOLUCIONES

a) $(29)_{10} = (-1)^0\,(11101)_2 = (011101)_{C_2^6}$.

b) $(0,625)_{10} = (-1)^0 \cdot (0,101)_2 \cdot 2^0 = (-1)^{0} \cdot (1,01)_2 \cdot 2^{-1}$.

- $\sigma = 0$.
- $E = -1 + 127 = 126 = (01111110)_2$.
- $m = (01000000000000000000000)_2$.
- $(0,625)_{10} = \left(\underbrace{0}_{\sigma}\underbrace{01111110}_{E}\underbrace{01000000000000000000000}_{m}\right)_{IEEE754}$.

c) $(0,1)_{10} = (-1)^0\,(0,000\overline{1100})_2 \cdot 2^0 = (-1)^{0}\,(1,\overline{1001})_2 \cdot 2^{-4}$.

- $\sigma = 0$.
- $E = -4 + 127 = 123 = (01111011)_2$.
- $m = (1001100110011001100110\mathbf{1})_2$.
- $(0,1)_{10} = \left(\underbrace{0}_{\sigma}\underbrace{01111011}_{E}\underbrace{1001100110011001100110\mathbf{1}}_{m}\right)_{IEEE754}$.

d) $(5,75)_{10} = (-1)^0\,(101,11)_2 \cdot 2^0 = (-1)^{0}\,(1,0111)_2 \cdot 2^{2}$.

- $\sigma = 0$.
- $E = 2 + 127 = 129 = (10000001)_2$.
- $m = (01110000000000000000000)_2$.
- $(5,75)_{10} = \left(\underbrace{0}_{\sigma}\underbrace{10000001}_{E}\underbrace{01110000000000000000000}_{m}\right)_{IEEE754}$.

e) $(-138)_{10} = (-1)^1\,(10001010)_2 = (101110110)_{C_2^9}$.

f) $(-15,125)_{10} = (-1)^1\,(1111,001)_2 \cdot 2^0 = (-1)^{1}\,(1,111001)_2 \cdot 2^{3}$.

- $\sigma = 1$.
- $E = 3 + 127 = 130 = (10000010)_2$.
- $m = (11100100000000000000000)_2$.
- $(15,125)_{10} = \left(\underbrace{1}_{\sigma}\underbrace{10000010}_{E}\underbrace{11100100000000000000000}_{m}\right)_{IEEE754}$.

10. Considere el conjunto de números de punto flotante $\mathbb{F}\,(2,3,-1,2)$:

    a) Determinar $x_{\text{mín}}, x_{\text{máx}}, \epsilon_M$ y el número de elementos de $\mathbb{F}$.

    b) Determinar los números de punto flotante positivos del conjunto $\mathbb{F}$.

    c) Graficar sobre la recta real los números de puntos flotantes determinados en el punto anterior.

Ayuda Recordar la notación $\mathbb{F}\,(\beta, t, L, U)$, donde β es la base, t es la cantidad de dígitos significativos de la mantisa, L es el valor mínimo del exponente y U es el valor máximo del exponente.

Soluciones

a) Para $\mathbb{F}(2, 3, -1, 2)$ resulta:
- $x_{\text{mín}} = 2^{-1} \cdot 2^{-1} = 2^{-2} = 1/4 = 0,25$.
- $x_{\text{máx}} = 2^{2}\left(1 - 2^{-3}\right) = 4\,(1 - 1/8) = 28/8 = 7/2 = 3,5$.
- $\epsilon_M = 2^{1-t} = 2^{1-3} = 2^{-2} = 1/4 = 0,25$.
- $|\mathbb{F}(2, 3, -1, 2)| = 2\,(2-1)\,2^{3-1}\,(2 - (-1) + 1) + 1 = 33$.

b) COMPLETAR.

c) COMPLETAR.

11. Determinar si el número $(2,89)_{10} \cdot 10^{10}$ es representable en un formato de coma flotante de 16 bits, con mantisa normalizada de la forma $1, b_{-1}b_{-2}\ldots$, bit implícito y 7 bits para el exponente.

Solución Observemos que

$$\begin{aligned}(2,89)_{10} \cdot 10^{10} &\approx (2,89)_{10} \cdot 2^{33,22} = (2,89)_{10} \cdot 2^{33} \cdot 2^{0,22} \approx (3,3661)_{10} \cdot 2^{33} \\ &\approx (11,010111011011100011)_2 \cdot 2^{33} \\ &= (1,1010111011011100011)_2 \cdot 2^{34}\end{aligned}$$

luego como tenemos reservados 7 bits para el exponente el rango es $[-63, 64]$ con sesgo 63 por lo que:

$$(2,89)_{10} \cdot 10^{10} \approx \left(\underbrace{0}_{\sigma}\ \underbrace{1100001}_{E}\ \underbrace{10101111}_{m}\right)$$

12. Dados los siguientes números representados en punto flotante IEEE 754 simple precisión, indicar a qué número en formato decimal corresponden y analizar si son números normalizados:

a) $N_1 = (11000010111011010100000000000000)_{IEEE754}$.

b) $N_2 = (0x40600000)_{IEEE754}$.

c) $N_3 = (0x00600000)_{IEEE754}$.

Soluciones

a) Para $N_1 = (11000010111011010100000000000000)_{IEEE754}$ tenemos:
- $(10000101)_2 = 133$.
- $e = 133 - 127 = 6$.
- $(1,110110101)_2 = 2^0 + 2^{-1} + 2^{-2} + 2^{-4} + 2^{-5} + 2^{-7} + 2^{-9} = 1,853515625$
- $N_1 = -\,(1,853515625) \cdot 2^6 = -118,625$.

b) Para $N_2 = (0x40600000)_{IEEE754} = (01000000011000000000000000000000)_{IEEE754}$ tenemos:

- $(10000000)_2 = 128$.
- $e = 128 - 127 = 1$.
- $(1,11)_2 = 2^0 + 2^{-1} + 2^{-2} = 1,75$
- $N_2 = 1,75 \cdot 2^1 = 3,5$.

c) Para $N_3 = (0x00600000)_{IEEE754} = (00000000011000000000000000000000)_{IEEE754}$ tenemos:

- $(00000000)_2 = 0$.
- Como N_3 es un numero denormalizado ($E = 0$) entonces $e = -126$.
- $(0,11) = 2^{-1} + 2^{-2} = 0,75$.
- $N_3 = 0,75 \cdot 2^{-126} \approx 8,81620763 \cdot 10^{-39}$.

13. Determinar la representación de punto flotante de π utilizando 4 dígitos. Utilizar la norma IEEE 754 simple precisión.

Solución Redondeamos $\pi \approx 3,1416$, luego observemos que $(3)_{10} = (11)_2$ y $(0,1416)_{10} \approx (0,001001000011111111001)_2$, luego:

- $\pi \approx (11,001001000011111111001)_2 \cdot 2^0 = (1,10010010000111111111001)_2 \cdot 2^1$.
- $\sigma = 0$.
- $E = 1 + 127 = 128 = (10000000)_2$.
- $m = (10010010000111111111001)_2$.
- $(3,1416)_{10} \approx \left(\underbrace{0}_{\sigma} \underbrace{10000000}_{E} \underbrace{10010010000111111111001}_{m} \right)_{IEEE754}$.

14. Realizar la suma 0,1 + 0,2 utilizando aritmética en punto flotante con norma IEEE 754 simple precisión. Luego realizar la suma 0,1 + 0,4. ¿Qué se puede observar?

Ayuda Al realizar las conversiones se puede reducir la cantidad de operaciones observando la periodicidad de los resultados

Solución Observemos que:

- $$\begin{aligned}(0,1)_{10} &\approx (00111101110011001100110011001101)_{IEEE754} = \\ &= (-1)^0\, 2^{-4}\, (1,10011001100110011001101)_2 \approx \\ &\approx (-1)^0\, 2^{-3}\, (0,11001100110011001100110)_2 \approx \\ &\approx (-1)^0\, 2^{-2}\, (0,01100110011001100110011)_2\end{aligned}$$
- $$\begin{aligned}(0,2)_{10} &\approx (00111110010011001100110011001101)_{IEEE754} = \\ &= (-1)^0\, 2^{-3}\, (1,10011001100110011001101)_2\end{aligned}$$

luego

$$\begin{array}{cccccccccccccccccccccccc} & 0, & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ + & 1, & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ \hline 1 & 0, & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{array}$$

por lo que

$$\begin{aligned}(0,1)_{10} + (0,2)_{10} &\approx (-1)^0\, 2^{-3}\, (10,01100110011001100110011) = \\ &= (-1)^0\, 2^{-2}\, (1,001100110011001100110011) \approx \\ &\approx (00111110100110011001100110011010)_{IEEE754} \approx \\ &\approx (0,300000011921)_{10}\end{aligned}$$

Además

- $$\begin{aligned}(0,4)_{10} &\approx (00111110110011001100110011001101)_{IEEE754} \approx \\ &\approx (-1)^0\, 2^{-2}\, (1,10011001100110011001101)_2\end{aligned}$$

luego

$$\begin{array}{cccccccccccccccccccccccc} & 0, & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ + & 1, & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ \hline 1 & 0, & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}$$

por lo que

$$\begin{aligned}(0,1)_{10} + (0,4)_{10} &= (-1)^0\, 2^{-2}\, (10) = (-1)^0\, 2^{-1}\, (1,0) = 0,5 = \\ &= (00111111000000000000000000000000)_{IEEE754}\end{aligned}$$

15. Efectuar los siguientes cálculos utilizando aritmética en punto flotante con norma IEEE 754 simple precisión, siendo $a = 12345$, $b = 0,0001$, $c = 45,5$:

    - $(a \oplus b) \oplus c$
    - $a \oplus (b \oplus c)$

    Al comparar los resultados en cada una de las operaciones, ¿qué puede concluirse?

Solución Observemos que:

- $$\begin{aligned}(12345)_{10} &\approx (01000110010000001110010000000000)_{IEEE754} \approx \\ &\approx (-1)^0\, 2^{13}\, (1,1000000111001)_2\end{aligned}$$
- $$\begin{aligned}(0,0001)_{10} &\approx (00111000110100011011011100010111)_{IEEE754} \approx \\ &\approx (-1)^0\, 2^{-14}\, (1,10100011011011100010111)_2 \approx \\ &\approx (-1)^0\, 2^{5}\, (0,00000000000000000011010)_2 \neq \\ &\neq (-1)^0\, 2^{13}\, (0,00000000000000000000000)_2\end{aligned}$$
- $$\begin{aligned}(45,5)_{10} &\approx (01000010001101100000000000000000)_{IEEE754} = \\ &= (-1)^0\, 2^{5}\, (1,011011)_2 = (-1)^0\, 2^{13}\, (0,00000001011011)_2\end{aligned}$$

luego:

- $(a \oplus b) \oplus c = a \oplus c = (01000110010000011001101000000000)_{IEEE754} = (12390,5)_{10} \approx (a+b)+c = (12390{,}5001)_{10}$
- $a \oplus (b \oplus c) = (12345)_{10} \oplus (01000010001101100000000000011010)_{IEEE754} = {} = (-1)^0\, 2^{13}\, (1,1000000111001)_2 + (-1)^0\, 2^{5}\, (1,01101100000000000011010)_2 \approx {} \approx (-1)^0\, 2^{13}\, (1,1000000111001)_2 + (-1)^0\, 2^{13}\, (0,00000001011011000000000)_2 = {} = (-1)^0\, 2^{13}\, (1,10000011001101)_2 = (12390,5)_{10} \approx a + (b+c) = (12390{,}5001)_{10}$

16. Repetir el ejercicio anterior pero ahora utilizando doble precisión.

 Ayuda No es necesario volver a hacer todo el procedimiento para hacer las conversiones.

 Solución Puede apreciarse que en este caso, ambas operaciones dan el resultado exacto.