1. Converta os seguintes números para o sistema de numeração em binário com 8-bits em complemento para 2 utilizando a representação indicada.
   a) X = +13,375 utilizando a representação Q5.3
   b) Y = -15,5 utilizando a representação Q6.2
   c) Realize a operação Z = X + Y com a máxima precisão, mas garantindo a inexistencia de overflow. Normalize o resultado para a representação Q4.4

2. Converta os seguintes números para a base decimal, tendo em consideração o sistema de representação de 8-bits em complemento para 2 indicado:
   a) $X_{Q5.3} = 10110111_2$
   b) $X_{Q6.2} = 00101111_2$

3. Determine o valor médio dos seguintes números, assumindo um sistema de representação de 8 bits em complemento para 2:
   - 10011011_2, em Q5.3
   - 10100101_2, em Q6.2
   - 00101001_2, em Q6.2
   - 00110001_2, em Q7.1

   a) Qual o resultado desta operação assumindo a máxima precisão possível?
   b) Qual o resultado desta operação nas seguintes representações: Q7.1, Q6.2, Q5.3? Converta o resultado da alínea anterior, para os formatos indicados. Comenta o que observa.

1) a)

13 | 2
1 6 | 2
0 3 | 2
1 1 | 2
1 0

P. Int:
1101

$0,375 \times 2 = 0,75$
$0,75 \times 2 = 1,5$
$0,5 \times 2 = 1,0$

P. Frac:
011

Logo
01101,011

b) −15,5

15 | 2
1 7 | 2
1 3 | 2
1 1 | 2
1 0

001111

$0,5 \times 2 = 1,0$
$0,0 \times 2 = 0,0$

P. Frac
10

001111,10

110000,01 → 110000,10

c) X+Y = 13,375 − 15,5

  01101,011
+ 110000,10
  11101,111 → 1101,1110

2) a) $10110,111_2 = -9,125_{10}$
↓
$01001,001_2 \to 2^3 + 2^0 + 2^{-3}$
↓
9,125

b) $001011,11_2 = 2^{-2} + 2^{-1} + 2^0 + 2^1 + 2^3 = 11,75$

3.
  1110011,011
  0001010,010
  1101001,010
+ 0011000,100
  1111111,011

÷4
↓
11111,11011 ← Mais precisão possível

b) Q7.1 = 1111111,1
Q6.2 = 111111,11
Q5.3 = 11111,110

↑
Perde-se muita precisão

4. Pretende-se calcular o perímetro de uma mesa redonda com um raio de 2,75 metros. Realize todos os cálculos considerando um sistema de 8-bits sem sinal usando a representação em virgula fixa (Qx.y) que permita obter a máxima precisão sem a ocorrência de overflows. Para sua referência, assuma a seguinte aproximação do número PI utilizando numa representação Q4.10: $\pi_{Q4.10} = 00110010010000_2$

5. Um determinado programa foi implementado para calcular a área de um quadrado num sistema de numeração binário de 8-bits sem sinal utilizando uma representação em vírgula fixa Q2.6 à entrada.

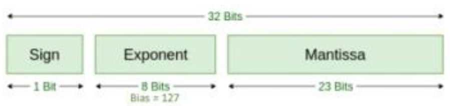
   a) Determine qual a representação do resultado que maximiza a precisão e garante a inexistência de overflow.

   b) Determine e comente os resultados do programa quando o lado do quadrado (L) toma os seguintes valores:

   i) L = 2,5 ii) L = 1,875 iii) L = 0,125

6. Represente os seguintes números utilizando o sistema de representação em vírgula flutuante baseado no formato IEEE-754 com resolução de 32-bits (float):

   a) X = 34,375
   b) X = -70,75
   c) X = 0,09375 (NOTA: $0.09375 = 2^{-4}+2^{-5}$)
   d) X = -1,0



7. Converta para a base decimal os seguintes números representados no sistema de representação em vírgula flutuante baseado no formato IEEE-754 com resolução de 32-bits (float):

   a) X = 00111110101100000000000000000000₂
   b) X = 11000001010100000000000000000000₂
   c) X = 10111111100000000000000000000000₂
   d) X = 00000000000000000000000000000000₂

4) $\pi_{Q4,10}$ = 0011,0010010000

$Q_{2,12}$ 11,001001000000

$Q_{2,6}$ 11,001001 //

2,75 → 10,110000 $Q_{2,6}$

P = 2 × 10,11000 × 11,001001 =
= 10001,010 //

1100100 1
× 1011 0000
1100100 1
1100100 1
0000000 0
+ 1 1001001
1 0001010 0 0 1 1 0000
10001,010

5 L = 3 = 11_2

11,111111 × 11,111111 = 1111,1110

Representação Q4.4

b) $2,5_{10}$ = 10,100000

10100000
× 101 00000
10100000
+10100000
01100100000 00000

0110,0100₂ → 6,25 //

ii) 1,875 = 01,111000

01111000
× 01111000
0 1 1 1 1 0000
0 1111000
0 1 111000
0 1 1 1 1 000
0 1 1 1 1000
+0
00 11 1000 1000 000

0011,1000₂ → 3,5
Perde precisão
$(1,875)^2 = 3,515625$

iii) 0,125 = 00,001000

00001000
× 00001000
0000 00 00 01 000000

0000,0000₂ → 0
Perde Precisão
$(0,125)^2 = 0,015625$

6) a) Sign = +
Mant = 34,375
Exp = 127 + 5 = 132 = 1000 0100

34 | 2
0 17 | 2
1 8 | 2
0 4 | 2
0 2 | 2
0 1 | 2
1 0

100010,011 = 1,00010011 × 2^5

Logo 0 | 1000 0100 | 00010011 0000 0000 0000 000
S Ex Mant

b) Sign = −

70 | 2
0 35 | 2
1 17

70 = 1000110 0,75 = 0,11

1000110,11 = 1,00011011 × 2^6

127+6 = 133 = 1000 0101

Logo 1 | 1000 0101 | 0001 1011 0000 0000 0000 000
S Ex Mant

c) 0,09375
0,00011 = $1,1 \times 2^{-4}$

Sign +

Exp = 127 − 4 = 123 = 0111 1011

0 | 0111 1011 | 1000 0000 0000 0000 0000 000
S Ex Mant

d) Sign: −

$1,0_{10} = 1,0_2$

1 | 0111 1111 | 0000 0000 0000 0000 0000 000
S Ex Mant

7 a) 0 01111101 0110 0000 0000 0000 0000 000
+ 125 1,011

125 − 127 = −2

$1,011 \times 2^{-2}$ = 0,01011 → 0,34375

b) 1 10000010 1010 000
− 130 1,1010

130 − 127 = 3

$1,1010 \times 2^3$ = 1101 → −13

c) 1 0111 1111 0000 000
− 127 1,0

$-1 \; 1,0 \times 2^0 = 1$

d) $+ 1,0 \times 2^{-127} \approx 0$

8. Considere os seguintes números representados no sistema de representação em vírgula flutuante baseado no formato IEEE-754 com resolução de 32-bits (float): A = 3D900000h, B = BEB00000h. Calcule, neste sistema de representação, os seguintes valores:

a) A + B b) A – B c) A x B

9. De acordo com a expressão definida por Albert Einstein para a relação entre massa e energia, um determinado corpo com massa m possui uma energia associada dada por $E=mc^2$, em que c é a velocidade da luz. Pretende-se utilizar esta relação para calcular a energia associada a um grão de arroz, com a massa de 25mg. Discuta as vantagens/desvantagens da utilização de vírgula-fixa e vírgula-flutuante (com resolução de 32-bits) para a implementação desta operação.

NOTAS: c = 299792458 m/s = 11DE784Ah em Q32.0
m = 0.000025 Kg = 0001A36Eh em Q0.32

8) A = 3D900000h

$\Downarrow$

$\underbrace{0011\ 1101\ 1}001\ 0 \cdots 0$

B = BEB00000h

$1\ \underbrace{01111101}\ 011\ 0 \ldots 0$

$A = 1.001 \times 2^{-4} \rightarrow 0{,}01001 \times 2^{-2}$

$B = -1.011 \times 2^{-2}$

$$
\begin{array}{r}
0{,}01001 \times 2^{-2} \\
+\ 0{,}10100 \times 2^{-2} \\
\hline
0{,}11101 \\
1{,}00010
\end{array}
$$

$A+B = -1{,}00011 \times 2^{-2}$

1 01111101 00011 0...

b) A-B

$$
\begin{array}{r}
0{,}01001 \times 2^{-2} \\
+\ 1{,}01100 \times 2^{-2} \\
\hline
1{,}10101 \times 2^{-2}
\end{array}
$$

0 01111101 101010...0

c) A x B =

(+ x -)

$\Downarrow$

–

$$
\begin{array}{r}
1{,}001 \\
\times\ 1{,}011 \\
\hline
1001 \\
1001\ \\
+\ 1001\ \ \ \\
\hline
01{,}100011
\end{array}
$$

$2^{-4} \times 2^{-2} = 2^{-6}$

127-6 = 121

$-1{,}100011 \times 2^{-6}$

1 01111001 100011 0...0

9) Fixa Flutuante

– Precisão

↑ vai levar a overflow!