Exercise 2: The following table shows decimal numbers

| a | -1609.5 |
|---|-----------|
| b | -938.8125 |

1. Write down the binary representation of the decimal number, assuming the IEEE 754 **single** precision format.

| 31 | 30 | 29 | 28 | 27 | 26 | 25 | 24 | 23 | 22 | 21 | 20 | 19 | 18 | 17 | 16 | 15 | 14 | 13 | 12 | 11 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
|-------|----|----|----|------|------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|-------|----|---|---|---|---|---|---|---|---|---|---|
| s | | | (| expo | nen | t | | | | | | | | | | | | | fr | actio | n | | | | | | | | | | |
| 1 bit | | | | 8 t | oits | | | | | | | | | | | | | | 2 | 3 bit | s | | | | | | | | | | |

2. Write down the binary representation of the decimal number, assuming the IEEE 754 **double** precision format.

| 31 | 30 | 29 | 28 | 27 | 26 | 25 | 24 | 23 | 22 | 21 | 20 | 19 | 18 | 17 | 16 | 15 | 14 | 13 | 12 | 11 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
|-------|----|----|----|----|----|-------|-----|----|----|----|----|----|----|-------|------|-------|------|----|----|----|------|------|---|---|---|---|---|---|---|---|---|
| s | | | | | ex | pone | ent | | | | | | | | | | | | | | frac | tion | | | | | | | | | |
| 1 bit | : | | | | 1 | 1 bit | ts | | | | | | | | | | | | | | 20 | bits | | | | | | | | | |
| | | | | | | | | | | | | | fr | actio | n (c | ontir | nued | i) | | | | | | | | | | | | | |

32 bits

Answer:

a)

```
-1609.5 = -11001001001.1_2
        = -1.110010010011 \times 2^{10}
```

1. Biểu diễn theo độ chính xác đơn

```
sign = -1
exponent = 10 + 127 = 137 = 10001001_2
fraction = 110010010011
```

| 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
|---|---|-----|----|----|------|------|------|---|---|---|---|---|---|---|---|---|----|-----|-----|------|------|-----|----|---|---|---|---|---|---|---|---|
| S | | exp | on | en | t (8 | B bi | its) | | | | | | | | | | fr | act | ior | 1 (2 | 23 1 | bit | s) | | | | | | | | |

2. Biểu diễn theo độ chính xác kép

```
sign = -1
exponent = 10 + 1023 = 1033 = 10000001001_2
fraction = 110010010011
```

| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
|---|---|---|----|----|-----|------|----|----|-----|---|---|---|---|---|----|---|------|-----|------|-----|-----|------|-----|------|------|-----|----|----|---|-------|
| S | | | Ex | po | nei | nt (| 11 | bi | ts) | | | | | | | | | | Fra | act | ion | (5 | 2 t | oits | | | | | | |
| | | | | | | | | | | | | | | | Vć | | 12 t | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | 4 | 40 | bit | s tl | hấp | c | on l | lại | = (|) to | oàn | bá | (6 | | |

```
b)
    -938.8125 = -1110101010.1101 \times 2^{0}
                = -1.1101010101101 \times 2^9
```

1. Biểu diễn theo đô chính xác đơn

```
exponent = 9 + 127 = 136 = 10001000_2
fraction = 1101010101101_2
```

| 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
|---|---|-----|----|-----|----|------|------|---|---|---|---|---|---|---|---|---|----|-----|-----|------|----|------|----|---|---|---|---|---|---|---|---|
| S | | exp | on | ent | 3) | 3 bi | its) | | | | | | | | | | fr | act | ior | n (2 | 23 | bits | s) | | | | | | | | |

2. Biểu diễn theo độ chính xác kép

```
sign = -1
exponent = 9 + 1023 = 1032 = 10000001000_2
fraction = 1101010101101_2
```

| 1 | 1 0 0 0 0 0 0 1 0 0 0 | 1 1 0 1 0 1 0 1 0 1 0 1 0 1 0 0 0 0 0 0 |
|---|-----------------------|---|
| S | Exponent (11 bits) | Fraction (52 bits) |
| | | Với 13 bits cao = 1101010101101 |
| | | 39 bits thấp còn lại = 0 toàn bộ) |

Exercise 4: The following table shows decimal numbers

| a | 5.00736125 x 10 ⁵ |
|---|------------------------------------|
| b | -2.691650390625 x 10 ⁻² |

- 1. Write down the binary representation of the decimal number, assuming the IEEE 754 **single** precision format.
- 2. Write down the binary representation of the decimal number, assuming the IEEE 754 **double** precision format.

Answer:

```
= 1.111010010000000000001 \times 2^{18}
```

Trình bày: Trần Thị Như Nguyệt

1. Biểu diễn theo độ chính xác đơn

```
sign = 0
exponent = 18 + 127 = 145 = 10010001_2
fraction = 111010010000000000001_2
```

| 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
|---|---|-----|----|-----|------|------|------|---|---|---|---|---|---|---|---|---|----|-----|-----|------|----|------|----|---|---|---|---|---|---|---|---|
| S | | exp | on | ent | t (8 | B bi | its) | | | | | | | | | | fr | act | ior | n (2 | 23 | bits | s) | | | | | | | | |

3. Biểu diễn theo độ chính xác kép

```
sign = -1
exponent = 18 + 1023 = 1032 = 10000010001_2
fraction = 111010010000000000001<sub>2</sub>
```

| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1110100100000000000010000000 |
|---|---|---|---|-----|-----|------|----|------|---|---|---|--|
| S | | | | Exp | one | nt (| 11 | bits |) | | | Fraction (52 bits) |
| | | | | | | | | | | | | với: 21 bits cao =111010010000000000001 |
| | | | | | | | | | | | | 31 bits thấp còn lại bằng 0 |

```
b)
-2.691650390625 \times 10^{-2} = 0.02691650390625 = 441/2^{14} = 1101111001 \times 2^{-14}
                                                                       = 1.101111001 \times 2^{-6}
```

1. Biểu diễn theo đô chính xác đơn

```
sign = -1
exponent = -6 + 127 = 121 = 1111001_2
fraction = 10111001_2
```

| 1 | L | 0 | 1 1 | . 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
|---|---|---|-----|-----|-------|-----|------|---|---|---|---|---|---|---|---|---|----|-----|-----|------|------|------|----|---|---|---|---|---|---|---|---|
| S | 3 | e | xpo | ner | nt (8 | 3 b | its) | | | | | | | | | | fr | act | ior | ı (2 | 23 1 | bits | s) | | | | | | | | |

2. Biểu diễn theo độ chính xác kép

```
sign = -1
exponent = -6 + 1023 = 1014 = 11111111001_2
fraction = 10111001_2
```

| 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
|---|---|---|----|----|-----|------|-------------------|-----|-----|---|---|---|---|---|---|---|---|---|-----|-----|----|-----|-----|------|---|---|---|---|---|-------|
| S | | | Ex | po | nei | nt (| $\overline{(11)}$ | bit | ts) | | | | | | | | | I | Fra | cti | on | (52 | 2 b | its) |) | | | | | |

Trình bày: Trần Thị Như Nguyệt

| 8 bits cao =10111001 |
|-----------------------------|
| 44 bits thấp còn lại bằng 0 |

Exercise 5: The following table shows bit patterns expressed in hexademical notation.

| a | 0x24A60004 |
|---|------------|
| b | 0xAFBF0000 |

- 1. What decimal number does the bit pattern represent if it is a two's complement integer? An unsigned integer?
- 2. If this bit pattern is placed into the Instruction Register, what MIPS instruction will be executed?
- 3. What decimal number does the bit pattern represent if it is a floating point number? Use the IEEE 754 standard.

Answers:

Đề bài cho một chuỗi số 8 chữ số đang biểu diễn trong hệ 16 (tương ứng với chuỗi 32 bits trong hệ 2)

- 1. Nếu 32 bits này là số nguyên có dấu bù 2, thì đó là số nào? Nếu là không dấu, thì đó là số nào?
- 2. Nếu chuỗi 32 bits này là lệnh MIPS thì đây là lệnh gì?
- 3. Nếu đây là số dấu chấm động (floating-point) thì là bao nhiều?

Đáp số:

1. a)

 $0x24A60004 = 0010\ 0100\ 1010\ 0110\ 0000\ 0000\ 0000\ 0100$

Nếu đây là số có dấu dạng bù 2, số tương ứng = 614858756 Nếu đây là số có không dấu, số tương ứng = 614858756

b)

 $0xAFBF0000 = 1010\ 1111\ 1011\ 1111\ 0000\ 0000\ 0000\ 0000$

Nếu đây là số có dấu dạng bù 2, số tương ứng = -1346437120 Nếu đây là số có không dấu, số tương ứng = 2948530176

2. a)

 $0x24A60004 = 0010\ 0100\ 1010\ 0110\ 0000\ 0000\ 0000\ 0100$

Nếu đây là lệnh assemply, lệnh tương ứng: *addiu* \$6, \$5, 4 Hay *addiu* \$a2, \$a1, 4

b) **0**x**AFBF0000** = **1010 1111 1011 1111 0000 0000 0000 0000**

Nếu đây là lệnh assemply, lệnh tương ứng: sw \$31, 0(\$29)Hay sw \$ra, 0(\$sp)

3. a)

$0x24A60004 = 0010\ 0100\ 1010\ 0110\ 0000\ 0000\ 0000\ 0100$

Nếu đây là số floating-point với độ chính xác đơn

0<mark>01001001</mark>0100110 0000 0000 0000 0100

Bit thứ 31 = 0, là bit dấu của số floating-point → số floating-point này là số dương

Bit thứ 30 tới $23 = 01001001 = 73_{(10)}$, là phần mũ của số floating-point sau khi đã công thêm $127 \rightarrow$ số floating-point này có số mũ = 73 - 127 = -54

Bit thứ 22 tới 0 = 0100110~0000~0000~0100, là phần thập phân của floating-point

 \Rightarrow số floating-point = 1.0100110 0000 0000 0000 0100 x 2⁻⁵⁴

b)

Nếu đây là số floating-point với độ chính xác đơn

 \Rightarrow số floating-point = -1.011 1111 x 2⁻³²

cuu duong than cong

Sửa lại thành -1.278 x 10³

Exercise 6:

The following table shows pairs of decimal numbers

| | A | В |
|----|-----------------------------|--------------------------------|
| a. | -1278 x 10 ³ | -3.90625 x 10 ⁻¹ |
| b. | 2.3109375 x 10 ¹ | 6.391601562 x 10 ⁻¹ |

- 1. Calculate the sum of A and B by hand, assuming that we keep 11 bits of significand and 5 bits of the exponent. (Rounding rule: add 1 if the bits to the right of the desired point is larger or equal to $100_{(2)}$). Show all the steps.
- 2. Calculate the sum of A and B by hand, assuming A and B are stored in the IEEE-754 single precision format. Show all the steps.

Answer:

1. Đề bài yêu cầu tính tổng A và B bằng tay (tức chạy từng bước) với giả sử số floating-point chỉ cho phép dùng 11 bits cho phần significand và 5 bits cho phần exponent

a)

A = -1.278 x
$$10^3$$
 = -1278 = -100111111110
= -1.00111111110 x 2^{10}

(kiểm tra số floating-point A đã đúng chuẩn chưa: 10 nằm trong phạm vi số 5 bits của phần mũ

Trình bày: Trần Thị Như Nguyệt

```
và phần significand '1.0011111110' đúng 11 bits cho phép
              → Đúng chuẩn)
          B = -3.90625 \times 10^{-1} = -0.390625 = -25/2^{6} = -11001 \times 2^{-6}
                                                 = -1.1001 \times 2^{-2}
          (kiểm tra số floating-point B này đã đúng chuẩn chưa:
              -2 nằm trong pham vi số 5 bits của phần mũ
              và phần significand '1.1001' cũng không vượt quá 11 bits
              → Đúng chuẩn)
          A + B = -(1.00111111110 \times 2^{10} + 1.1001 \times 2^{-2})
                = - (1.00111111110 \times 2^{10} + 0.000000000011001 \times 2^{10})
               = -1.00111111110011001 \times 2^{10}
                Do phần significand chỉ được phép chứa 11 bits, nên A + B phải được
          làm tròn, phần cắt bỏ là 011001_{(2)} > 100_{(2)} nên 1.00111111110011001 \approx
          1.0011111111
          V_{ay} A + B = -1.00111111111 \times 2^{10}
                     = -100111111111_{(2)} = 1279
          A = 2.3109375 \times 10^{1} = 23.109375
              =23+7/2^6
                               = 10111.000111_{(2)}
                               = 1.0111000111 \times 2^4
          (kiểm tra số floating-point A đã đúng chuẩn chưa:
              4 nằm trong phạm vi số 5 bits của phần mũ
              và phần significand '1.0111000111' cũng không vượt quá 11 bits
              → Đúng chuẩn)
          B = 6.391601562 \times 10^{-1} = 0.6391601562
                                 = 1309/2^{11} = 10100011101 \times 2^{-11}
                                             = 1.0100011101 \times 2^{-1}
          (kiểm tra số floating-point B này đã đúng chuẩn chưa:
              -1 nằm trong phạm vi số 5 bits của phần mũ
              và phần significand '1.0100011101' cũng không vượt quá 11 bits
              → Đúng chuẩn)
          A + B = 1.0111000111 \times 2^4 + 1.0100011101 \times 2^{-1}
                = 1.0111000111 \times 2^4 + 0.000010100011101 \times 2^4
               Do phần significand chỉ được phép chứa 11 bits, nên A + B phải được làm tròn,
Vay A + B = 1.01111110000 \times 2^4
                     = 23.75_{(10)}
```

Đề bài yêu cầu tính tổng A và B bằng tay (tức chạy từng bước) với giả sử số 2. floating-point dùng format IEEE đô chính xác đơn

```
a)
   A = -1.278 \times 10^3 = -1278 = -100111111110
                            = -1.00111111110 \times 2^{10}
(kiểm tra số floating-point A đã đúng chuẩn chưa:
       (10 + 127) nằm trong phạm vi số 8 bits của phần mũ
       và phần fraction '0011111110' cũng không vượt quá 23 bits
       → Đúng chuẩn)
   B = -3.90625 \times 10^{-1} = -0.390625 = -25/2^{6} = -11001 \times 2^{-6}
                                             = -1.1001 \times 2^{-2}
   (kiểm tra số floating-point B này đã đúng chuẩn chưa:
       (-2 + 127) nằm trong phạm vi số 8 bits của phần mũ
       và phần fraction '1001' cũng không vượt quá 23 bits
       → Đúng chuẩn)
   A + B = -(1.001111111110 \times 2^{10} + 1.1001 \times 2^{-2})
          = -(1.00111111110 \times 2^{10} + 0.000000000011001 \times 2^{10})
         = -1.001111111110011001 \times 2^{10}
          Phần fraction này chứa 16 bits, không vượt quá 23 bits của IEEE độ
   chính xác đơn, nên:
    Vav A + B = -1.00111111110011001 \times 2^{10}
b)
   A = 2.3109375 \times 10^{1} = 23.109375
                         = 23 + 7/2^6
                          = 10111.000111_{(2)}
                          = 1.0111000111 \times 2^{4}
   (kiểm tra số floating-point A đã đúng chuẩn chưa:
       (4 + 127) nằm trong phạm vi số 8 bits của phần mũ
       và phần fraction '0111000111' cũng không vượt quá 23 bits
       → Đúng chuẩn)
   B = 6.391601562 \times 10^{-1} = 0.6391601562
        = 1309/2^{11} = 10100011101 \times 2^{-11}
                                         = 1.0100011101 \times 2^{-1}
   (kiểm tra số floating-point B này đã đúng chuẩn chưa:
       (-1+127) nằm trong phạm vi số 8 bits của phần mũ
       và phần fraction '0100011101' cũng không vượt quá 23 bits
       → Đúng chuẩn)
   A + B = 1.0111000111 \times 2^4 + 1.0100011101 \times 2^{-1}
          = 1.0111000111 \times 2^4 + 0.000010100011101 \times 2^4
```

Phần fraction này chứa 15 bits, chưa vượt quá 23 bits nên

Exercise 7:

The following table shows pairs of decimal numbers

| A | | В |
|----|------------------------------|----------------------------|
| a. | 5.66015625 x 10 ⁰ | 8.59375 x 10 ⁰ |
| b. | 6.18 x 10 ² | 5.796875 x 10 ¹ |

- 1. Calculate A x B by hand, assuming that we keep 11 bits of significand and 5 bits of the exponent. (Rounding rule: add 1 if the bits to the right of the desired point is larger or equal to $100_{(2)}$). Show all the steps.
- 2. Calculate A x B by hand, assuming A and B are stored in the IEEE-754 single precision format. Show all the steps.

Answer:

Đề bài yêu cầu tính A x B bằng tay (tức chạy từng bước) với giả sử số floating-1. point chỉ cho phép dùng 11 bits cho phần significand và 5 bits cho phần exponent a)

> $A = 5.66015625 \times 10^0 = 1.0110101001 \times 2^2$ (kiểm tra số floating-point A đã đúng chuẩn cho phép chưa: → Đúng chuẩn)

 $B = 8.59375 \times 10^{0} = 1.0001001100 \times 2^{3}$

(kiểm tra số floating-point B này đã đúng chuẩn cho phép chưa: → Đúng chuẩn)

 $A \times B = (1.0110101001 \times 2^2) \times (1.0001001100 \times 2^3)$

Exponent của A x B = 2 + 3 = 5Significand

```
1.0110101001
         \times 1.0001001100
             00000000000
            00000000000
           10110101001
         10110101001
        00000000000
       00000000000
      10110101001
    0000000000
   00000000000
  00000000000
 10110101001
1.10000101001000101100
```

Do phần significand chỉ được phép chứa 11 bits, nên significand của A x B phải được làm tròn, phần cắt bỏ là $1000101100_{(2)} > 100_{(2)}$ nên $1.1000010100 \frac{1000101100}{(2)} \approx 1.1000010101$ $V_{ay} A \times B = 1.1000010101 \times 2^5$

| | A | В |
|----|------------------------------|----------------------------|
| a. | 5.66015625 x 10 ⁰ | 8.59375 x 10 ⁰ |
| b. | 6.18 x 10 ² | 5.796875 x 10 ¹ |

$$A = 6.18 \times 10^{2} = 618 = 1001101010_{(2)}$$
$$= 1.001101010 \times 2^{9}$$

(kiểm tra số floating-point A đã đúng chuẩn cho phep chưa:

→ Đúng chuẩn)

$$B = 5.796875 \times 10^{1} = 57.96875 = 1.11001111111 \times 2^{5}$$

(kiểm tra số floating-point B này đã đúng chuẩn cho phép chưa:

→ Đúng chuẩn)

A x B =
$$(1.001101010 \times 2^9) \times (1.11001111111 \times 2^5)$$

Exponent của A x B = $9 + 5 = 14$
Significand

```
1.0011010100
               × 1.1100111111
                   10011010100
                 10011010100
                10011010100
               10011010100
             10011010100
            10011010100
           00000000000
         00000000000
        10011010100
       10011010100
     10011010100
    10001011111110000101100
Significand của A x B = 10.001011111110000101100 \rightarrow \text{phải chuẩn hóa lại}
A \times B = 10.001011111110000101100 \times 2^{14}
     = 1.00010111111110000101100 \times 2^{15}
Do phần significand chỉ được phép chứa 11 bits, nên significand của A x B phải
được làm tròn, phần cắt bỏ là
                                          10000101100_{(2)} > 100_{(2)} nên
1.00010111111110000101100_{(2)} \approx 1.0001100000
```

2. Đề bài yêu cầu tính A x B bằng tay (tức chạy từng bước) với giả sử số floatingpoint dùng format IEEE độ chính xác đơn

```
A = 5.66015625 x 10<sup>0</sup> = 1.01101010101 x 2<sup>2</sup>
(kiểm tra số floating-point A đã đúng chuẩn cho phép chưa:

→ Đúng chuẩn)

B = 8.59375 x 10<sup>0</sup> = 1.0001001100 x 2<sup>3</sup>
(kiểm tra số floating-point B này đã đúng chuẩn cho phép chưa:

→ Đúng chuẩn)

A x B = (1.0110101001 x 2<sup>2</sup>) x (1.0001001100 x 2<sup>3</sup>)

Exponent của A x B = 2 + 3 = 5
Significand
```

Vây A x B = $1.0001100000 \times 2^{15}$

Do phần fraction được phép chứa 23 bits, nên fraction của A x B từ kết quả trên thỏa mãn, không cần làm tròn

 $V_{ay} A \times B = 1.10000101001000101100 \times 2^5$

b)

| | A | В |
|----|------------------------------|----------------------------|
| a. | 5.66015625 x 10 ⁰ | 8.59375 x 10 ⁰ |
| b. | 6.18 x 10 ² | 5.796875 x 10 ¹ |

```
A = 6.18 \times 10^2 = 618 = 1001101010_{(2)}
= 1.001101010 x 2<sup>9</sup>
```

(kiểm tra số floating-point A đã đúng chuẩn cho phep chưa:

→ Đúng chuẩn)

$$B = 5.796875 \times 10^{1} = 57.96875 = 1.11001111111 \times 2^{5}$$

(kiểm tra số floating-point B này đã đúng chuẩn cho phép chưa:

→ Đúng chuẩn)

A x B =
$$(1.001101010 \times 2^9) \times (1.11001111111 \times 2^5)$$

Exponent của A x B = $9 + 5 = 14$
Significand

```
1.0011010100
              × 1.1100111111
                  10011010100
                 10011010100
                10011010100
              10011010100
             10011010100
            10011010100
           00000000000
         00000000000
        10011010100
       10011010100
     10011010100
    10001011111110000101100
Significand của A x B = 10.001011111110000101100 → phải chuẩn hóa lại
A x B = 10.0010111111110000101100 \times 2^{14}
     = 1.00010111111110000101100 \times 2^{15}
Do phần fraction được phép chứa 23 bits, nên fraction của A x B trên hop lệ
Vậy A x B = 1.0001011111110000101100 \times 2^{15}
```