Девяткин Арсений P3115

**Домашнее задание №6.**

**Сложение чисел с плавающей точкой.**

**Вариант 17.**

A = 256,8; B = 244,6

1. Формат *Ф1* (число разрядов мантиссы *m* =12).

а) A>0 B>0

*А* = (256,8)10 =(100,(С))16 = (0,1(С)16 · 163

Характеристика = 64 + 3 = 67

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

**1** округление к ближайшему

*В* = (244,6)10 = (F4,(9))16 = (0,F4(9))16 · 162

Характеристика = 64 + 2 = 66

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |

1) *XA* = \_1 0 0 0 0 1 1

*XB* = 1 0 0 0 0 1 0

(*XA-XB*)пр .= 0 0 0 0 0 0 1

(*XA-XB*). = 1; *ХС = ХA =* 3

а) Оба операнда положительные (*А*>0, *B*>0):

4

2,3) М*А* = . 0 0 0 1 0 0 0 0 0 0 0 1

+

М*В* = . 0 0 0 0 1 1 1 1 0 1 0 0

М*С* = 0. 0 0 0 1 1 1 1 1 0 1 0 1

Результат сложения нормализован.

(*ХС* = *ХА* = 3).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 |

*С*

0 1  7 8 19

*С*\* = М*С* · 16Р*с* = (0,1F5)16 · 163 = (1F5)16 = 50110.

(в то время как 256,8 + 244,6 = 501,4)

Несмотря не то, что первый операнд за счет округления был представлен с избытком, результат получился представленным с недостатком. Этот факт можно объяснить потерей значащих младших разрядов у второго операнда при выравнивании порядков.

Итого абсолютная погрешность Δ*С* = *С*Т - *С*\* = 501,4 – 501 = 0,4

Относительная погрешность δ*С* = · 100% =. 100% = 0,08%,

Данную погрешность можно объяснить:

1. Неточным представлением операндов;
2. Потерей значащих разрядов мантиссы одного из операндов при уравнивании порядков;

б) A<0 B>0

Сложение мантисс будем проводить их прямым вычитанием. В качестве уменьшаемого используем мантиссу положительного операнда (*В*);



4

2,3) М*В* = . 0 0 0 0 1 1 1 1 0 1 0 0

М*А* = . 0 0 0 1 0 0 0 0 0 0 0 1

М*Сдк* = . 1 1 1 1 1 1 1 1 0 0 1 1

М*Спк* = . 0 0 0 0 0 0 0 0 1 1 0 1

Результат денормализован вправо.

 8

4) МС = . 1 1 0 1 0 0 0 0 0 0 0 0

Т.к. выполнен сдвиг мантиссы влево, то характеристику результата нужно уменьшить на 2 (*ХС* = *ХС* -2 = 1).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

*С*

*С*\* = М*С* · 16Р*с* = -(0,D)16 · 161 = -(D)16 = -1310.

(в то время как (-256,8) + 244,6 = -12,2)

Первый операнд (отрецательный) за счет округления был представлен с избытком, из - за этого результат получился представленным с избытком (по модулю). Также результат можно объяснить потерей значащих младших разрядов у второго операнда при выравнивании порядков.

Итого абсолютная погрешность Δ*С* = *С*Т - *С*\* = -12,2 – (-13) = 0,8

Относительная погрешность δ*С* = · 100% =. 100% = 6,55%,

Данную погрешность можно объяснить:

1. Неточным представлением операндов;
2. Потерей значащих разрядов мантиссы одного из операндов при уравнивании порядков;

в) A>0 B<0

Сложение мантисс будем проводить их прямым вычитанием. В качестве уменьшаемого используем мантиссу положительного операнда (*A*);



4

2,3) М*A* = . 0 0 0 1 0 0 0 0 0 0 0 1

М*B* = . 0 0 0 0 1 1 1 1 0 1 0 0

М*С* = . 0 0 0 0 0 0 0 0 1 1 0 1

Результат денормализован вправо.

 8

4) МС = . 1 1 0 1 0 0 0 0 0 0 0 0

Т.к. выполнен сдвиг мантиссы влево, то характеристику результата нужно уменьшить на 2 (*ХС* = *ХС* -2 = 1).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

*С*

*С*\* = М*С* · 16Р*с* = (0,D)16 · 161 = (D)16 = 1310.

(в то время как 256,8 - 244,6 = 12,2)

Первый операнд (положительный) за счет округления был представлен с избытком, из - за этого результат получился представленным с избытком. Также результат можно объяснить потерей значащих младших разрядов у второго операнда при выравнивании порядков.

Итого абсолютная погрешность Δ*С* = *С*Т - *С*\* = 12,2 – 13 = - 0,8

Относительная погрешность δ*С* = · 100% =. 100% = 6,55%,

Данную погрешность можно объяснить:

1. Неточным представлением операндов;
2. Потерей значащих разрядов мантиссы одного из операндов при уравнивании порядков;
3. Формат *Ф2*.

а) A>0, B>0

*А* = (256,8)10 =(100,(С))16 = (0.10000000011001100110)2 · 29

Характеристика = 128 + 9 = 137

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |

*В* = (244,6)10 = (F4,(9))16 = (0.1111010010011001100)2 · 28

Характеристика = 128 + 8 = 136

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 |

**1** округление к ближайшему

1) *XA* = \_1 0 0 0 1 0 0 1

*XB* = 1 0 0 0 1 0 0 0

(*XA-XB*)пр .= 0 0 0 0 0 0 0 1

(*XA-XB*). = 1; *ХС = ХА* = 9*.*

а) Оба операнда положительные (*А*>0, *B*>0):



1

2,3) М*А* = . 1 0 0 0 0 0 0 0 0 1 1 0

+

М*В* = . 0 1 1 1 1 0 1 0 0 1 0 0

М*С* = . 1 1 1 1 1 0 1 0 1 0 1 0

Результат сложения нормализован.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |

*С*

*С*\* = М*С* · 2Р*с* = (0,111110101010)2 · 29 = (111110101,010)2 = 501,25

(в то время как 256,8 + 244,6 = 501,4)

Δ*С* = *С*Т - *С*\* = 501,4 – 501,25 = 0,15

δ*С* = · 100% = . 100% = 0,03%

б) *А*<0, *B*>0.

2,3) М*B* = . 0 1 1 1 1 0 1 0 0 1 0 0

-

М*A* = . 1 0 0 0 0 0 0 0 0 1 1 0

М*Сдк* = . 1 1 1 1 1 0 0 1 1 1 1 0

М*Спк* = . 0 0 0 0 0 1 1 0 0 0 1 0

Результат денормализован вправо.

 5

4) МС = . 1 1 0 0 0 1 0 0 0 0 0 0

Т.к. выполнен сдвиг мантиссы влево, то характеристику результата нужно уменьшить на 5 (*ХС* = *ХС* -5 = 4).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

*C*

*С*\* = М*С* · 2Р*с* = -(0,110001)2 · 24 = -(1100,01)2 = -12,25.

(в то время как (-256,8) + 244,6 = -12,2)

Δ*С* = *С*Т - *С*\* = -12,2 – (-12,25) = 0,05

δ*С* = · 100% = . 100% = 0,4%

в) *А*>0, *B*<0.

2,3) М*А* = . 1 0 0 0 0 0 0 0 0 1 1 0

-

М*В* = . 0 1 1 1 1 0 1 0 0 1 0 0

М*С* = . 0 0 0 0 0 1 1 0 0 0 1 0

Результат денормализован вправо.

 5

4) МС =. 1 1 0 0 0 1 0 0 0 0 0 0

Т. к. выполнен сдвиг мантиссы влево, то характеристику результата нужно уменьшить на 5 (*ХС* = *ХС* -5 = 4).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

*C*

*С*\* = М*С* · 2Р*с* = (0,110001)2 · 24 = (1100,01)2 = 12,25.

(в то время как 256,8 - 244,6 = 12,2)

Δ*С* = *С*Т - *С*\* = 12,2 – 12,25 = -0,05

δ*С* = · 100% = . 100% = 0,4%

Как мы видим, формате *Ф2* результаты получились точнее (погрешность 0,03% против 0,08% у Ф1 и 0,4 % против 6,55%). Это связанно с тем, что операнды представлены точнее и при нормализации результата сдвиг производился на один двоичный разряд, а не на четыре.