Занятие 4

# Представление чисел в формате с плавающей запятой

Число состоит из мантиссы (m) и порядка (p).

*A* = *m qp****,***

где ***m*** — мантисса числа, ***р*** — порядок, ***q*** — основание.

В машинном представлении формат числа с плавающей запятой (рис. 3.30) задается двумя полями — полем мантиссы ***m*** и полем порядка ***р***, причем каждое поле имеет свой разряд знака. Значение основания системы счисления ***q*** в формате числа не указывается — оно подразумевается одинаковым для всех чисел.

Мантисса и порядок представляются в формате с фиксированной запятой, причем обычно порядок— целое число со знаком (запятая фиксирована по­сле младшего разряда), а мантисса — правильная дробь (запятая фиксирована между знаковым разрядом и старшим разрядом модуля). С целью увеличения точности представления числа в заданном формате мантиссу представляют в *нормализованной* форме, когда старший разряд модуля мантиссы — не ноль (для прямых кодов):

,

Рассмотрим два числа: 0,2364 ⋅104 ≈ 0.0024 ⋅ 106.

Оба эти числа содержат одинаковое число разрядов, однако во втором случае мантисса не нормализована и точность представ­ления числа — всего два десятичных разряда вместо четырех разрядов у первого числа.

Для двоичной системы счисления условие нормализации принимает вид:

Пример формата изображения числа с плавающей запятой (п.з.)



Рис. 1. Представление числа в формате с плавающей запятой

Пример: Записать числа в форме с плавающей запятой в нормальной форме.

*А*=-10110,1111; *В*=+0,000110010111

Мантисса должна удовлетворять условию , следовательно

; 

Формат представления числа с п.з. определяет способ кодирования полей двоичного представления с п.з. Например, мантисса может быть представлена в прямом, обратном и дополнительном коде, порядок – в прямом, обратном и дополнительном или со смещением. От этих параметров зависит формат представления числа с п.з в памяти компьютера.

В примере ниже знак числа кодируется 2-мя двоичными разрядами, мантисса – 9-ю двоичными разрядами, знак порядка – 1 разряд, порядок – 4 разряда. И порядок и мантисса представлены в прямом коде:

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

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

Знак m Мантисса m Знак p Порядок p

Примеры: Записать числа в форме с плавающей запятой в нормальной форме.

A= -11000,1101 

B= -0,00001001 

C= +101,101001 

D= +0,10101 

**Обобщение:** число уменьшается, степень увеличивается,

число увеличивается, степень уменьшается

## Особенности сложения чисел с плавающей запятой

**Нормализация**

Нарушение нормализации – не выполнение условия . Нарушение нормализации может быть справа и слева.

Нормализованная мантисса положительного числа имеет вид: 0**.**x1x2…xn, причем x1=1, нормализованная мантисса отрицательного числа в дополнительном или обратном коде –

1. x1x2…xn, где x1=0.

Нарушение условия нормализации: 0.0x2…xn или 1.1x2…xn

Операция нормализации состоит из совокупности сдвигов и проверки наличия признаков нарушения нормализации.

### Правила выполнения сдвигов

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Исх. число** | **Сдвиг на 1 влево**  **←** | **Сдвиг на 1 вправо**  **→** |
| +Х | 0**.**x1x2…xn | x1**.**x2…xn0 | 0**.**0x1x2…xn-1 |
| (-Х) доп.код | 1**.**x1x2…xn | x1**.**x2…xn0 | 1**.**1x1x2…xn-1 |
| (-Х) обр.код | 1**.**x1x2…xn | x1**.**x2…xn1 | 1**.**1x1x2…xn-1 |

# Нарушение нормализации

**Влево** если мантисса модуля числа равна или больше 1 **Коррекция:** сдвиг вправо

**Вправо**, если мантисса числа меньше q-1. Причем:

* Отрицательное число в обр. или доп. коде начинается с 1 после точки – 11.**1…**
* Положительное число начинается с 0 после точки – 00.**0**…

## Сложение чисел с плавающей точкой (запятой)

1. Запись мантисс в прямом коде и заданной разрядной сетке
2. Запись мантисс в требуемом коде (обратном или дополнительном)
3. Уравнять порядки чисел – порядок меньшего числа привести к порядку большего, сдвигая мантиссу меньшего числа
4. Сложить мантиссы по обычным правилам
5. Проверить полученную сумму на нарушение нормализации

Операнды для сложения всегда должны быть представлены в нормализованном виде.

Для выравнивания порядков необходимо порядок меньшего числа увеличить на , т.е. сдвинуть мантиссу меньшего числа вправо на количество разрядов, равное ∆p. При выравнивании порядков может случиться нарушение нормализации, в этом случае коррекция не проводится.

Примеры: Выполнить сложение чисел, представленных в формате с плавающей запятой

1. ;  в дополнительном коде, m=6 разрядов, p=4 разряда.

Запись мантисс в прямом коде  

Проверка и выравнивание порядков  ; ∆p=0, 

Запись мантисс в дополнительном коде  

Сложение мантисс и проверка на нарушение нормализации результата



Ответ: 

2. ;  m=8 разрядов, p=4 разряда.

Запись мантисс в прямом коде  

Проверка и выравнивание порядков  ; ∆p=0, 

Сложение мантисс и проверка на нарушение нормализации результата



Ответ: 

3. ;  m=8 разрядов, p=4 разряда.

Запись мантисс в прямом коде.  

Проверка и выравнивание порядков  ; ∆p=0, 

Сложение мантисс и проверка на нарушение нормализации результата



Ответ: 

4. ;  в обратном коде, m=8 разрядов, p=4 разряда.

Запись мантисс в прямом коде.  

Запись мантисс в обратном коде  

Проверка и выравнивание порядков  ; ∆*p*=2, 

Порядок числа А меньше, следовательно требуется выполнить сдвиг мантиссы *mA* на ∆*p* разрядов вправо, т.е. выполнить увеличение порядка:, 

Сложение мантисс и проверка на нарушение нормализации результата



Ответ: 

Выполнить сложение самостоятельно.

1. ;  в обратном коде, m=8 разрядов, p=4 разряда.
2. ;  в обратном коде, m=8 разрядов, p=4 разряда.
3. ;  в дополнительном коде, m=8 разрядов, p=4 разряда.
4. ;  в обратном коде, m=8 разрядов, p=4 разряда.
5. ;  в обратном коде, m=8 разрядов, p=4 разряда.
6. ;  в обратном коде, m=8 разрядов, p=4 разряда.

Решение:

1. ;  в обратном коде, m=8 разрядов, p=4 разряда.

Запись мантисс в прямом коде  

Запись мантисс в обратном коде  

Проверка и выравнивание порядков  ; ∆*p*=2, 

Порядок у числа B меньше, следовательно требуется выполнить сдвиг мантиссы на ∆*p* разряда вправо, т.е. выполнить увеличение порядка:, 

Сложение мантисс и проверка на нарушение нормализации результата



Ответ: 

2. ;  в обратном коде, m=8 разрядов, p=4 разряда.

Запись мантисс в прямом коде  

Запись мантисс в обратном коде  

Проверка и выравнивание порядков  ; ∆p=3, 

Порядок у числа A меньше, следовательно требуется выполнить сдвиг мантиссы на ∆p разрядов вправо, т.е. выполнить увеличение порядка:, 

Сложение мантисс и проверка на нарушение нормализации результата



Ответ: 

3. ;  в дополнительном коде, m=8 разрядов, p=4 разряда.

Запись мантисс в прямом коде  

Запись мантисс в дополнительном коде  

Проверка и выравнивание порядков  ; ∆p=2, 

Порядок у числа A меньше, следовательно, требуется выполнить сдвиг мантиссы на ∆p разрядов вправо, т.е. выполнить увеличение порядка:, 

Сложение мантисс и проверка на нарушение нормализации результата



Ответ: 

4. ;  в обратном коде, m=8 разрядов, p=4 разряда.

Запись мантисс в прямом коде  

Проверка и выравнивание порядков  ; ∆p=4, 

Порядок у числа A меньше, следовательно требуется выполнить сдвиг мантиссы на ∆*p* разрядов вправо, т.е. выполнить увеличение порядка: , 

Сложение мантисс и проверка на нарушение нормализации результата



Ответ: 

5. ;  в обратном коде, m=8 разрядов, p=4 разряда.

Запись мантисс в прямом коде  

Запись мантисс в обратном коде  

Проверка и выравнивание порядков  ; ∆p=2, 

Порядок у числа B меньше, следовательно требуется выполнить сдвиг мантиссы на ∆p разрядов вправо, т.е. выполнить увеличение порядка: , 

Сложение мантисс и проверка на нарушение нормализации результата



Ответ: 

6. ;  в обратном коде, m=8 разрядов, p=4 разряда.

Запись мантисс в прямом коде  

Запись мантисс в обратном коде  

Проверка и выравнивание порядков  ; ∆p=1, 

Порядок у числа A меньше, следовательно требуется выполнить сдвиг мантиссы на ∆p разрядов вправо, т.е. выполнить увеличение порядка: , 

Сложение мантисс и проверка на нарушение нормализации результата



Ответ: 

# Формальное представление чисел в стандарте IEEE 754 для любого формата точности

Операции над числами в формате с плавающей запятой имеют существенные отличия от аналогичных операций целочисленной арифметики, поэтому их обычно реализуют с помощью самостоятельного операционного устройства. Как и целочисленное ОПУ, операционное устройство для чисел в формате с плавающей запятой.

Рассмотрим основные положения записи чисел в стандарте IEEE 754.

Чтобы записать число в стандарте IEEE 754 или восстановить его, необходимо знать три параметра:

* S- бит знака (31-й бит)
* E- смещенная экспонента (30-23 биты)
* M - остаток от мантиссы (22-0 биты)

Это целые числа, которые записанные в числе IEEE 754 в двоичном виде.

Приведём формулу для получения десятичного числа из числа IEEE754 одинарной точности:

, **(1)**

где *A* – десятичное число.

Основное применение в технике и программировании получили форматы 32 и 64 бита. Например, в VB используют типы данных single (32 бита) и double (64 бита). В Си аналогично используют float (32 бита) и double (64 бит).

Чтобы представить число в формате IEEE 754 необходимо привести его к двоичному нормализованному виду.

Описание преобразования в 32 битный формат IEEE 754:

1. Число может быть положительным или отрицательным .  
   Поэтому отводится 1 бит для обозначения знака числа:   
   0 ­­- положительное  
   1 - отрицательное  
   Этот самый старший бит в 32 битной последовательности.
2. Далее пойдут биты экспоненты, для этого выделяют 1 байт (8 бит).  
   Экспонента может быть, как и число, со знаком “+” или “-“.  
   Для определения знака экспоненты, чтобы не вводить ещё один бит знака, добавляют смещение к экспоненте в половину байта +127(0111 1111). То есть, если наша экспонента = +7 (+111 в двоичной), то смещенная экспонента = 7+127=134. А если бы, наша экспонента была -7 , то смещенная экспонета=127-7 =120. Смещенную экспоненту записывают в отведенные 8 бит. При этом, когда нам будет нужно получить экспоненту двоичного числа, мы просто отнимем 127 от этого байта.
3. Оставшиеся 23 бита отводят для мантиссы.

Но, у нормализованной двоичной мантиссы первый бит всегда равен 1, так как число лежит в диапазоне 1<=M<2.

Нет смысла записывать единицу в отведенные 23 бита, поэтому в отведенные 23 бита записывают остаток от мантиссы.

(sign bit)

Бит знака

W – бит (bit)

S

E

M

1

b – бит (bit)

n – бит (bit)

(offset exponent)

Смещенная экспонента

(Mantissa)

мантисса

**Рис. 2.** Представление числа в формате IEEE 754

где:

* S - бит знака, если S=0 - положительное число; S=1 - отрицательное число
* E - смещенная экспонента двоичного числа;  
  exp2 = E - (2(b-1) - 1) - экспонента двоичного нормализованного числа с плавающей точкой   
  (2(b-1) -1) - заданное смещение экспоненты (в 32-битном IEEE 754 оно равно +127 см.выше)
* M - остаток мантиссы двоичного нормализованного числа с плавающей точкой

## Представление нормализованного числа в формате IEEE 754

Формула вычисления десятичных чисел с плавающей точкой, из чисел представленных в стандарте IEEE754:

**(2)**

Используя формулу (1) вычислим формулы для нахождения десятичных чисел из форматов одинарной (32 бита) и двойной (64 бита) точности IEEE 754:

0

22

30

31

1

8

23

32

0

51

32

63

1

11

52

64

**Рис.2**. Формат числа одинарной (32бита) и двойной точности (64 бита)

## Представление денормализованного числа и других чисел в формате IEEE 754

Если применить формулу (2) для вычисления минимального и максимального числа одинарной точности представленного в IEEE754, то получим следующие результаты:

* 00 00 00 00hex= 5,87747175411144e-39 (минимальное положительное число)
* 80 00 00 00hex=-5,87747175411144e-39 (минимальное отрицательное число)
* 7f ff ff ffhex= 6,80564693277058e+38 (максимальное положительное число)
* ff ff ff ffhex=-6,80564693277058e+38 (максимальное отрицательное число)

Отсюда видно, что невозможно представить число нуль или бесконечность в заданном формате.

#### Поэтому формула (2) не применяется в следующих случаях:

1. Число IEEE754=00 00 00 00hex считается числом +0

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 31 | |  |  |  |  |  |  | 22 | |  |  |  |  |  |  | 15 | |  |  |  |  |  |  | 7 |  |  |  |  |  |  | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 00hex | | | | | | | | 00hex | | | | | | | | 00hex | | | | | | | | 00hex | | | | | | | |

Число IEEE754=80 00 00 00hex считается числом -0

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 31 | |  |  |  |  |  |  | 22 | |  |  |  |  |  |  | 15 | |  |  |  |  |  |  | 7 |  |  |  |  |  |  | 0 |
| 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 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 00hex | | | | | | | | 00hex | | | | | | | | 00hex | | | | | | | | 00hex | | | | | | | |

1. Число IEEE754=7F 80 00 00hex считается числом +∞

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 31 | |  |  |  |  |  |  | 22 | |  |  |  |  |  |  | 15 | |  |  |  |  |  |  | 7 |  |  |  |  |  |  | 0 |
| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7Fhex | | | | | | | | 80hex | | | | | | | | 00hex | | | | | | | | 00hex | | | | | | | |

Число IEEE754=FF 80 00 00hex считается числом -∞

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 31 | |  |  |  |  |  |  | 22 | |  |  |  |  |  |  | 15 | |  |  |  |  |  |  | 7 |  |  |  |  |  |  | 0 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FFhex | | | | | | | | 80hex | | | | | | | | 00hex | | | | | | | | 00hex | | | | | | | |

1. Числа IEEE754=FF (1xxx)X XX XXhex не считается числами (NAN), кроме случая п.2  
   Числа IEEE754=7F (1xxx)X XX XXhex не считается числами (NAN), кроме случая п.2  
   Число представленное в битах с 0...22 могут быть любым числом кроме 0 (т.е.+∞ и -∞ ).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 31 | |  |  |  |  |  |  | 22 | |  |  |  |  |  |  | 15 | |  |  |  |  |  |  | 7 |  |  |  |  |  |  | 0 |
| x | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
|  | | | | | | | |  | | | | | | | |  | | | | | | | |  | | | | | | | |

1. Числа IEEE754=(x000) (0000) (0xxx)X XX XXhex считаются денормализованными числами, за исключением чисел п.1 (то есть -0 и +0)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 31 | |  |  |  |  |  |  | 22 | |  |  |  |  |  |  | 15 | |  |  |  |  |  |  | 7 |  |  |  |  |  |  | 0 |
| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
|  | | | | | | | |  | | | | | | | |  | | | | | | | |  | | | | | | | |
| 31 | |  |  |  |  |  |  | 22 | |  |  |  |  |  |  | 15 | |  |  |  |  |  |  | 7 |  |  |  |  |  |  | 0 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
|  | | | | | | | |  | | | | | | | |  | | | | | | | |  | | | | | | | |

### Формула расчета денормализованных чисел:

**(3)**

**Пояснения к исключительным числам:**

1. **- ∞/ +∞** числа, которые больше границ диапазона представления чисел считаются бесконечными.
2. **Не числа NAN(No a Numbers)** к ним относятся символы, или результаты недопустимых операций.
3. **Денормализованные числа**. Это числа, мантиссы которых лежат в диапазоне 0.1 <= M <1.  
   Денормализованные числа находятся ближе к нулю, чем нормализованные. Денормализованные числа как бы разбивают минимальный разряд нормализованного числа на некоторое подмножество. Сделано так потому, что в технической практике чаще встречаются величины близкие к нулю.

# Диапазон чисел с плавающей запятой

## Вычисление границ диапазона для чисел одинарной точности IEEE 754

Зная формат чисел с одинарной точностью стандарта IEEE 754 можно посчитать границы диапазона представления действительных чисел в этом формате. Для этого подставим значения максимальных и минимальных абсолютных чисел IEEE 754 в формулы (2) и (3).

Минимальное нормализованное число по (абсолютное):

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 31 | |  |  |  |  |  |  | 22 | |  |  |  |  |  |  | 15 | |  |  |  |  |  |  | 7 |  |  |  |  |  |  | 0 |
| x | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 80 или 00 | | | | | | | | 00hex | | | | | | | | 00hex | | | | | | | | 00hex | | | | | | | |

00 80 00 00 = 2-126∙(1+0/223)= 2-126 ≈ 1,17549435∙e-38  
80 80 00 00 = -2-126∙(1+0/223)=-2-126 ≈ -1,17549435∙e-38  
  
Максимальное денормализованое число (абсолютное):

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 31 | |  |  |  |  |  |  | 22 | |  |  |  |  |  |  | 15 | |  |  |  |  |  |  | 7 |  |  |  |  |  |  | 0 |
| x | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 80 или 00 | | | | | | | | 7F | | | | | | | | FF | | | | | | | | FF | | | | | | | |

00 7F FF FF = 2-126∙(1-2-23) ≈ 1,17549421∙e-38  
80 7F FF FF = -2-126∙(1-2-23) ≈ -1,17549421∙e-38  
Отсюда видно что минимальное нормализированное число граничит с максимальным денормализированным.

Минимальное денормализованное число (абсолютное):

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 31 | |  |  |  |  |  |  | 22 | |  |  |  |  |  |  | 15 | |  |  |  |  |  |  | 7 |  |  |  |  |  |  | 0 |
| x | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 80 или 00 | | | | | | | | 00 | | | | | | | | 00 | | | | | | | | 01 | | | | | | | |

00 00 00 01 = 2-126∙ 2-23= 2-149 ≈ 1,40129846∙e-45  
80 00 00 01 = -2-126∙ 2-23= 2-149 ≈ -1,40129846∙e-45  
Это число граничит с нулем.

Максимальное нормализированное число (абсолютное):

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 31 | |  |  |  |  |  |  | 22 | |  |  |  |  |  |  | 15 | |  |  |  |  |  |  | 7 |  |  |  |  |  |  | 0 |
| x | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| FF или 7F | | | | | | | | 7F | | | | | | | | FF | | | | | | | | FF | | | | | | | |

7F 7F FF FF = 2127∙(2-2-23) ≈ 3,40282347∙e+38  
FF 7F FF FF = -2127∙(2-2-23) ≈ -3,40282347∙e+38  
Это число граничит с бесконечностью.

## Полный диапазон чисел одинарной точности (32 бит) по стандарту IEEE754

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Отрицательные числа | | | | | | | | | | | | | | Положительные числа | | | | | | | | | | | | | |
| FFFFFFFF | FF800001 | FF800001 | FF800000 | FF7FFFFF | FF800001 | FF800001 | FF800001 | FF800001 | 80800000 | 807FFFFF | FF800001 | 80000001 | 80000000 | 00000000 | 00000001 | FF800001 | 007FFFFF | 00800000 | FF800001 | FF800001 | FF800001 | FF800001 | 7F7FFFFF | 7F800000 | 7F900001 | FF800001 | 7FFFFFFF |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  | —∞ | —3,40282347e+38 |  |  |  |  | —1,17549435e—38 | —1,17549421e—38 |  | —1,40129846e—45 | —0 | +0 | 1,40129846e—45 |  | 1,17549421e—38 | 1,17549435e—38 |  |  |  |  | 3,40282347e+38 | +∞ |  |  |  |
|  | | — НУЛЕВЫЕ ЧИСЛА | | | | | | | | | | | | | | | | | | | | | | | | | |
|  | | — ДЕНОРМАЛИЗОВАННЫЕ ЧИСЛА (РАССЧИТЫВАЮТСЯ ПО ФОРМУЛЕ №3) | | | | | | | | | | | | | | | | | | | | | | | | | |
|  | | — НОРМАЛИЗОВАННЫЕ ЧИСЛА (РАССЧИТЫВАЮТСЯ ПО ФОРМУЛЕ №2) | | | | | | | | | | | | | | | | | | | | | | | | | |
|  | | — БЕСКОНЕЧНОСТЬ | | | | | | | | | | | | | | | | | | | | | | | | | |
|  | | — НЕ ЧИСЛА (NANs) | | | | | | | | | | | | | | | | | | | | | | | | | |

**Рис.4 .**Диапазон чисел формата одинарной точности (32 бита) представленных по стандарту

IEEE 754

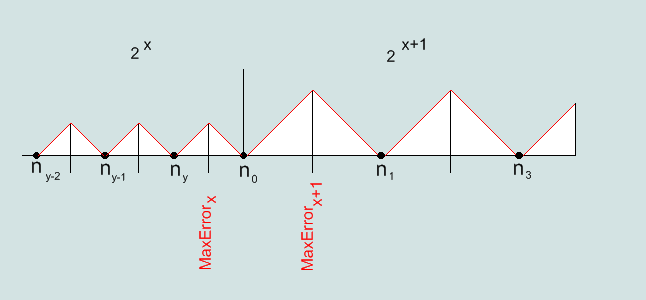
## Полный диапазон чисел двойной точности (64 бит) по стандарту IEEE754

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Отрицательные числа | | | | | | | | | | | | | | Положительные числа | | | | | | | | | | | | | |
| FFFFFFFF FFFFFFFF | FF800001 | FFF00000 00000001 | FFF00000 00000000 | FFFEFFFF FFFFFFFF | FF800001 | FF800001 | FF800001 | FF800001 | 80100000 00000000 | 800FFFFF FFFFFFFF | FF800001 | 80000000 00000001 | 80000000 00000000 | 00000000 00000000 | 00000000 00000001 | FF800001 | 000FFFFF FFFFFFFF | 00100000 00000000 | FF800001 | FF800001 | FF800001 | FF800001 | 7FEFFFFF FFFFFFFF | 7FF00000 00000000 | 7FF00000 00000001 | FF800001 | 7FFFFFFF FFFFFFFF |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| NAN | | | —∞ | —1,797693134862315708e+308 |  |  |  |  | —2,225073858507201383e—308 | —2,225073858507200889e —308 |  | —4,940656458412465441e—324 | —0 | +0 | 4,940656458412465441e—324 |  | 2,225073858507200889e —308 | 2,225073858507201383e—308 |  |  |  |  | 1,797693134862315708e+308 | +∞ | NAN | | |
|  | | — НУЛЕВЫЕ ЧИСЛА | | | | | | | | | | | | | | | | | | | | | | | | | |
|  | | — ДЕНОРМАЛИЗОВАННЫЕ ЧИСЛА (РАССЧИТЫВАЮТСЯ ПО ФОРМУЛЕ №3) | | | | | | | | | | | | | | | | | | | | | | | | | |
|  | | — НОРМАЛИЗОВАННЫЕ ЧИСЛА (РАССЧИТЫВАЮТСЯ ПО ФОРМУЛЕ №2) | | | | | | | | | | | | | | | | | | | | | | | | | |
|  | | — БЕСКОНЕЧНОСТЬ | | | | | | | | | | | | | | | | | | | | | | | | | |
|  | | — НЕ ЧИСЛА (NANs) | | | | | | | | | | | | | | | | | | | | | | | | | |

**Рис.5 .**Диапазон чисел формата двойной точности (64 бита) представленных по стандарту IEEE 754

# Точность представления вещественных чисел в формате IEEE754.

Числа представленные в формате IEEE754 представляют конечное множество, на которое отображается бесконечное множество вещественных чисел. Поэтому исходное число может быть представлено в формате IEEE754 с ошибкой.



**Рис.5** Функция ошибки точности представления числа в IEEE754

Абсолютная максимальная ошибка для числа в формате IEEE754 равна в пределе половине шага чисел. Шаг чисел удваивается с увеличением экспоненты двоичного числа на единицу. То есть, чем дальше от нуля, тем шире шаг чисел в формате IEEE754 по числовой оси.  
 Шаг денормализованных чисел равен **2(E-149)** (Single) и **2(E-1074)**(Double). Соответственно предел макс. абсолютной ошибки будет равен 1/2 шага числа: **2(E-150)** (Single) и **2(E-1075)** (Double).  
 Относительная ошибка в % будет равна: **(2(E-150)/F)\*100%**(Single) и **(2(E-1075)/F)\*100%** (Double).  
 Шаг нормализованных чисел равен **2(E-150)** (Single) и **2(E-1075)**(Double). Соответственно предел макс. абсолютной ошибки будет равен 1/2 шага числа: **2(E-151)** (Single) и **2(E-1076)** (Double).  
 Относительная ошибка в % будет равна: **(2(E-151)/F)\*100%**(Single) и **(2(E-1076)/F)\*100%** (Double).

Максимальная относительная ошибка для денормализованного числа(single/double):

Максимальная относительная ошибка нормализованного числа(single):

Максимальная относительная ошибка нормализованного числа(double):

Максимальная возможная ошибка для чисел Single:

|  |  |  |  |
| --- | --- | --- | --- |
| IEEE754, hex | число, dec | абсолютная ошибка, dec | относительная, % |
| 00000001 | 2-149 ≈1,401298e-45 | 2-150≈0,700649e-45 | =50 |
| 00000002 | 2-148 ≈2,802597e-45 | 2-150≈0,700649e-45 | =25 |
| 00000032 | ≈7,00649e-44 | 2-150≈0,700649e-45 | =1 |
| 007FFFFF | ≈1,175494e-38 | 2-150≈0,700649e-45 | ≈5,96e-6 |
| 00800001 | ≈1,175494e-38 | 2-149 ≈1,401298e-45 | ≈11,9209e-6 |
| 0DA24260 | ≈1,0e-30 | 2-123 ≈9,4039e-38 | ≈9,4039e-6 |
| 1E3CE508 | ≈1,0e-20 | 2-90 ≈8,0779e-28 | ≈8,0779e-6 |
| 2EDBE6FF | ≈1,0e-10 | 2-57 ≈6,9389e-18 | ≈6,9389e-6 |
| 3F800000 | ≈1,0 | 2-23 ≈1,192e-7 | ≈11,9209e-6 |
| 41200000 | ≈10,0 | 2-20 ≈9,5367e-7 | ≈9,5367e-6 |
| 42C80000 | ≈1,0e+2 | 2-17 ≈7,6294e-6 | ≈7,62939e-6 |
| 501502F9 | ≈1,0e+10 | 210 ≈1,024e+3 | ≈10,24e-6 |
| 60AD78EC | ≈1,0e+20 | 243 ≈8,7961e+12 | ≈8,7961e-6 |
| 7149F2CA | ≈1,0e+30 | 276 ≈7,5558e+22 | ≈7,5558e-6 |
| 7F7FFFFF | ≈+3,40282e+38 | 2104 ≈2,02824e+31 | ≈5,96e-6 |

Максимальная возможная ошибка для чисел Double:

|  |  |  |  |
| --- | --- | --- | --- |
| IEEE754, hex | число, dec | абсолютная ошибка, dec | относительная, % |
| 00000000 00000001 | 2-1074 ≈4,940656e-324 | 2-1075≈2,470328e-324 | =50 |
| 00000000 00000002 | 2-1073 ≈9,881313e-324 | 2-1075≈2,470328e-324 | =25 |
| 00000000 00000032 | ≈2,470328e-322 | 2-1075≈2,470328e-324 | =1 |
| 000FFFFF FFFFFFFF | ≈2,225073e-308 | 2-1075≈2,470328e-324 | ≈1,110223e-14 |
| 00100000 00000001 | ≈2,225074e-308 | 2-1074 ≈4,940656e-324 | ≈2,220446e-14 |
| 2B2BFF2E E48E0530 | ≈1,0e-100 | 2-385 ≈1,268971e-116 | ≈1,268971e-14 |
| 3FF00000 00000000 | =1,0 | 2-52 ≈2,220446e-16 | ≈2,220446e-14 |
| 54B249AD 2594C37D | ≈1,0e+100 | 2280 ≈1,942669e+84 | ≈1,942669e-14 |
| 6974E718 D7D7625A | ≈1,0e+200 | 2612 ≈1,699641e+184 | ≈1,699641e-14 |
| 7FEFFFFF FFFFFFFF | ≈1,79769e+308 | 2971 ≈1,99584e+292 | ≈1,110223e-14 |

Из выше приведенного следует, что основная масса чисел в формате IEEE754 имеет стабильную небольшую относительную погрешность: Максимально возможная относительная погрешность для числа Single составляет 2-23\*100% =11,920928955078125e—6 %  
Максимально возможная относительная погрешность для числа Double составляет 2-52\*100% =2,2204460492503130808472633361816e— 14 %

**Домашнее задание.**

Вариант выбирается по формуле K=(N-1) mod S +1, где N-ваш номер в списке группы, S- число вариантов (20), K-номер варианта

Задание 1. Числа A и B заданы в прямом коде. Сложить A+B в обратном коде, m=8 разрядов, p=4 разряда

Задание 2. Вычесть A-B в дополнительном коде, m=8 разрядов, p=4 разряда.

Задание 3. Представить десятичное число в формате с плавающей запятой стандарта IEEE 754, четырехбайтный формат в прямом коде; определить абсолютную и относительную погрешность числа (можно пользоваться калькулятором)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Вари-ант | Задание 1,2 | Задание 3 | Вари-  ант | Задание 1,2 | Задание 3 |
| 1 | A=–0,11010110⋅23  B= 0,10010111⋅25 | 143,41 | 11 | A=–0,11010010⋅21  B= 0,11110111⋅20 | 0,00056 |
| 2 | A=–0,11010110⋅2-2  B= 0,10100111⋅2-2 | -167,892 | 12 | A=–0,11010110⋅2-6  B= 0,10110111⋅2-7 | -5,377 |
| 3 | A=–0,11010100⋅2-1  B= 0,10011011⋅24 | 15,70 | 13 | A=–0,11010100⋅2-15  B= 0,10010101⋅2-14 | -12,18 |
| 4 | A=–0,11110110⋅23  B= 0,10010001⋅23 | -80,34 | 14 | A=–0,10010110⋅213  B= 0,10010111⋅215 | 44,82 |
| 5 | A=–0,11010110⋅23  B= 0,10010111⋅25 | 11,38 | 15 | A=–0,11110110⋅23  B= 0,11010100⋅25 | 16,9 |
| 6 | A=–0,10100110⋅2-4  B= 0,10011011⋅2-6 | 7,15 | 16 | A=–0,11010110⋅2-4  B= 0,11010111⋅2-5 | -10,4 |
| 7 | A=–0,11001110⋅2-1  B= 0,10000111⋅2-1 | -0,021 | 17 | A=–0,11010110⋅24  B= 0,11000111⋅23 | 5643 |
| 8 | A=–0,11001000⋅22  B= 0,10110011⋅25 | 167235 | 18 | A=–0,11010110⋅233  B= 0,10011111⋅234 | -8,67 |
| 9 | A=–0,11010110⋅2-2  B= 0,10010111⋅2-1 | -163,3 | 19 | A=–0,11110110⋅23  B= 0,10010111⋅25 | -10,2 |
| 10 | A=–0,11111010⋅21  B= 0,11110101⋅23 | -0,0003 | 20 | A=–0,11010110⋅2-13  B= 0,10010101⋅215 | 0,02 |