Алгоритм умножения float point чисел

**Оглавление**

[Базовые понятия 1](#_Toc1321153905)

[Введение 2](#_Toc1201259857)

[Умножение 2](#_Toc1032915010)

[Перемножение мантисс 3](#_Toc387407195)

[Суммирование экспонент 3](#_Toc1689205972)

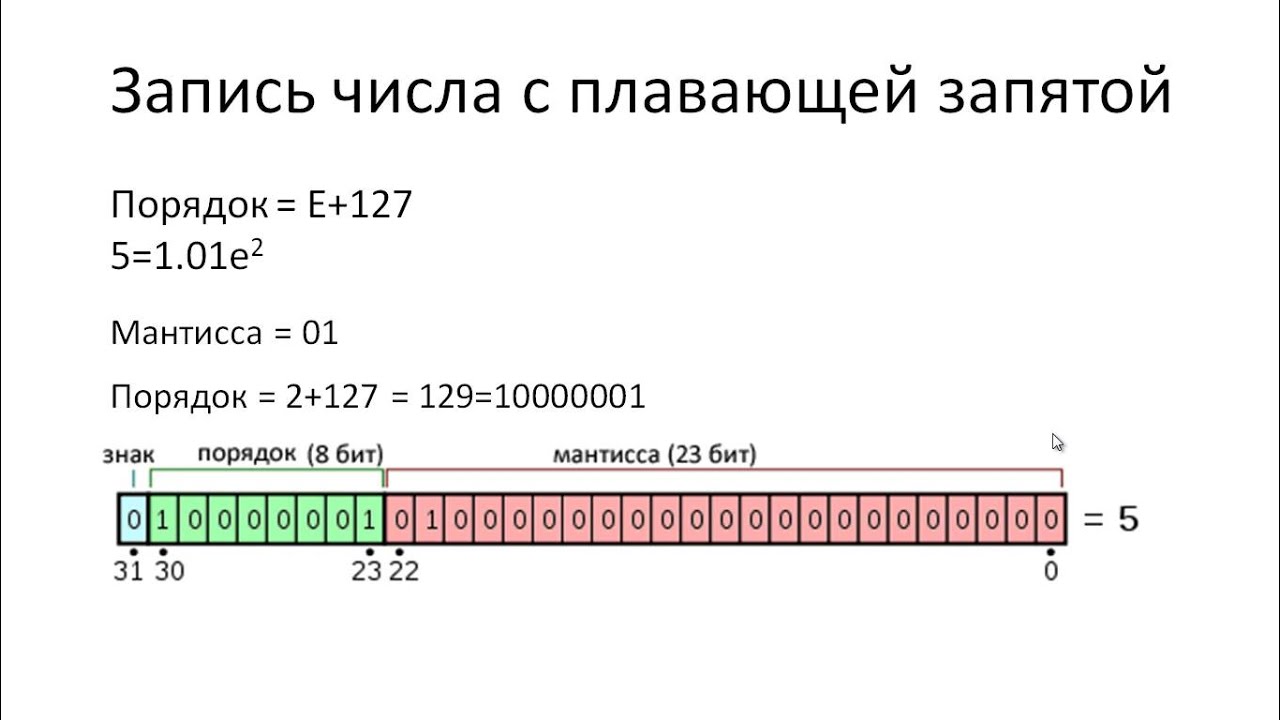
[Определение знака 4](#_Toc2076118339)

[Запись в результат 4](#_Toc1785373249)

[Краткий алгоритм 4](#_Toc368375346)

# Базовые понятия

**Число с плавающей запятой** — экспоненциальная форма представления вещественных (действительных) чисел, в которой число хранится в виде мантиссы и порядка (показателя степени).



**IEEE 754** — широко используемый стандарт IEEE, описывающий формат представления чисел с плавающей точкой. Используется в программных (компиляторы разных языков программирования) и аппаратных (CPU и FPU) реализациях арифметических действий.

# Введение

Cтандарт **IEEE 754**, описывает как должны быть записаны числа с плавающей запятой и как выполнять арифметические операции над ними. Его можно найти в открытом доступе.

Оригинал документа: [IEEE Std 754™-2008 (Revision of IEEE Std 754-1985), IEEE Standard for Floating-Point Arithmetic (andrewt0301.github.io)](https://andrewt0301.github.io/hse-acos-course/part1ca/07_FP/ieee-754-2008.pdf)

Для удобного представления float point чисел, использую онлайн конвертер: [IEEE-754 Floating Point Converter (h-schmidt.net)](https://www.h-schmidt.net/FloatConverter/IEEE754.html)

# Умножение

Для умножения двух чисел с плавающей запятой нужно: перемножить их мантиссы, суммировать экспоненты и выполнить операцию xor над знаками. В зависимости от размера числа алгоритм не меняется.

# Краткий алгоритм

1. Переводим числа в формат **IEEE 754**
2. Разделяем их на: знак, экспоненту и мантиссу
3. Перемножаем мантиссы
   1. Добавляем мнимые единицы
   2. Выполняем операцию умножения
   3. Проверяем на *переполнение*
      1. Если переполнение есть – сдвигаем мантиссу на 1 бит вправо, чтобы слева от точки не было лишнего разряда
      2. Запоминаем, что нужно будет увеличить порядок (добавить единицу к результирующей экспоненте)
   4. Проверяем на округление
      1. Округляем, если выполняется условие: (i(24) and i(23)
      2. Округляем, если выполняется условие: (i(23) and (i(22-0))
   5. Удаляем лишние нули и мнимую единицу
   6. Сохраняем результат
4. Суммируем экспоненты
   1. Суммируем экспоненты
   2. Если было *переполнение* в пункте 3.3 добавляем единицу
   3. Отнимаем максимальную экспоненту (для fp32 это - 127)
   4. Сохраняем результат
5. Находим знак
   1. Выполняем операцию xor между знаками
   2. Сохраняем результат
6. Соединяем сохраненные результаты
7. Готово

# Пример

Будем умножать на примере. Возьмем два числа: 5,5 и 5,5. Мы знаем, что 5,5 \* 5,5 = 30,25.

Переводим в формат IEEE 754, fp32:

5,5 = 101,1 = 1,011e2;

Мантисса = 011;

Экспонента = 2;

Результат в формате **IEEE 754:**



## Перемножение мантисс

Это самый трудоемкий и сложный для понимания этап умножения, на который стоит максимально обратить внимание.

Наши мантиссы: 011 и 011. Обратим внимание, что у мантисс существует мнимая единица, описанная в стандарте IEEE 754, добавим ее и получим: 1.011 и 1.011.

Перемножим числа, результат:

1.11100100000000000000000 000000000000000000000000,

обратим внимание, что число увеличилось в два раза.

Проверим результат на *переполнение*, в нашем случае его нет.

Если мы попробуем умножить мантиссы чисел 12 и 11,5, то получим такой результат:

1.10000000000000000000000 \* 1.01110000000000000000000 = 10.001010000000000000000000 0000000000000000000000.

В **этом** примере *переполнение* есть, нормализуем число, сдвинув его на 1 бит вправо:

10.00101 >> 1 = 1.000101

Если была *нормализация*, в будущем, не забудем *инкрементировать* экспоненту.

В **нашем** примере, получился результат:

1.1110010000000000000000000 0000000000000000000000, который не нужно *нормализовывать.*

Теперь надо выполнить округление.

Условие: ((i(24) and i(23)) or (i(23) and (i(22-0))).

Если оно **истинно**, то делаем сдвиг вправо на размер мантиссы и *инкрементируем* результат. Если условие **ложно** — делаем только сдвиг.

На нашем примере условие **ложно**:

1.1110010000000000000000**0 0**00000000000000000000000

(если оба бита равны 1, то округляем)

или

1.11100100000000000000000 **000000000000000000000000**

(если самый левый, выделенный бит равен 1 и хотя бы один бит правее от него равен 1, то округляем).

Пример **истинных** условий:

1.010011111110011101101100 **1 0101110010101100000010**

- сработала *вторая* часть условия.

Или

1.01001111111001110110110**1** **1**0000000000000000000000

- сработала *первая* часть условия.

В нашем случае условие **ложно**, поэтому делаем сдвиг:

1.1110010000000000000000000 0000000000000000000000 >> 23 = 1.1110010000000000000000000

Удаляем мнимую единицу и получаем результат:

1110010000000000000000000



## Суммирование экспонент

Наши экспоненты = 10000001, в десятичном формате = 129. Суммируем: 129 + 129 = 258;

Отнимаем максимальный размер экспоненты, для fp32 - это = 127, выполняем: 258 - 127 = 131.

Формула максимального размера экспоненты

(Если была *нормализация* мантиссы, *инкрементируем* результат)

Переведем результат в двоичный формат: 131 = 10000011.

## Определение знака

Выполняем операцию xor между нашими знаками: 0 xor 0 = 0.

## Запись в результат

Соединяем результаты: 0 10000011 11100100000000000000000.



Проверяем: 1.111001e4 = 11110,01 = 30,25.

Результат - правильный.