К компьютерной технике вещественные алгебраические числа могут быть представлены в формате с фиксированной точкой (запятой) или в формате с плавающей точкой (запятой).

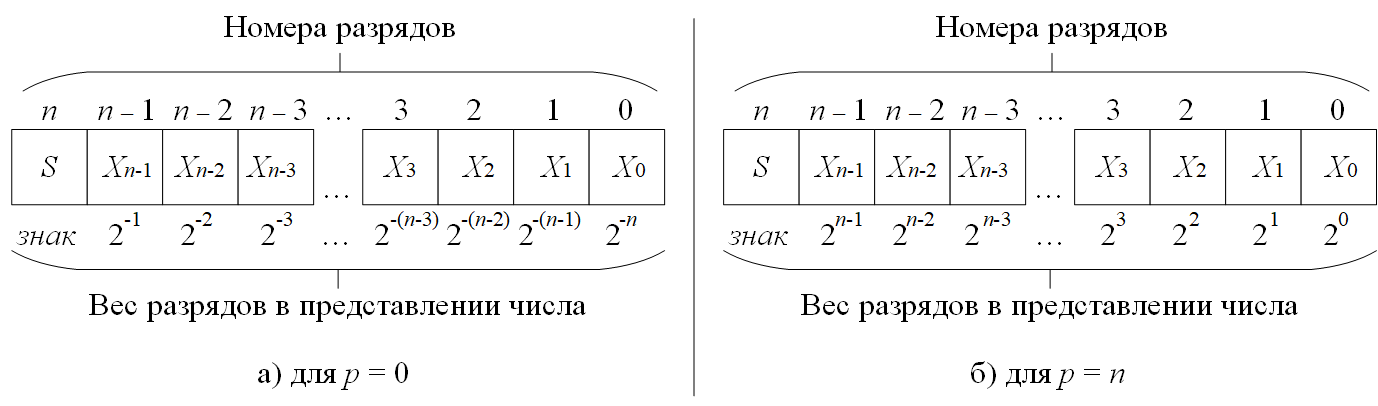
**Вещественные числа в формате с фиксированной точкой** также представляются в кодированной записи (прямой, обратный, дополнительный коды) и содержат поле знака и поле модуля. Поле модуля в данном случае хранит целую и дробные части, разделенную не обозначаемым, но подразумеваемым разделителем (точкой (запятой)).

При таком формате представления вещественного числа положение разделителя строго фиксировано, т. к. порядок p числа постоянен. Выбор величины порядка p числа в общем случае произволен. При этом положение разделителя целой и дробной части закрепляется (только подразумевается и никак не обозначается) в определенном месте относительно всех разрядов числа и сохраняется неизменным для всех чисел, представляемых в выбранной разрядной сетке вычислительной машины.

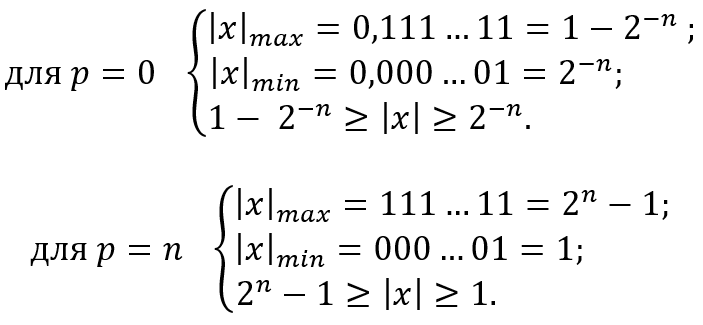
Наиболее распространены два способа закрепления разделителя:

* - перед старшим разрядом (p = 0), тогда все числа по модулю меньше единицы;
* - после младшего разряда (p = n), где n – количество разрядов цифровой части числа; тогда представляемые числа знаковые целые (с нулевой дробной частью).

Разрядную сетку чисел, представляемых в формате с фиксированной точкой, можно представить как показано на рисунке 1.13.

Рисунок 1.13 – Разрядная сетка чисел в формате с фиксированной точкой

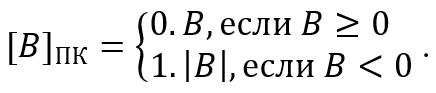
Тогда диапазоны представления таких n-разрядных чисел имеют вид:



Кодирование вещественных числе в формате с фиксированной точкой происходит аналогично, как для целых чисел.

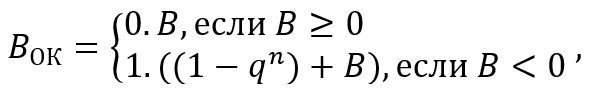
**Прямой базовый код**

При кодировании числа в прямой базовый код запись вещественного числа (правильной дроби) в формате с фиксированной точкой B формируется по следующему правилу:



**Обратный базовый код**

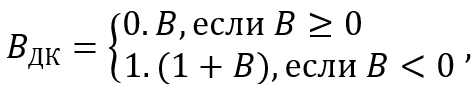
При кодировании числа в обратный базовый код запись дробного числа (правильной дроби) B формируется по следующему правилу:



где: 1 – qn – максимальная включенная граница диапазона изменения представляемых чисел.

**Дополнительный базовый код**

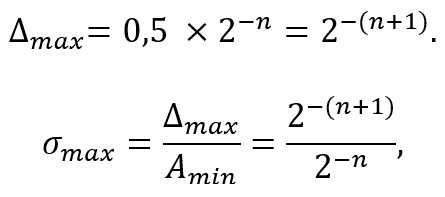
При кодировании числа в дополнительный базовый код запись дробного числа (правильной дроби) B формируется по следующему правилу:



где:   1 – максимальная невключенная граница диапазона изменения представляемых чисел.

Аналогично с целыми числами при выполнении арифметических операций используются модифицированные коды.

Одним из важнейших параметров представления вещественных чисел является ошибка представления, т. е. точность. Ошибка представления может быть абсолютной (Δ) или относительной (σ). Для формата с фиксированной точкой максимальные значения этих ошибок определяются следующим образом:



где Amin – минимальное, отличное от нуля, значение числа.

Таким образом, в худшем случае относительная ошибка может достигать сравнительно большого значения – 50 %.

Достоинства формата представления вещественных чисел с фиксированной точкой (запятой):

* требуются сравнительно несложные операционные арифметические устройства с высоким быстродействием;
* операции сложения и вычитания выполняются, как с целыми знаковыми числами;
* поддержка старых процессоров, микроконтроллеров и т. п.

В настоящее время формат представления вещественных чисел с фиксированной точкой (запятой) почти не используется из-за ряда таких недостатков:

* в таком формате существующими в языках программирования вещественными типами могут храниться сравнительно небольшие значения;
* происходит быстрый сдвиг точки (запятой) при умножении;
* необходимость предварительного масштабирования исходных чисел и необходимость слежения за положением точки (запятой)в ходе выполнения вычислений;
* относительно невысокая точность представляемых чисел.

**Вещественное число в форме с плавающей точкой (запятой)**, в общем случае, представляет собой смешанную дробь и имеет формат, представленный на рисунке 1.14.

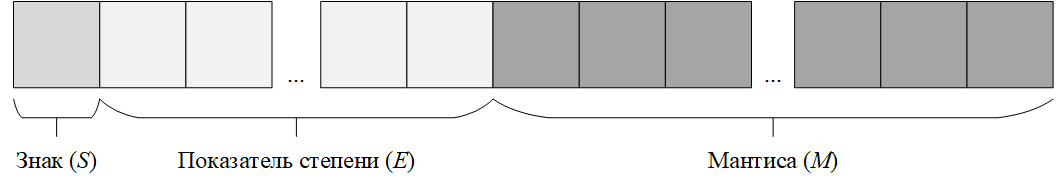


Рисунок 1.14 – Формат представления числа в форме с плавающей точкой

Местоположение условной точки в записи числа может быть различным, а т. к. сама точка в записи числа не присутствует, то для однозначного задания числа необходима не только его запись, но и информация о том, где в записи числа располагается условная точка, отделяющая целую и дробную части.

При представлении в форме с плавающей точкой (запятой) число Q представляется в виде трех частей:

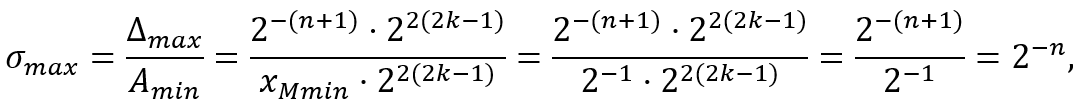
* знака числа (S);
* показателя степени (E) (экспонента), отображающий местоположение условной точки в записи числа, который представляется в виде целого числа с форматом фиксированной точки;
* мантиссы (M), отображающей запись числа, которая представляется в виде правильной дроби с форматом фиксированной точки. Таким образом целая часть числа должна быть равна единице, которая не фигурирует в записи, а записывается только дробная часть.

При этом количественная оценка числа Q определяется как:



где: q – основание системы счисления;  
       S – знак всего числа (нуль для положительных, единица для отрицательных);  
       E – показатель степени числа;  
       M – мантисса числа.

Максимальная относительная ошибка σmax представления чисел в формате с плавающей точкой определяется как:



где n – количество разрядов в записи модуля мантиссы числа,  
      k – количество разрядов в записи показателя степени числа.

Относительная ошибка при представлении чисел в форме с плавающей точкой существенно меньше, чем в случае с фиксированной точкой. Это, а также больший диапазон изменения представляемых чисел, является основным преимуществом представления чисел с плавающей точкой.

**Международный стандарт IEEE Standard for Binary Floating-Point Arithmetic 754-2008**

В настоящее время представление и обработка вещественных чисел в формате с плавающей точкой (запятой) регламентируется международным стандартом IEEE 754-2008 (IEEE Standard for Binary Floating-Point Arithmetic 754-2008).

Стандарт IEEE 754-2008 регламентирует такие вопросы как:

* представление нормализованных и денормализованных (субнормализованных) вещественных чисел с плавающей точкой;
* форматы вещественных чисел (к ним, по сути, приравниваются типы вещественных данных в типизированных языках программирования высокого уровня);
* смещение показателя степени для разных форматов вещественных чисел с плавающей точкой;
* представление специальных величин (±нуля, ± бесконечности, «не числа» (NaN/NaNs));
* правила округления вещественных чисел с плавающей точкой;
* точность представления вещественных чисел с плавающей точкой.

**Нормальная и нормализованная форма вещественных чисел с плавающей точкой**

Нормальная форма представления вещественного числа с плавающей точкой (ВЧсПТ) – это такая форма представления числа, при которой мантисса ВЧсПТ находится в интервале от нуля включительно до единицы не включительно (т. е. [0, 1)).

Нормальная форма представления ВЧсПТ имеет существенный недостаток – одно и то же число может быть представлено несколькими вариантами (т. е. неоднозначно).

Например, даже такое небольшое десятичное число 0,001 можно записать во множестве вариантов: 0,001×100 или 0,01×10−1 или 0,1×10−2 или 0,0001×101 или 0,00001×102 или 0,000001×103 и т. д.

Поэтому, в соответствии со стандартом IEEE 754-2008, ВЧсПТ должны быть представлены в нормализованной форме.

Нормализованная форма представления ВЧсПТ подразумевает, что мантисса десятичного числа принимает значения от 1 (включительно) до 10 (не включительно), а мантисса двоичного числа принимает значения от 1 (включительно) до 2 (не включительно) – т. е. может быть только единицей. Таким образом, в мантиссе слева от запятой (целая часть) до применения порядка должен находиться только один разряд, причем единичный. И, поскольку, старший двоичный разряд (целая часть) мантиссы ВЧсПТ в нормализованном виде всегда равен «1», то стандарт IEEE 754-2008 предписывает не записывать его (справедливо только для двоичных чисел), сэкономив таким образом один бит.

Для двоичной системы счисления – нормализованная мантисса должна иметь в старшем разряде модуля прямого кода значение равное единице, т. е. для двоичной системы счисления мантисса должна удовлетворять неравенству

1 > |M| ≥ 0,5

В нормализованной форме любое число представляется единственным образом. Исключения: ±нуль, ±∞ и неопределенность (не число, NaN).

Примеры нормализации ВЧсПТ

Нормальная форма записи чисел:   
5,010=101,02=101,0e0=10,1e1=1,01e2=0,101e3=1010,0e(-1);  
2,510=10,12=10,1e0=1,01e1=0,101e2=101,0e(-1).

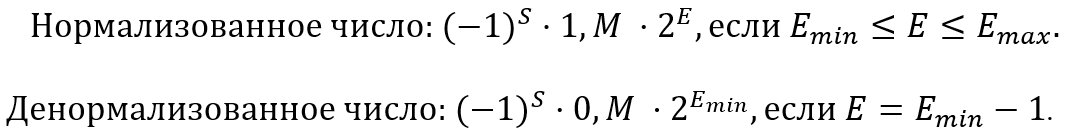
Нормализованная форма записи чисел:  
5,010=1,01e2;  
2,510=1,01e1.

Как видно из примеров, перемещая плавающую точку (запятую) на n разрядов влево – необходимо увеличить показатель степени e на значение n, а перемещая плавающую точку (запятую) на n разрядов вправо – необходимо уменьшить показатель степени e на значение n. При этом плавающая точка (запятая) двигается влево или вправо так, чтоб в целой части был только один и ненулевой разряд (единичный).

**Денормализованная (субнормализованная) форма вещественных чисел с плавающей точкой**

*Денормализация*– это способ повышения точности представления чисел, т. е. увеличение количества представимых ВЧсПТ значений около нуля. Таким образом, каждое значение денормализованного числа меньше самого маленького нормализованного значения числа.

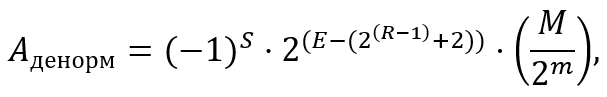
Отличия представления нормализованного и денормализованного чисел можно наглядно записать как:



Денормализованные числа позволяют более точно обрабатывать очень маленькие значения, однако, ввиду сложности реализации на аппаратном уровне, такая обработка чаще всего реализуется программно и выполняется в десятки раз медленнее, чем обработка нормализованных чисел.

Значения мантиссы денормализованных чисел лежит в диапазоне 0,1 ≤ M < 1.

Формула для расчета денормализованых чисел может быть записана как:



где Aденорм – представление денормализованного числа,  
      S – знак всего числа (нуль для положительных, единица для отрицательных),  
      E – показатель степени со смещением,  
      R – количество разрядов для представления показателя степени E,  
      (2(R–1) + 2) – заданное смещение для денормализации,  
      M – дробная часть мантиссы,  
      m – количество разрядов для представления мантисы M.

**Форматы представления вещественных чисел с плавающей точкой. Смещение показателя степени для разных форматов**

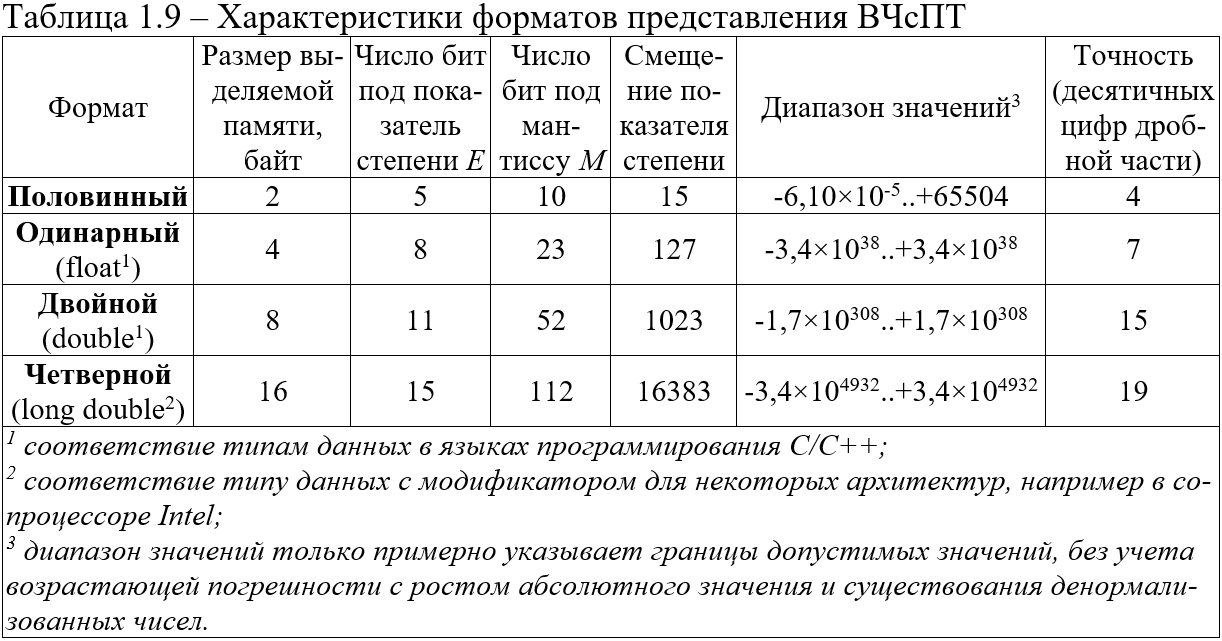
Стандарт IEEE 754-2008 определяет следующие форматы представления ВЧсПТ:

* формат половинной точности (Binary16, Half Precision);
* формат одинарной точности (Binary32, Single Precision, float);
* формат двойной точности (Binary64, Double Precision, double);
* формат четверной точности (Binary128, Quadruple Precision).

Форматы определяют такие характеристики вещественных чисел как:

* размер выделяемой памяти под хранения значения числа;
* точность представления вещественных чисел (число разрядов дробной части числа);
* количество разрядов для записи показателя степени;
* количество разрядов для записи мантиссы;
* допустимый диапазон изменения значений чисел;
* значение смещения показателя степени для возможности представления знаковых значений.

Наглядное сравнение характеристик форматов при представлении знаковых ВЧсПТ на 32-разрядной вычислительной машине представлено в таблице 1.9.



Рассмотри данные характеристики подробнее на примере формата одинарной точности (типа данных single в языке программирования Delphi или float в языках программирования С/С++).

Под хранение значений ВЧсПТ в формате одинарной точности выделяется 4 байта памяти, т. е. 4 байта × 8 = 32 бита. Из которых:

* под представление значения мантиссы выделяется 23 бита (с 0 по 22 биты);
* под представление значения показателя степени выделяется 8 бит (с 23 по 30 биты);
* под представление знака всего числа S выделяется 1 бит (старший, 31-й);
* целая часть мантиссы, равная единице для нормализованных чисел не записывается ни в какой бит, а только подразумевается.

Наглядно формат одинарной точности проиллюстрирован на рисунке 1.15.

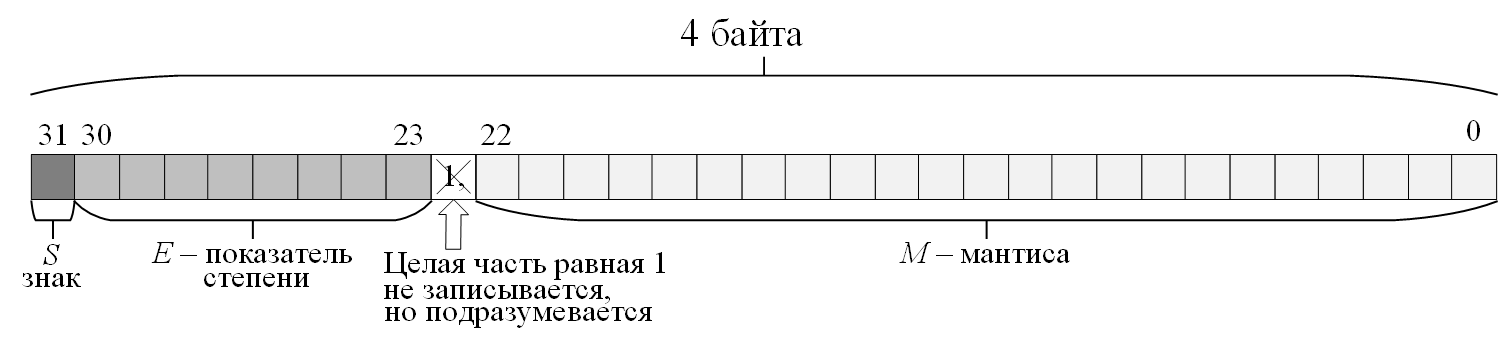
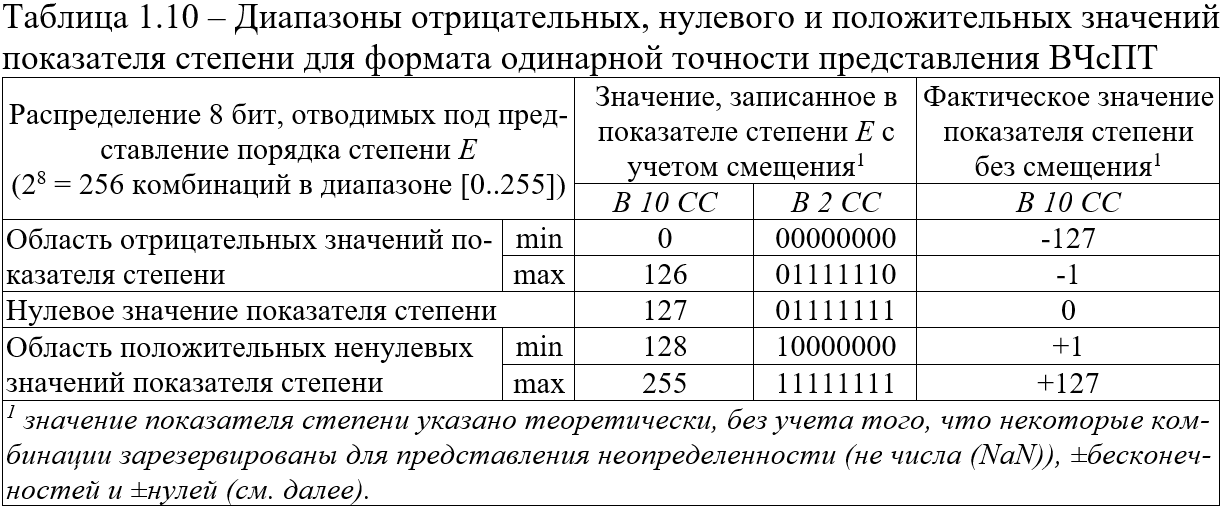


Рисунок 1.15 – Представление формата одинарной точности

Поскольку старший (31-й) бит хранит знак всего числа, т. е. по сути является знаком мантиссы, то как видно из рисунка 1.15 нигде не предусмотрен знак показателя степени, который может быть как положительный, так и отрицательный, в зависимости от того, каким получится число после нормализации.

Для того, чтоб учесть знак показателя степени, стандарт IEEE 754-2008 вводит понятие смещения показателя степени. В рассматриваемом примере формата одинарной точности под представление порядка степени выделяется 8 бит. Восемь двоичных разрядов (бит) могут представлять 28 = 256 уникальных двоичных комбинаций (эквивалентных в десятичной системе счисления множеству [0..255]). Разделив это множество поровну на отрицательную и положительную области, с учетом того, что нуль также является неотрицательным числом, получим, что значение 127 является серединой данного множества (то есть нулем). Тогда двоичные комбинации (00000000..01111110), соответствующие десятичному подмножеству от 0 до 126, будут соответствовать области отрицательных значений показателя степени; двоичная комбинация (01111111), соответствующая десятичному значению 127, будет являться началом положительной области (нулевым значением показателя степени); а двоичные комбинации (10000000..11111111), соответствующие десятичному подмножеству от 128 до 255, будут принадлежать области положительных ненулевых значений показателя степени. Наглядно данное смещение для формата одинарной точности показано в таблице 1.10.



Аналогичным образом рассчитываются смещения и диапазоны для других форматов точности представления ВЧсПТ.

*Пример представления числа с положительной степенью:*

Представим десятичное число +155,62510 в формате одинарной точности по стандарту IEEE 754-2008.

Поскольку заданное число представляет собой неправильную дробь, отдельно переведем в двоичную систему целую и дробные части:

+15510 = +100110112

+0,62510= +0,1012.

Таким образом, +155,62510 = +10011011,1012.

Обратная проверка:

10011011,1012 = 1∙27 + 0∙26 + 0∙25 + 1∙24 + 1∙23 + 0∙22 + 1∙21 + 1∙20 + 1∙2-1 +

+ 0∙2-2 + 1∙2-3 = 128 + 0 + 0 + 16 + 8 + 0 + 2 + 1 + 0,5 + 0 + 0,125 = 155,62510.

Для наглядности приведем полученное число к нормализованному виду в десятичной системе счисления:

+155,625*e*0 = +1,55625*e*2 (поскольку мы переместили плавающую точку на 2 разряда влево – увеличиваем показатель степени на значение равное двум).

Приведем полученное число к нормализованному виду в двоичной системе счисления:

10011011,101*e*0= 1,0011011101*e*7 (поскольку мы переместили плавающую точку на 7 разрядов влево – увеличиваем показатель степени на значение равное семи).

Таким образом, имеем мантиссу *M* = 1,0011011101 и показатель степени *E* = 7.

Выполним корректировку мантиссы смещением на значение 127 (для одинарной точности по условию):

127 + *e* = 127 + 7 =13410 = 100001102.

Поскольку заданное число положительное, то старший (31-й) знаковый бит *S* будет нулевым.

Далее слева направо (с 30-го по 23-й) бит будет записано значение показателя степени со смещением *E* равное 10000110.

Далее слева направо (с 22-го по 13-й биты, т. е. в старшие разряды мантиссы) будет записана дробная часть мантиссы *M*, т. к. целая часть (единица) не записывается с целью экономии бита, а подразумевается (поскольку число нормализованное, то в целой части двоичного числа всегда будет единица), при этом младшие 13 бит (с нулевого по 12-й) заполняются нулями, поскольку для дробной части младшие нулевые разряды являются незначащими в любой системе счисления.

Результат представления числа +155,62510 в формате одинарной точности по стандарту IEEE 754-2008 проиллюстрирован на рисунке 1.16.

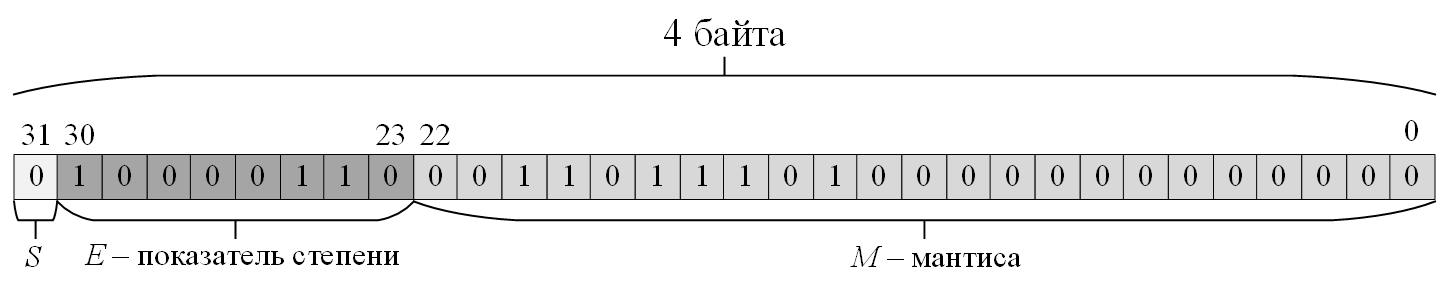


Рисунок 1.16 – Результат представления числа +155,62510

*Пример представления числа с отрицательной степенью:*

Представим десятичное число –0,12510 в формате одинарной точности по стандарту IEEE 754-2008.

Переведем в двоичную систему дробную часть:

–0,12510= –0,0012.

Обратная проверка:

0,0012 = 0∙20 + 0∙2-1 + 0∙2-2 + 1∙2-3 = 0 + 0 + 0 + 0,125 = 0,12510.

Для наглядности приведем полученное число к нормализованному виду в десятичной системе счисления:

0,125*e*0 = 1,25*e*-1 (поскольку мы переместили плавающую точку на 1 разряда вправо – уменьшаем показатель степени на значение равное единице).

Приведем полученное число к нормализованному виду в двоичной системе счисления:

0,001*e*0= 1,0*e*-3 (поскольку мы переместили плавающую точку на 3 разрядов вправо – уменьшаем показатель степени на значение равное трем).

Таким образом, имеем мантису *M* = 1,0 и показатель степени *E* = –3.

Выполним корректировку мантисы смещением на значение 127 (для одинарной точности по условию):

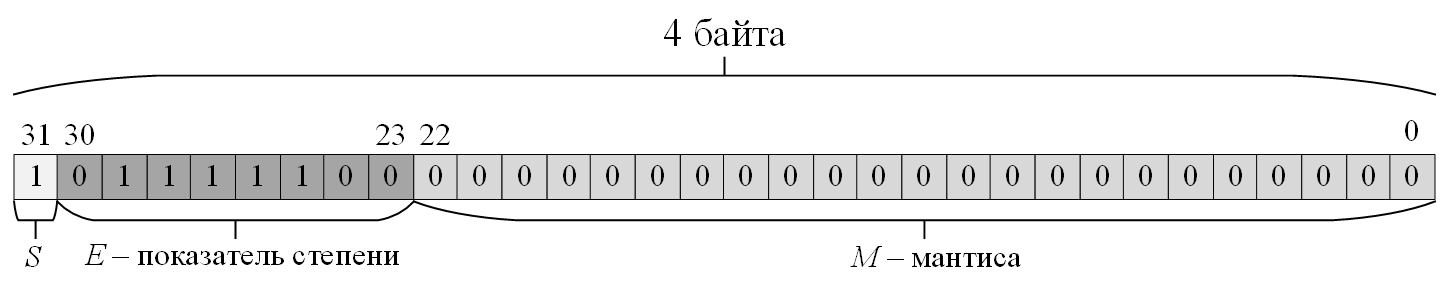
127 + *e* = 127 + (–3) =12410 = 011111002.

Поскольку заданное число отрицательное, то старший (31-й) знаковый бит *S* будет единичным.

Далее слева направо (с 30-го по 23-й) бит будет записано значение показателя степени со смещением *E* равное 01111100.

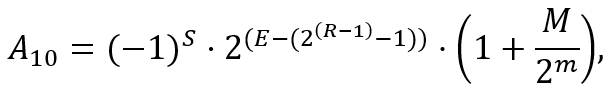
Далее слева направо (с 22-го по нулевой биты, т. е. в старшие разряды мантисы) будет записана дробная часть мантисы *M*, т. к. целая часть (единица) не записывается с целью экономии бита, а подразумевается (поскольку число нормализованное, то в целой части двоичного числа всегда будет единица), причем в данном случае дробная часть мантисы равна нулю.

Результат представления числа –0,12510 в формате одинарной точности по стандарту IEEE 754-2008 проиллюстрирован на рисунке 1.17.

Рисунок 1.17 – Результат представления числа –0,12510

*Пример обратного перевода ВЧсПТ, представленного по стандарту IEEE 754-2008, в десятичное*

Для обратного перевода можно воспользоваться следующей формулой:



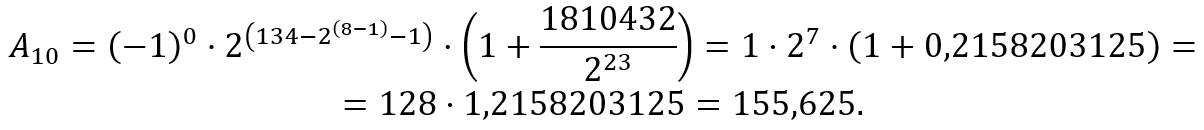
где *A*10 – представление числа в десятичной системе счисления,

*S* – знак всего числа (нуль для положительных, единица для отрицательных),  
      *E* – показатель степени со смещением,  
      *R* – количество разрядов для представления показателя степени *E*,  
      (2(*R*–1) – 1) – заданное смещение, в соответствии с форматом,  
      *M* – дробная часть мантиссы,  
      *m* – количество разрядов для представления мантиссы *M*.

Переведем обратно в десятичную систему счисления число, представленное на рисунке 1.16 в формате одинарной точности.

*S* = 0;   
*E* = 100001102 = 13410;  
         *M* = 001101110100000000000002 = 181043210;  
         *R* = 8;  
         *m* = 23.

Подставив значения, получим:



В результате перевода мы получили верное значение в десятичной системе счисления равное +155,62510.

*Представление особых значений чисел с плавающей точкой*

К особым значениям ВЧсПТ относятся:

* плюс нуль;
* минус нуль;
* плюс бесконечность;
* минус бесконечность;
* неопределенность (не число (*NaN*/*NaNs*)).

***Нуль (со знаком)***

В нормализованной форме ВЧсПТ невозможно представить как однозначный нуль. Для представления нуля зарезервированы специальные комбинации мантиссы и показателя степени.

Число считается нулем, если все его биты (и мантиссы и показателя степени), кроме знакового, равны нулю. При этом в зависимости от значения бита знака нуль может быть как положительным, так и отрицательным (рисунок 1.18).

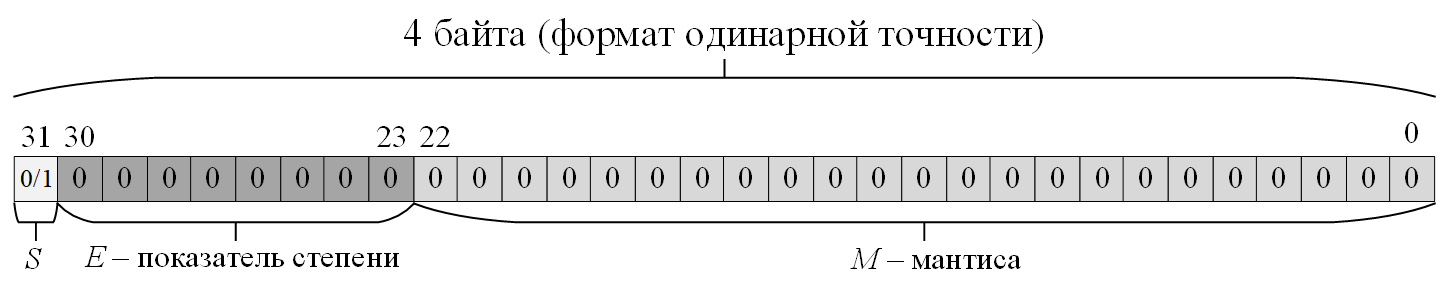
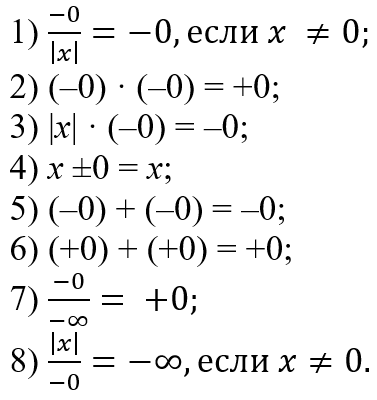


Рисунок 1.18 – Представление нуля по стандарту IEEE 754-2008

Кроме того, для нуля со знаком справедливы следующие правила арифметики:



***Бесконечность (со знаком)***

Вещественное число в формате с плавающей точкой считается равным бесконечности, если все двоичные разряды его показателя степени равны единице, а мантисса равна нулю. Знак бесконечности определяется знаковым битом числа (рисунок 1.19).

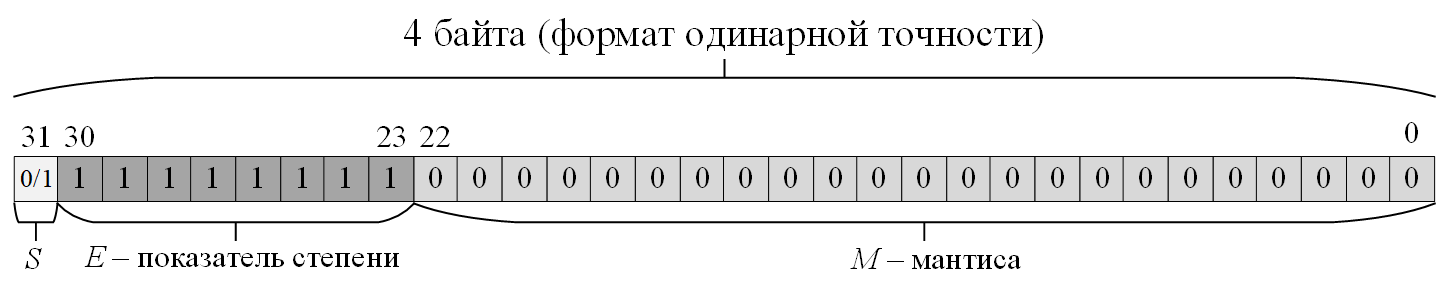
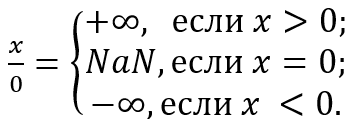


Рисунок 1.19 – Представление бесконечности по стандарту IEEE 754-2008

Получить бесконечность можно при переполнении и при делении ненулевого числа на ноль. При этом:



***Неопределенность (не число (NaN/NaNs))***

Неопределенность является результатом арифметических операций, если во время их выполнения произошла ошибка. Неопределенность также называется «не числом» (от англ. «Not a Number», *NaN*) В стандарте IEEE 754-2008 представлена как число, в котором все двоичные разряды показателя степени единичные, а мантисса не равна нулю (т. е. могут быть различные комбинации разрядов, но все разряды мантисы не должны быть нулевыми).

Исходя из этого одна неопределенность ≠ другая неопределенность.

Представление неопределенности проиллюстрировано на рисунке 1.20.

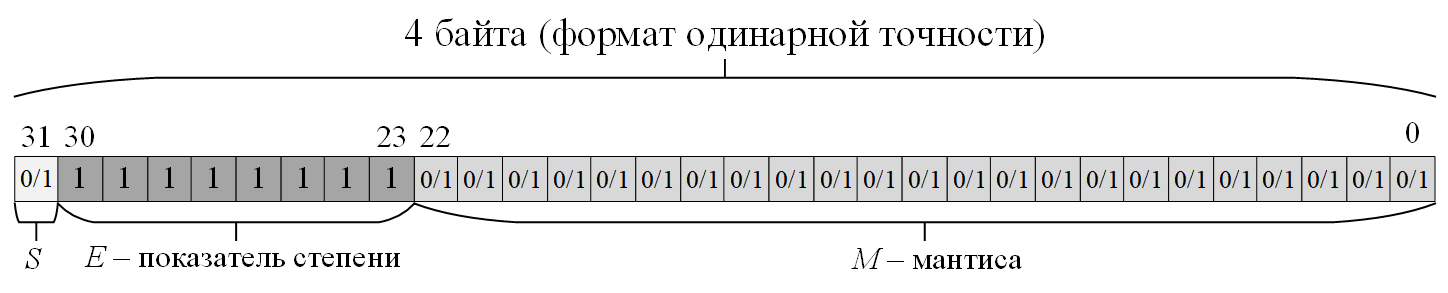
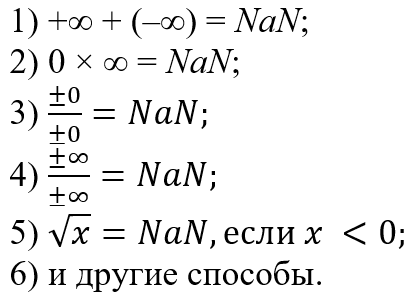


Рисунок 1.20 – Представление неопределенности (*NaN*) по стандарту IEEE 754-2008

Неопределенность (не число/*NaN*) можно получить в результате следующих операций:



Таким образом, полный диапазон представления вещественных чисел с плавающей точкой в формате с одинарной точность можно представить на рисунке 1.21 (значения записаны в формате полубайтов в шестнадцатеричной системе счисления и в формате с плавающей точкой в десятичной системе счисления).

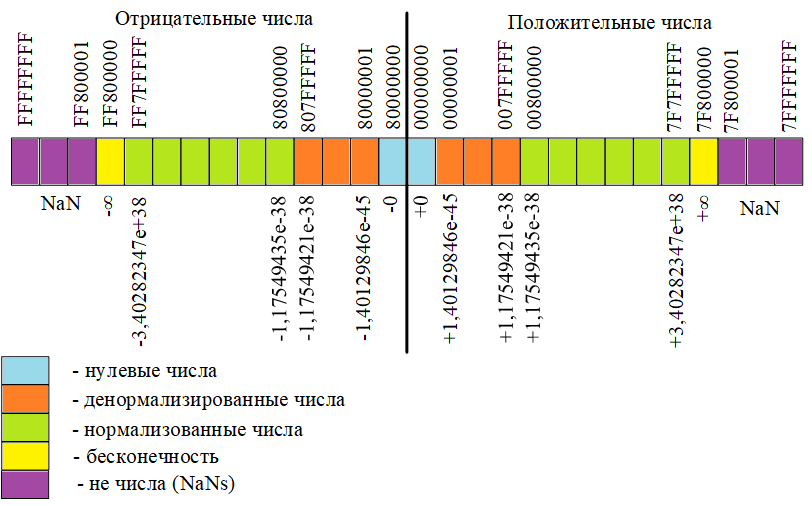


Рисунок 1.21 – Диапазон представления чисел в формате с одинарной точностью

Полный диапазон представления вещественных чисел с плавающей точкой в формате с двойной точность можно представить на рисунке 1.22 (значения записаны в формате полубайтов в шестнадцатеричной системе счисления и в формате с плавающей точкой в десятичной системе счисления).

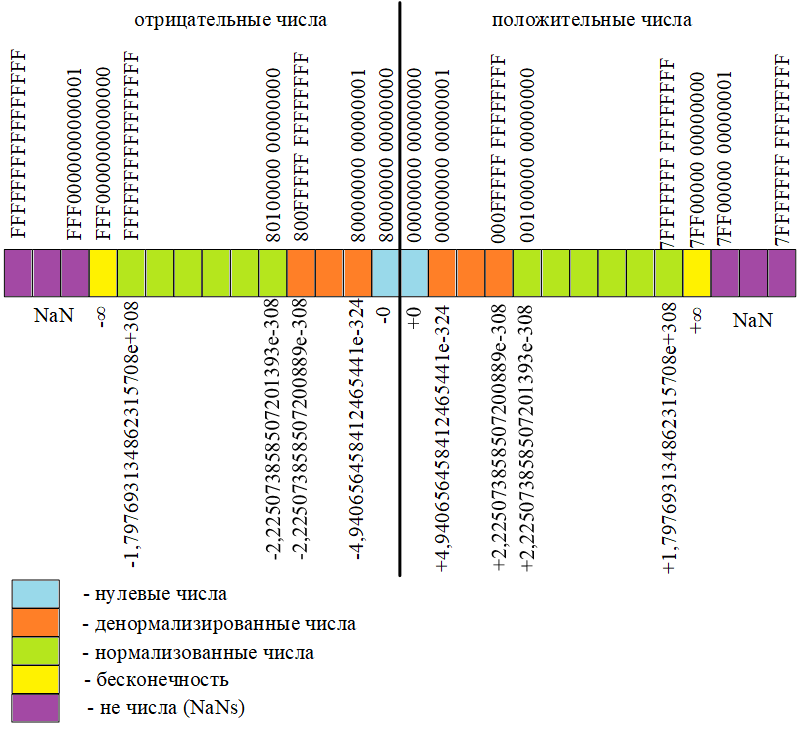


Рисунок 1.22 – Диапазон представления чисел в формате с двойной точностью

На рисунке 1.23 проиллюстрированы представления наименьших и наибольших нормализованных положительных и отрицательных ВЧсПТ.

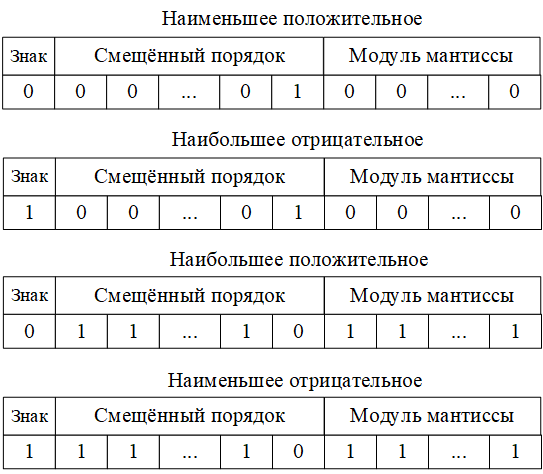


Рисунок 1.23 – Представления знаковых наименьших и наибольших ВЧсПТ

***Точность представления вещественных чисел с плавающей точкой***

Абсолютная максимальная погрешность для ВЧсПТ будет равна в пределе половине шага чисел. При этом шаг чисел будет удваивается с увеличением показателя степени двоичного числа на одну единицу. Таким образом, чем дальше от нуля, тем шире будет шаг чисел по числовой оси и тем больше максимальная ошибка представления. Наглядно это проиллюстрировано на рисунке 1.24.

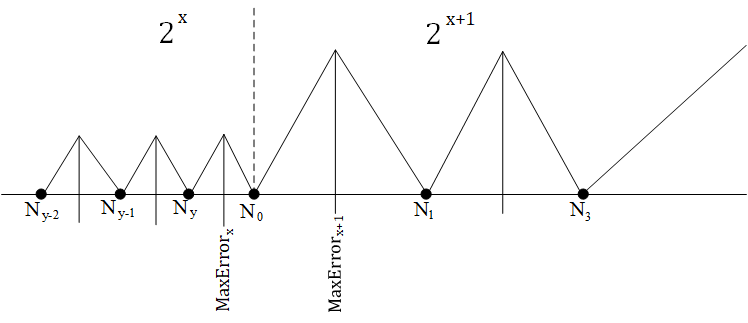


Рисунок 1.24 – Зависимость максимальной абсолютной ошибки от шага чисел и показателя степени

Характеристики ошибок представления нормализованных и денормализованных ВЧсПТ представлены в таблице 1.11.



Из таблицы 1.11 следует, что основная масса ВЧсПТ в форматах IEEE 754-2008 имеет стабильную небольшую относительную погрешность:

Максимально возможная относительная погрешность для числа одинарной точности составляет 2-23 ·100% = 11,920928955078125*e*(-6) %.

Максимально возможная относительная погрешность для числа двойной точности составляет 2-52 · 100% = 2,2204460492503130808472633361816*e*(-14) %.

***Округление вещественных чисел с плавающей точкой***

Стандарт IEEE 754-2008 предусматривает четыре способа округления вещественных чисел с плавающей точкой:

* округление, стремящееся к ближайшему целому;
* округление, стремящееся к нулю;
* округление, стремящееся к +∞;
* округление, стремящееся к –∞.

По умолчанию используется округление к ближайшему целому, однако довольно часто используется и округление к нулю, поскольку при таком округлении необходимо просто отбросить незначащие разряды числа, поэтому он самый легкий с точки зрения аппаратной реализации. Способы округления продемонстрированы в таблице 1.12.

Таблица 1.12 – Примеры округления различными способами

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Исходное десятичное число | Способ округления | | | |
| К ближайшему целому | К нулю | К –∞ | К +∞ |
| 1,33 | 1,3 | 1,3 | 1,4 | 1,3 |
| –1,33 | –1,3 | –1,3 | –1,3 | –1,4 |
| 1,37 | 1,4 | 1,3 | 1,4 | 1,3 |
| –1,37 | –1,4 | –1,3 | –1,3 | –1,4 |
| 1,35 | 1,4 | 1,3 | 1,4 | 1,3 |
| –1,35 | –1,4 | –1,3 | –1,3 | –1,4 |