Володин Владимир Евгеньевич
студент
ФГАОУ ВО «Национальный исследовательский университет
«Московский институт электронной техники»
г. Москва

DOI 10.21661/r-112846

ЧИСЛА С ПЛАВАЮЩЕЙ ТОЧКОЙ КАК ИНСТРУМЕНТ В ПРОЕКТИРОВАНИИ ЦИФРОВЫХ УСТРОЙСТВ

Аннотация: *в статье обосновано применение чисел с плавающей точкой в проектировании цифровых устройств. В работе приведено описание данного представления чисел. Актуальность использования данного представления обусловлена тем, что в современных вычислительных системах происходят вычислительные операции как над очень большими, так и над очень маленькими числами. Так как существует необходимость в представлении и проведении операций на конечном множестве нулей и единиц, в качестве компромисса между скоростью, размером и точностью представления ученые предложили использовать формат чисел с плавающей точкой.*

Ключевые слова: *двоичное число, числа с плавающей точкой, float point, знак, мантисса, основание, порядок, форматы одинарной точности, форматы двойной точности.*

Числа с плавающей запятой (float) соответствуют экспоненциальному представлению чисел. В рамках данного представления преодолены возникающие ограничения в строгом количестве целых и дробных бит числа, а значит, что возможно представление как очень маленьких, так и очень больших чисел. Ровно как и в экспоненциальном представлении числа с плавающей точкой имеют знак, мантиссу, основание и порядок.

$\pm M \times B^e$ – представление числа с плавающей точкой;

M – мантисса;

B – основание;

E – порядок.

Например: число $1.2 * 10^3$ будет являться десятичным экспоненциальным представлением числа 1200. Мантиссой в таком случае будет 1.2, основание 10, порядок 3. У чисел с плавающей точкой основание будет равняться 2, а мантисса и порядок будут представлять собой двоичное число. Существуют различные стандарты представления чисел с плавающей точкой. Рассмотрим 32-х битное представление чисел.

Форматы представления чисел с плавающей точкой

Пример: Дать представление числа 220 в формате числа с плавающей точкой

Для начала преобразуем число и десятичного в двоичное:

$$220_{10} = 11011100_2 = 1.1011100_2 * 2^7$$

Приведем это число к 32-битному представлению:

0 | 00000111 | 10111000000000000000000,

где знаковый бит является положительным и равен 0, в 8 битах порядка находится значение 7 и 23 бита – мантисса.

В двоичных числах с плавающей точкой первый бит мантиссы всегда будет равен 1, поэтому его можно опустить. Этот бит называется неявной старшей единицей, и обычно в представлении она не входит в 23 бита мантиссы.

0 | 00000111 | 10111000000000000000000 – модифицированное представление числа 220 с учетом неявной старшей единицы.

Опуская неявную старшую единицу, мы тем самым освобождаем место под еще один бит данных.

Проведем модификацию для представления порядка числа. Показатель степени может быть как положительным, так и отрицательным. Для этого в формате представления чисел с плавающей точкой используется смещенный порядок, который формируется из исходного порядка плюс постоянное положительное смещение. В 32-х битном представлении чисел с плавающей точкой используется постоянное смещение 127. При данном исходном порядке = 7, итоговый порядок будет следующим: $7 + 127 = 134 = 10000110_2$.

0 | 10000110 | 10111000000000000000000 – представление числа 220 в формате 32-х битного числа с плавающей точкой с неявной старшей единицей и смещенным порядком. Такое представление чисел с плавающей точкой соответствует стандарту IEEE 754.

Стандартом IEEE 754 предусмотрены так же особые случаи для таких чисел, как 0, бесконечность и результаты, которые недопустимы. Например, мы не сможем представить 0 в формате числа с плавающей точкой из-за существования неявной старшей единицы. Для таких «особых» случаев зарезервированы специальные коды, в которых порядок и мантисса состоит только из 0 или только из 1.

Обозначение «особых» случаев в соответствии со стандартом IEEE 754

Number	Sign	Exponent	Fraction
0	X	00000000	00000000000000000000000
∝	0	11111111	00000000000000000000000
±∝	1	11111111	00000000000000000000000
NaN	X	11111111	Non-zero

Рис. 1

Существует множество различных представлений числе с плавающей точкой. Много лет разные производители использовали различные несовместимые форматы. В таком случае результат от одного компьютера не мог быть интерпретирован другим. Проблема была решена введением единого стандарта IEEE 754 Институтом инженеров электротехники и электроники (Institute of Electrical and Electronics Engineers, IEEE) в 1985 году. На данный момент применение этого стандарта повсеместно.

Форматы одинарной и двойной точности

До этого были рассмотрены 32-х битные числа с плавающей точкой. Так же такой формат называют форматом одинарной точности. Стандартом IEEE 754 так же определяются 64-х битные числа с плавающей точкой двойной точности. Они позволяют представить более большой диапазон чисел с гораздо большей точностью. В формате представления чисел с плавающей точкой двойной точности используется 11 бит порядка и 52 бита мантиссы.

Округление двоичных чисел с плавающей точкой

Часто бывает так, что полученные результаты вычислений выходят за пределы доступной нам точности. В таком случае нужно производить округление до наиболее близких чисел к полученному результату. Существует несколько способов округления: округление в большую сторону, округление в меньшую сторону, округление до 0 и округление к ближайшему целому числу. Способом округления по умолчанию принято округление к ближайшему числу. При таком округлении если 2 числа находятся на одинаковом расстоянии, то выбирается то, у которого будет 0 в младшем разряде дробной части. Так же существуют моменты, когда число переполняется если его величина слишком велика для конкретно выбранного представления. Таким же образом число будет исчезающе малым при ситуации, когда оно слишком мало для представления. При округлении исчезающе малые числа округляются до $\pm\infty$, а исчезающе малые числа округляются до 0.

Итоги по стандарту представления чисел с плавающей точкой

Стандарт IEEE 754–1985 определяет:

1. Как осуществлять представление нормализованных положительных и отрицательных чисел с плавающей точкой.

2. Как осуществлять представление денормализованных положительных и отрицательных чисел с плавающей точкой.

3. Как осуществлять представление нулевых чисел.

4. Как осуществлять представление специальной величины бесконечность (Infinity).

5. Как осуществлять представление специальной величины «Не число» (NaN или NaNs).

6. Как производить округление чисел с плавающей точкой (4 типа).

IEEE 754–1985 определяет четыре формата представления чисел с плавающей запятой:

1) с одинарной точностью (single-precision) 32 бита;

2) с двойной точностью (double-precision) 64 бита;

3) с одинарной расширенной точностью (single-extended precision) >=43 бит (редко используемый);

4) с двойной расширенной точностью (double-extended precision) >= 79 бит (обычно используют 80 бит).

Преобразование числа с плавающей точкой в 32 битный формат IEEE 754:

1. Число может быть «+» или «–».

Для этого выделяется 1 бит для обозначения знака числа:

0 – положительное;

1 – отрицательное.

Этот бит будет являться старшим в итоговой 32-х битной последовательности.

2. Далее пойдут биты экспоненты, для их представления выделяют 1 байт (8 бит).

Экспонента может являться, как и числом, со знаком «+», так и числом со знаком «–».

Для того, чтобы определить знак экспоненты, чтобы не прибегать к введению еще одного бита знака, осуществляют добавление к смещению 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 экспонента=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. Остается 23 бита, и они уходят под мантиссу.

У нормализованной двоичной мантиссы первый бит всегда равен 1, так как число лежит в диапазоне 1 < = M < 2. Нет смысла заносить единицу от мантиссы в выделенные 23 бита. В 23 бита мантиссы заносят биты остатка мантиссы, тем самым появляется дополнительный бит для описания большей точности числа.

Список литературы

1. Глушков В.М. Кибернетика, вычислительная техника, информатика. Математические вопросы кибернетики. – 1990.

2. Харрис Д.М. Цифровая схемотехника и архитектура компьютера (перевод) / Д.М. Харрис, С.Л. Харрис. – Morgan Kaufman, 2013.