УДК 004.67

Оценка параметров многоразрядных чисел с плавающей точкой для выполнения операций высокой точности

Бабков В.С., Пехотин Е.В.
Донецкий национальный технический университет
victor.babkov@gmail.com

Abstract

Babkov V., Pehotin E. Estimation of parameters of multidigit floating point numbers for implementation high precision calculations. *In work approach to the estimation of parameters for multidigit numbers in a format with a floating point is offered. A job performance is the system of linked criteria for the selection of effective presentation of multidigit number at implementation of calculations of high precision.*

Введение

Выполнение вычислений с высокой точностью – задача, которая возникает в инженерных и научных расчетах в различных областях [1]: моделирование в физике, химии, астрономии и т.д. На практике наиболее часто возникающей проблемой является выбор между диапазоном представления величин и точностью вычислений (т.е. потерей точности и накоплением ошибок при выполнении вычислительных операций).

Общепринятым стандартом для представления чисел в формате с плавающей точкой и выполнения операций над ними является стандарт IEEE754 [2].

За последние два-три десятка лет в литературе неоднократно поднималась проблема достоверности компьютерных вычислений [3-6]. Несмотря на существование стандарта, есть исследования [7-9], которые показывают, что использование современных подходов к представлению в компьютере действительных чисел приводит к:

- нерациональному расходованию вычислительных ресурсов;
- малой достоверности результатов сложных математических расчетов, особенно с накоплением ошибки;
- плохим соответствием модели представления действительных чисел реальным потребностям различных отраслей знаний.

Особенно серьезными данные проблемы становятся тогда, когда вычисления – это составляющая процесса моделирования сложных, дорогостоящих и опасных объектов в области энергетики, космических исследований, и т.д. В этом случае недостоверность результата вычислений может приводить к аварийным ситуациям, техногенным катастрофам и т.п. Как упоминается в [10], «математическая модель будет адекватной и работоспособной лишь в том случае, если погрешности вычислений контролируются».

В связи с вышесказанным, тема работы, посвященная отысканию путей для определения эффективно–представляющих форматов действительных чисел, является актуальной.

Особенности чисел с плавающей точкой

Рисунок 1 демонстрирует общее представление чисел с плавающей точкой.

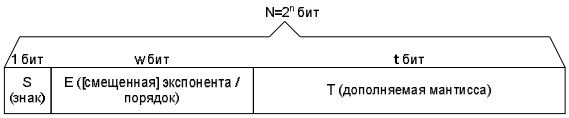

Рисунок 1 – Общий вид числа с плавающей точкой

Обычно разрядность ячейки данных $n = 2^L$, при этом L – кол-во удвоений единичного объема, а также кол-во комбинаций L элементов, имеющих 2 состояния (3:8, 4:16, 5:32, 6:64, 7:128). Также 1+w+t = N.

Для обычной экспоненты значения порядка лежат в диапазоне: $D[E] = [-2^{w-1}; 2^{w-1}-1]$

Обычно используется смещенная экспонента. Пусть:

p_{min}, p_{max} – минимальное и максимальное значения порядка;

w – размерность порядка (кол-во бит под порядок);

bias – прибавляемое смещение.

На рис. 2 показан переход от порядка к характеристике.

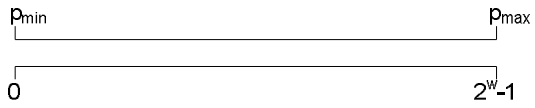

Рисунок 2 – Преобразование порядка в характеристику

Из рисунка 2 можно сделать следующие выводы.

Знаковый диапазон порядка превратился в правильный беззнаковый, т.е. отрицательные числа в обоих случаях меньше положительных, либо $\forall$ a,b $\in$ D[E] & a $\le$ b $\Rightarrow$ Proj[a] $\le$ Proj[b] (Proj – операция проецирования на новый диапазон).

Можно использовать и другие диапазоны порядков, т.к. благодаря смещению все они все равно будут переведены в положительную полуплоскость и, соответственно, представлены на беззнаковом аппарате АЛУ процессора, достаточно лишь, чтобы выполнялось условие $p_{max} - p_{min} = 2^w - 1$ (или $p_{max} - p_{min} + 1 = 2^w$).

Для получения смещения необходимо составить и решить систему уравнений аффинного преобразования одномерного пространства. Т.к. нам известны 2 точки и оба интервала равномерны (что означает, что имеется равенство модулей разностей любых двух соседних элементов любых двух подинтервалов данного интервала: $\forall$ K,S $\subseteq$ D[X] $\forall$ i,j: $|k_{i+1} - k_i| = |s_{j+1} - s_j|$, где X – интервал, либо $\forall$ K $\subseteq$ D[X] $\forall$ i,j: $|k_{i+1} - k_i| = |k_{j+1} - k_j|$), то преобразование будет выполняться с помощью прямой вида

$$x2 = b0 + b1 * x1$$

$$\begin{cases} x2 = b0 + b1 * x1 \\ x2(p_{min}) = 0 \\ x2(p_{max}) = 2^w - 1 \\ p_{max} - p_{min} = 2^w - 1 \end{cases} \Rightarrow \begin{cases} b0 + b1 * p_{min} = 0 \\ b0 + b1 * p_{max} = p_{max} - p_{min} \end{cases} \Rightarrow$$

$$\begin{cases} b0 = -b1 * p_{min} \\ b1(p_{max} - p_{min}) = p_{max} - p_{min} \end{cases} \Rightarrow$$

$$b1 = 1 \Rightarrow b0 = -p_{min} = 2^w - p_{max} - 1 \Rightarrow$$

$$x2 = x1 + bias,\ bias = -p_{min} = 2^w - p_{max} - 1$$

Итак, смещение bias = – p_{min}. Например, для смещения обычной экспоненты необходимо смещение bias = – (-2^{w-1}) = 2^{w-1}. Следует отметить, что при x1 = 0: x2 = x1 + bias = 0 + bias = bias.

Назовем **характеристикой** величину q = p + bias. Будем представлять значение экспоненты числа с плавающей точкой с помощью характеристики.

Рассмотрим один из возможных интервалов значений экспоненты: D[E] = [$1-2^{w-1}$; 2^{w-1}]

Он является зеркализацией обычного интервала, т.е. $p_{min} = -p^g_{max}$, $p_{max} = -p^g_{min}$. Дело в том, что эти интервалы идентичны на промежутке [$1-2^{w-1}$; $2^{w-1}-1$] и отличаются только восприятием числа 2^{w-1}. Это число, у которого в w-1 (самом старшем) бите стоит 1, а в остальных – 0. По правилам знаковых чисел, это число отрицательное и его абсолютное значение равно $2^w - 2^{w-1} = 2^{w-1}$, т.е. представление абсолютного значения числа совпадает с представлением самого числа. Поэтому можно сказать, что это число является своим абсолютным значением и записать его в положительный полуинтервал. В то же время, благодаря использованию смещенной мантиссы это число становится беззнаковым 2^w-1, что позволяет избежать нарушения правил перевода чисел из дополнительного кода в обычный (т.е. получения абсолютных значений знаковых чисел). Благодаря такому подходу удваивается диапазон представления целых чисел.

Утверждение 0. Применение смещения не изменяет свойств интервала.

Доказательство: пускай $\forall$ a,b $\in$ D[E] имеется операция P(a, b). Тогда, если утверждение S применимо к P(a, b), то оно применимо и к P(a’, b’). Т.к. a, b – любые, то мы можем задать a’ = bias + a, b’ = bias + b. ■.

Следует отметить, что в стандарте на кодирование чисел с плавающей точкой IEEE-754 [2] используют именно смещенную экспоненту на интервале [$1-2^{w-1}$; 2^{w-1}].

Мантисса может быть представлена в нормализованном виде 1.0 + ε, где ε < 1.

Постулат 1. Любое число с фиксированной точкой может быть представлено в виде числа с плавающей точкой в данном формате абсолютно без потери точности при условии, что разность между номерами самого старшего и самого младшего единичных бит будет меньше размера мантиссы, а показатель степени самого старшего единичного разряда числа лежит в диапазоне [p_{min}, p_{max}].

Доказательство. Пусть L[M] – размер используемой мантиссы (а через M будем обозначать любую допустимую мантиссу, т.е. любую двоичную комбинацию размером L[M]). Представляем число X в двоичном виде в формате с фиксированной точкой. У любого двоичного числа (кроме 0) будет хотя бы одна значащая 1 в представлении. Пускай номер самого старшего единичного разряда числа при отсчет с 0 от его самого младшего единичного бита равен n (т.е. в числе всего n+1 разрядов), а его показатель степени – z (рисунок 3).

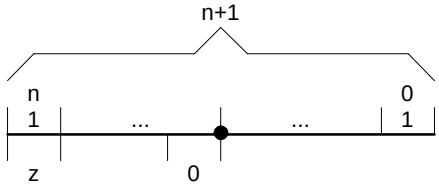

Рисунок 3 – Отсчеты в представляемом числе

Для 0: значение мантиссы возьмем 0.0, значение порядка может быть любое (в том числе и 0).

Для чисел, отличных от 0: номер разряда самой старшей значащей 1 целой части обозначен через z. По условию z ∈ [p_{min}, p_{max}], поэтому может быть допустимым для данного формата порядком представляемого числа. Разделим число на 2^z, или, что то же самое, сдвинем значение представление числа вправо через двоичную точку на z разрядов. Мы получим $m = X/2^z$, у которой целая часть будет равна 1 – старшему единичному биту X, а дробная – остальным битам X. Дополнив m справа нулями до L[M] (по условию их количество не превзойдет длины мантиссы, а т.к. один единичный бит будет у нас всегда, то количество дополняющих нулей будет лежать в интервале [0,L[M] – 1]), получим M – мантиссу числа. Итак,

$$X = \frac{X}{2^z} 2^z = M * 2^p$$

,

где

$M = X/2^z$ – мантисса представляемого числа,

p = z – порядок представляемого числа. ■

Следствие 1. Любое отличное от 0 представимое число может быть представлено в виде 1.0 + ε, где ε < 1.

Т.к. любое число можно представить в формате с плавающей точкой, причем мантисса такого числа может быть выражена в виде 1.0 + ε, то мы можем сэкономить 1 разряд, подразумевая 1.0 и не отображая его в представлении числа. Таким образом, размер мантиссы увеличится на 1 разряд, что приведет к увеличению точности на 1 разряд.

При этом появляется проблема – как представить бесконечности, 0 и ненормализованные числа (которые в данном FP–формате нельзя представить в виде 1.0 + ε, а только 0.0 + ε из-за ограничений хранения порядка). Можно пойти 2 путями:

- хранить полную мантиссу;
- использовать 2 граничных значения порядка, как специальные, и кодировать ими особые случаи.

Теперь необходимо описать форматы 0.0 + ε и 1.0 + ε для дальнейшей работы с ними.

Формат 1.0 + ε, называемый нормализованным, определяет представление числа a в виде $a = (1.0 + e_a) \cdot 2^{p_a}$, при этом $p_a \in [p_{min}, p_{max}]$ – целое, $0 \le e < 1$, при этом ε является мантиссой числа, а 1.0 подразумевается. Постулат 1 определяет числа, представимые в данном формате абсолютно без потери точности.

Формат 0.0 + ε, называемый полным или свободным, определяет представление числа a в виде $a = (0.0 + e_a) \cdot 2^{p_a}$, при этом $p_a \in [p_{min}, p_{max}]$ – целое. Для наложения ограничения на ε необходимо определить положение двоичной точки в представлении числа. Например, мы можем представить число 12.5 (десятичная с/с) в таких видах:

$$\begin{cases} 1250.0 \cdot 10^{-2} \\ 125.0 \cdot 10^{-1} \end{cases} \quad \begin{cases} 12.5 \cdot 10^{0} \\ 1.25 \cdot 10^{1} \end{cases} \quad \begin{cases} 0.125 \cdot 10^{2} \\ 0.0125 \cdot 10^{3} \end{cases}$$

Однако, при использовании левых форм у нас в мантиссе будут лишнии данные – лидирующие нули, при этом из-за них мы будем терять в точности (т.к. лидирующие нули будут требовать места в мантиссе). Поэтому левое представление не подходит для использования в качестве представляющего.

При использовании правого представления мы в числах порядка p_{min} будем терять значимые биты точности (т.к. $1250.0 \cdot 10^{p_{min}} = 1.25 \cdot 10^3 \cdot 10^{p_{min}} = 1.25 \cdot 10^{p_{min}+3}$, т.е. вид числа $1.xx \cdot 10^{yy}$ дает возможность при той же длине порядка представить несколько дополнительных порядков малых чисел).

К верхнему среднему мы можем применить те же рассуждения, взяв, например, 1.25 вместо 12.5, тогда остается только представление в виде $1.xx \cdot 10^{yy}$. Эти же рассуждения справедливы и для двоичной с/с с той поправкой, что вид числа становится $1.xxxx \cdot 2^{yy}$, т.е. в формате 1.0 + ε.

Итак, будем располагать двоичную точку между старшим и предыдущим разрядом мантиссы, т.е. между 2^{t-1} и 2^{t-2} разрядами (отсчет по рисунку 1 с 0 справа налево).

Теперь имеется некоторое различие между полным и свободным форматом 0.0 + ε. Для полного формата обязательно наличие лидирующей единицы в первом разряде, т.е. мантисса числа $e = 1.0 + e^*, 0 \le e^* < 1$ – это формат 1.0 + ε, но только с явно присутствующей в мантиссе лидирующей единицей. У свободного формата условие такого типа отсутствует.

Эффективно-представляющий формат для чисел с плавающей точкой

Теперь можно ввести понятие **эффективно-представляющего формата** – при данный значениях длин порядка $w = w_0$ и мантиссы $t = t_0$ (соответственно, общий размер числа $N = N_0 = 1 + t_0 + w_0$) формат является эффективно-представляющим, если он эффективно использует все w_0 битов порядка и t_0 битов мантиссы, т.е. имеет такое свойство: количество представимых форматом **различных** чисел равно общему числу двоичных чисел разрядностью N_0 (булеану от N_0).

У эффективно–представляющих форматов

имеется исходящее из их определения преимущество перед другими: они представляют максимальное количество различных чисел, которые можно представить имеющейся разрядностью.

Лемма 1. Все числа, представимые в формате 1.0 + ε, является различными или (что то же) любое число, представимое в формате 1.0 + ε, представимо в нем единственным образом.

Доказательство. Т.к. число a представимо в формате 1.0 + ε, то его можно представить в виде $a = (1.0 + e_a) \cdot 2^{p_a}$, причем $0 \le \varepsilon_a < 1.0$, p_a – целое (по определению формата).

Данное разложение является единственным, и его единственность вытекает именно из данного ограничения:

$$0 \le \varepsilon_a < 1.0 \Rightarrow$$

$$1 \le 1.0 + \varepsilon_a < 2 \Rightarrow$$

$$2^{p_a} \le (1.0 + \varepsilon_a) * 2^{p_a} < 2 * 2^{p_a} = 2^{p_a+1} \Rightarrow$$

$$2^{p_a} \le a < 2^{p_a+1} \Rightarrow$$

$$p_a \le \log_2 a < p_a + 1 \Rightarrow$$

$$0 \le \log_2 a - p_a < 1 \Rightarrow$$

пускай

$$\log_2 a = p_a + g \Rightarrow$$

$$0 \le p_a + g - p_a < 1 \Rightarrow$$

$$0 \le g < 1 \Rightarrow \lfloor g \rfloor = 0.$$

Но т.к. p_a – целое (по определению формата), то $\lfloor p_a \rfloor = p_a$ (по определению операции) и

$$\lfloor p_a + g \rfloor = \lfloor p_a \rfloor + \lfloor g \rfloor + \lfloor (p_a - \lfloor p_a \rfloor) + (g - \lfloor g \rfloor) \rfloor =$$

$$= p_a + 0 + \lfloor (p_a - p_a) + (g - 0) \rfloor =$$

$$= p_a + \lfloor 0 + g \rfloor = p_a,$$

с другой стороны $\lfloor p_a + g \rfloor = \lfloor \log_2 a \rfloor$, откуда $p_a = \lfloor \log_2 a \rfloor$ и т.к. a – единственно, то и p_a – единственно (причем эта единственность – односторонняя, т.к. $\nexists\, p_a^* : p_a^* = \lfloor \log_2 a \rfloor \,\&\, p_a^* \ne p_a$, но при этом $\exists a^* : \lfloor \log_2 a \rfloor = \lfloor \log_2 a^* \rfloor$ и $|a - a^*| < 1$.

Значит, мы можем по a единственным образом определить p_a. Однако по p_a мы не можем единственным образом определить a, но можем определить множество A: $A(p_a) = \{a_i \mid p_a \le \log_2 a_i < p_a + 1\}$.

Пусть для данного множества $\log_2 a_i = p_a + g_{a_i}$. Т.к. p_a однозначным образом определяется по a_i ($p_a = \lfloor \log_2 a_i \rfloor$), то и $g_{a_i} = \log_2 a_i - p_a = \log_2 a_i - \lfloor \log_2 a_i \rfloor$ однозначным образом определяется по a_i, а значит, и единственно для данного a_i

Следовательно, для любого a можно однозначно определить p_a и g_a.

Но т.к. $\lfloor \log_2 a \rfloor = \log_2 a - g_a$, то

$$\frac{a}{2^{p_a}} = \frac{a}{2^{\lfloor \log_2 a \rfloor}} = \frac{a}{2^{\log_2 a - g_a}} = \frac{a}{2^{\log_2 a} \cdot 2^{-g_a}} = \frac{a}{a} \cdot 2^{g_a} = 2^{g_a}$$

и т.к. g_a – уникально для данного a, то и 2^{g_a} уникально для данного $a \Rightarrow \frac{a}{2^{p_a}}$ также уникально для данного a. С другой стороны $\frac{a}{2^{p_a}} = 1.0 + e_a \Rightarrow e_a = \frac{a}{2^{p_a}} - 1.0 = 2^{g_a} - 1.0$, что означает, что ε_a также уникально для данного a.

По ε_a можно однозначно установить конкретный элемент множества $A(p_a)$: $e_a = 2^{g_a} - 1.0 \Rightarrow g_a = \log_2(1.0 + e_a)$, а g_a уникален для каждого элемента в пределах множества $A(p_a)$ и определяет этот элемент.

В результате получаем формулы расчета p_a и ε_a:

$$\begin{cases} p_a = \lfloor \log_2 a \rfloor \\ e_a = \frac{a}{2^{p_a}} - 1.0 \end{cases}$$

Итак, имеем p_a, для данного a однозначным образом определяющее множество $A(p_a)$, которое включает в себя a, и ε_a, однозначно определяющее элемент a в данном множестве $A(p_a)$.

При этом и p_a, и ε_a однозначным образом определяются по a, откуда следует, что:

- подтверждено однозначное соответствие $(p_a, e_a) \to a$, вытекающее из определения формата ($a = (1.0 + e_a) * 2^{p_a}$);

- доказано однозначное соответствие $a \to (p_a, e_a)$, тогда $a \leftrightarrow (p_a, e_a)$ и, следовательно, $\forall a \exists!\, p_a \in \mathrm{N}, e_a \in [0;1) : a = (1.0 + e_a) \cdot 2^{p_a}$. ■

Следствие 1. Алгоритм преобразования из Постулата 1 преобразует число a в формат 1.0 + ε единственным образом.

Утверждение 1. Формат 1.0 + ε является эффективно-представляющим.

Доказательство. По лемме 1 все числа формата 1.0 + ε являются различными. Т.к. формат не накладывает никаких ограничений ни на мантиссу, ни на порядок, ни на знак числа, это значит, что:

- для представления дробной части значения числа используются все биты мантиссы, а количество различных значений мантиссы равно булеану от ее длины, т.е. 2^{t_0};

- для представления порядка числа используются все биты порядка, а количество различных значений порядка равно соответственно булеану его длины, т.е. 2^{w_0};

- знак – это элемент множества {+, –}, мощность которого равна 2, при этом каждое представляемое число имеет знак.

Соответственно, всего в формате может быть представлено $2 \cdot 2^{w_0} \cdot 2^{t_0} = 2^{1+w_0+t_0} = 2^{N_0}$ различных чисел. ■

Утверждение 2. Формат 0.0 + ε не является эффективно-представляющим.

Доказательство. Формат 0.0 + ε – это формат полного представления числа, что предполагает наличие двоичной точки в мантиссе между старшим и предыдущим разрядами, что дает ограничение $0 \le \varepsilon < 2$. Действительно, в старшем разряде может быть либо 0, либо 1, а предельные случаи для оставшихся разрядов – это все 0 (левая граница), либо все 1 ($2 - 2^{-t}$). Итак, возможны два случая:

1) ε представляется свободно. Положим $a = 0$. С другой стороны $a = (0.0 + e) \cdot 2^p \Rightarrow 0 = e \cdot 2^p$, но т.к. $2^p \neq 0 \; \forall p \in N$, то ε = 0. С другой стороны, $\forall p \; 0 \cdot 2^p = 0$, т.е. число $a = 0$ можно представить во всех имеющихся порядках, а таких у нас $C_p = p_{\max} - p_{\min} + 1$, и 0 можно представить C_p образами, а не единственно, что нарушает определяющее свойство эффективно-представляющего формата;

2) ε представляется в виде $e = e^* + g$, при этом $e^* \cap g = 0$ – это означает, что некоторые разряды мантиссы задаются фиксированным образом. Здесь имеются два случая – часть чисел все равно продолжает представляться не единственным образом что сводится к случаю 1, либо все числа представимы единственным образом. Пусть $g = 1.0$. По следствию 1 постулата 1 любое представимое число можно представить в виде $1.0 + \varepsilon^*$. Получается представление в формате 1.0 + ε и по лемме 1 в данном представлении любое число представляется единственным образом. Однако, т.к. мы явным образом храним старшую 1 в первом разряде, то на самом деле для представления чисел используются не все t бит мантиссы, а t-1, а 1 разряд занимает 1. В этом случае мы сможем представить булеан не от N_0, а от N_0-1 различных чисел, что противоречит условию ЭП–формата. ■

Итак, в качестве базового формата представления чисел с плавающей точкой рекомендуется использовать эффективно-представляющий формат 1.0 + ε.

Следует отметить, что любое число формата 0.0 + ε может быть представлено в формате 1.0 + ε абсолютно без потери точности и доказательство данного утверждения базируется на инженерной особенности последнего [2].

Система связующих критериев

Далее необходимо составить систему, выражающую зависимость представления чисел с плавающей точкой от размеров порядка и мантиссы представления. Для этого введем определения:

$G_{\min}$ – минимальное представимое форматом значение;

$G_{\min}^{abs}$ – минимально представимое форматом абсолютное значение (т.е. наиболее близкое к 0 представимое значение, отличное от него);

$G_{\max}$ – максимальное представимое форматом значение; $G_{\max}^{abs} \equiv G_{\max}$, естественно, что $G_{\max} = -G_{\min}$;

$G_{\max}^{M}$ – максимально представимая мантисса, одновременно мантисса $G_{\min}$ и $G_{\max}$, из чего следует, что $G_{\max} = G_{\max}^{M} \cdot 2^{p_{\max}}$, $G_{\min} = -G_{\max} = -G_{\max}^{M} \cdot 2^{p_{\max}}$;

$G_{\min}^{M}$ – минимально представимая мантисса, одновременно мантисса $G_{\min}^{abs}$, из чего следует, что $G_{\min}^{abs} = G_{\min}^{M} \cdot 2^{p_{\min}}$;

G_{Δ}^{M} – минимальный мантиссный прирост, т.е. наименьшее возможное изменение мантиссы числа, отражаемое форматом; обычно, $G_{\Delta}^{M} = 2^{-t}$ для формата 1.0 + ε, и $G_{\Delta}^{M} = 2^{-t+1}$ при полном представлении мантиссы, одновременно она обычно является мантиссой $G_{\min}^{abs}$: $G_{\Delta}^{M} = G_{\min}^{M}$;

G_{Δ}^{p} – минимальный прирост чисел порядка p, фактически, являет собою абсолютное значение разницы между двумя соседними представимыми форматом числами (мантиссой данного числа будет G_{Δ}^{M}, а порядком – p $\Rightarrow$ $G_{\Delta}^{p} = G_{\Delta}^{M} \cdot 2^{p}$).

Примем $P_{\Delta}^{p} = G_{\Delta}^{p} \big/ 2$ – точность представления числа порядка p [11].

Можно рассчитать K – количество представимых данным форматом чисел.

В одном порядке мы имеем всего

$$K^{M} = \frac{\left(G_{\max}^{M} - G_{\min}^{M} + G_{\Delta}^{M}\right)}{G_{\Delta}^{M}} = \frac{\left(G_{\max}^{M} - G_{\min}^{M}\right)}{G_{\Delta}^{M}} + 1$$

комбинаций мантисс, т.е. чисел.

Всего комбинаций порядков $K^{E} = p_{\max} - p_{\min} + 1$.

У каждого числа имеется знак – элемент множества S = {+, –}, |S| = 2.

Следовательно всего чисел

$$K = |S| * K^{M} * K^{E} = 2 \cdot \left(\frac{G_{\max}^{M} - G_{\min}^{M}}{G_{\Delta}^{M}} + 1 \right) \cdot \left(p_{\max} - p_{\min} + 1\right) \cdot$$

При этом $G_{\min}^{M}$, $G_{\max}^{M}$ и G_{Δ}^{M} зависят от N –

размерности элемента множества хранения и S – системы отображения числа на множество хранения.

Составим соответствующую систему для общего случая:

$$\begin{cases} G_{\min} = -G^M_{\max} \cdot 2^{p_{\max}} \\ G_{\max} = G^{abs}_{\max} = +G^M_{\max} \cdot 2^{p_{\max}} \\ G^{abs}_{\min} = +G^M_{\min} \cdot 2^{p_{\min}+1} \\ G^M_{\min} = G^M_{\Delta} \\ G^p_{\Delta} = G^M_{\Delta} \cdot 2^p,\ p_{\min} \le p \le p_{\max} \\ P^p_{\Delta} = G^p_{\Delta} \Big/ 2 \\ K = 2 * K^M * K^E = 2 \cdot \left(\frac{G^M_{\max} - G^M_{\min}}{G^M_{\Delta}} + 1 \right) \cdot \left(p_{\max} - p_{\min} + 1 \right) \end{cases}$$

Добавив в нее для конкретного формата зависимости краевых значений порядка и мантиссы от их битовой длины, получим систему зависимостей между краевыми значениями формата и длинами его составляющих. Для формата 1.0 + ε это:

$$\begin{cases} G_{\min} = -\left(1.0 + G^M_{\max}\right) \cdot 2^{p_{\max}} \\ G_{\max} = G^{abs}_{\max} = +\left(1.0 + G^M_{\max}\right) \cdot 2^{p_{\max}} \\ G^{abs}_{\min} = +G^M_{\min} \cdot 2^{p_{\min}+1} \\ G^M_{\min} = G^M_{\Delta} \\ G^p_{\Delta} = G^M_{\Delta} \cdot 2^p,\ p_{\min} \le p \le p_{\max} \\ G^M_{\Delta} = 2^{-t} \\ G^M_{\max} = 1 - G^M_{\Delta} = 1 - 2^{-t} \\ P^p_{\Delta} = G^p_{\Delta} \Big/ 2 \\ K = 2 * K^M * K^E = 2 * \left(\frac{G^M_{\max} - G^M_{\min}}{G^M_{\Delta}} + 1 \right) \left(p_{\max} - p_{\min} + 1 \right) \\ p_{\max} - p_{\min} = 2^w - 1 \\ q = -p_{\min} \\ N = 1 + t + w \end{cases}$$

Здесь q – характеристика формата. В соответствии со стандартом IEEE-754 в эту систему могут быть добавлены еще 2 уравнения:

$$\begin{cases} p_{\min} = 1 - 2^{w-1} \\ p_{\max} = 2^{w-1} \end{cases}.$$

Выводы

В результате исследования получена система, выражающая зависимость граничных значений и точности формата представления 1.0 + ε от длин его составляющих – мантиссы и порядка. Полученные зависимости могут быть использованы на практике для реализации программных библиотек для математических расчетов, ориентированных на получение достоверных результатов с максимально эффективным представлением чисел заданной разрядности.

Литература

1. У. Кулиш, Д. Рац, Р. Хаммер, М. Хокс Достоверные вычисления. Базовые численные методы. Изд-во: «Регулярная и хаотическая динамика», - 1995. – 496 с.
2. IEEE Standard for Floating-Point Arithmetic (Revision of IEEE Std 754-1985) // IEEE Computer Society. – 2008. – P: 70.
3. Wilkinson J. H. Modern error analysis // SIAM Rev. — 1971. — Vol. 13, № 4. — P. 548–568.
4. Moore R. E. Interval analysis. — Englewood Cliffs; Prentice Hall, 1966. — 145 p.
5. Moore R. E. Methods and applications of interval analysis. — Philadelphia; SIAM, 1979. — xi, 190 p.
6. Alefeld G., Herzberger J. Introduction to interval computations. — New York etc.; Academic Press, 1983. — XVIII, 333 p.; Рус. перев.; Алефельд Г., Херцбергер Ю. Введение в интервальные вычисления: Пер. с англ. — М.: Мир, 1987. — 356 с.
7. Добронец Б. С., Шайдуров В. В. Двусторонние численные методы. — Новосибирск; Наука, 1990. — 208 с.
8. Yohe J. M. Portable software for interval arithmetic // Fundamentals of numerical computation (computer-oriented numerical analysis) / Ed.: G. Alefeld, R. D. Grigorieff. — Wien etc.: Springer-Verlag, 1980. — (Computing; Suppl. 2). — P. 211–229.
9. Klatte R., Kulisch U., Wiethoff A., Lawo C., Rauch M. C-XSC. A C++ class library for extended scientific computing. — Berlin etc.: Springer Verlag, 1993. — 270 p.
10. Nickel K. Can we trust the results of our computing? // Mathematics for Computer Science; Proc. Symposium held in Paris, March 16–18, 1982. — S. l.; Association francaise pour la cybernetique et technique (AFCET), 1982. — P. 167–175
11. Верещагин Н.К., Шень А. Математическая логика и теория алгоритмов: Часть 1. Начала теории множеств. – М.: Изд-во МЦНМО, - 2008. – 128 с.