УДК 004.222.3

# ОЦЕНКА ПОГРЕШНОСТИ ПРЕДСТАВЛЕНИЯ ВЕЩЕСТВЕННЫХ ЧИСЕЛ В ПОСТБИНАРНЫХ ФОРМАТАХ С ПЛАВАЮЩЕЙ ЗАПЯТОЙ

*Иваница С.В., Котов Е.И., Аноприенко А.Я.*
*Донецкий национальный технический университет, Украина*

*Рассмотрена точность представления вещественных чисел в бинарных и постбинарных форматах с плавающей запятой и представлены формулы для расчета абсолютной и относительной погрешностей. Проведен ряд экспериментов для точного определения относительной погрешности чисел в зависимости от способа округления и точности постбинарного формата числа с плавающей запятой.*



## Введение

Современные компьютеры и компьютерные системы используют два основных типа данных для хранения и обработки с числовых значений — целочисленные и вещественные. Целочисленный тип данных служит для представления целых чисел. Множество чисел этого типа представляет собой подмножество бесконечного множества целых чисел, ограниченное максимальным и минимальным значениями. Вещественный тип данных позволяет представить действительные числа в форме мантиссы и показателя степени. Такое представление числа определяет последнее как число с плавающей запятой. При этом число с плавающей запятой имеет фиксированную относительную точность и изменяющуюся абсолютную [1], поскольку конечное множество чисел с плавающей запятой отображает бесконечное множество вещественных чисел. Наиболее часто используемое представление таких чисел отражено в стандарте IEEE 754, первая версия которого была утверждена в 1985 году.

До сегодняшнего дня стандарт двоичной арифметики с плавающей запятой IEEE 754 остается актуальным, в виду отсутствия альтернативного и более точного представления вещественных чисел, как в программных реализациях арифметических действий, так и во многих аппаратных реализациях. Однако в 2011 году в работе [2] впервые были описаны принципы кодирования и обработки вещественных данных в постбинарных форматах чисел с плавающей запятой от одинарной до счетверенной (квадро-) точности с целью увеличения точности преставления чисел и повышения надежности вычислений.

Используемый в постбинарных форматах способ кодирования данных основан на принципах кодо-логического базиса [3]: в качестве системы кодирования выступает тетракод *T*, а в качестве единицы хранения одного разряда – тетрит *t*, кодирующий одно из четырех значений: 0 – двоичный ноль, 1 – двоичную единицу, A – значение неопределенности, M – значение множественности:

$$T = \{t\},\ t \in \{0, 1, A, M\}. \quad (1)$$

Такое гибкое кодирование количественных значений позволяет с высокой степенью точности представлять числа в форматах с плавающей запятой.

**Точность представления вещественных чисел в формате с плавающей запятой**

При представлении вещественного числа в виде двоичных полей порядка, мантиссы и знака на вещественной оси можно отложить конечный набор значений, в общем случае не превосходящий $2^{s+l+m}$, где $s$, $l$ и $m$ – разрядность знака, порядка и мантиссы соответственно. Вещественные числа, представленные этим конечным набором, называют базовыми числами, т. е. числами, представленными в формате с плавающей запятой без потери точности. В свою очередь такие числа представляют на вещественной оси конечный набор точек, которые также называют базовыми точками.

На рис. 1 приведена общая схема изменения абсолютной погрешности Δ.

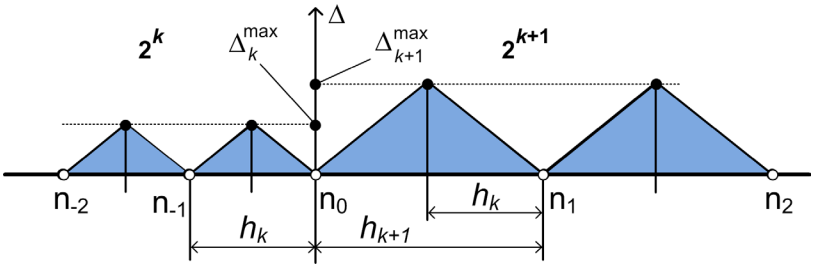


Рисунок 1. Общая схема изменения абсолютной погрешности представления нормализованного числа в формате с плавающей запятой с округлением к ближайшей точке $n_i$

При этом шаг чисел $h$ определяет расстояние между соседними базовыми точками одного порядка. Поэтому, для формата числа с плавающей запятой

$$h_k = 2^{k+off+m}; \tag{2}$$

$$h_{k+1} = 2 \cdot h_k, \tag{3}$$

где $off$ — смещение порядка $k$; $m$ — число разрядов поля мантиссы.

Равенства (2) и (3) справедливы для расчета шага нормализованных чисел (множество денормализованных чисел определено в пределах одного (минимального) значения порядка, и для нахождения шага чисел необходимо также учитывать отсутствие скрытой единицы в поле мантиссы). Из рис. 1 следует, что максимальная абсолютная погрешность числа в формате с плавающей запятой равна в пределе половине шага чисел:

$$\Delta_k^{\max} = \frac{1}{2} h_k = 2^{k+off+m-1}. \tag{4}$$

Учитывая формулы (3) и (4), можно получить зависимость

$$\Delta_{k+1}^{\max} = 2 \cdot \Delta_k^{\max}, \tag{5}$$

Следовательно, максимальная относительная ошибка нормализованного $\delta_k^{\max}$ числа в формате с плавающей запятой равна

$$\delta_k^{\max} = \frac{\Delta_k^{\max}}{|F_k|} \cdot 100\%, \tag{6}$$

где $F_k$ — десятичное число, представленное в формате с плавающей запятой.

Следует отметить, что при округлении к ближайшему числу, формулы (2)–(6)

справедливы и для постбинарных форматов чисел с плавающей запятой, однако в силу использования дополнительных вариантов округления, максимальная абсолютная ошибка для постбинарных форматов может быть уменьшена в 2 раза и достигать в пределе ¼ шага чисел.

**Результаты эксперимента**

Для подтверждения математических расчетов и наглядной демонстрации изменения ошибки представления чисел с плавающей запятой для постбинарных форматов, был проведен эксперимент по определению относительной погрешности для ряда числовых диапазонов с разным шагом. Для этого была создана программная модель, которая представляет исходное число в постбинарных форматах с плавающей запятой различной точности, извлекает число из полученного формата и по разности полученного и исходного значений определяет величину относительной погрешности. Эксперимент проводился с использованием двух видов округления – к нулю и к ближайшему числу.

На рис. 2 и 3 представлены гистограммы значений относительной погрешности для диапазонов чисел в пределах двух базовых точек для постбинарного формата одинарной точности (pbinary32) при различных значениях порядков. В заголовках графиков указаны числовые диапазоны, шаг чисел и количество замеров для данного диапазона. На обеих гистограммах наблюдаются границы скачкообразного увеличения и снижения значений погрешности. Причем, с увеличением порядка числа, граница скачка приближается к базовым точкам.

На рис. 4 представлены гистограммы значений относительной погрешности для диапазонов чисел в пределах трех базовых точек для постбинарного формата pbinary32. При этом базовые числа подобраны на границе увеличения порядка, а изменение погрешности точно соответствует схеме на рис. 1 и формулам (3) и (5).

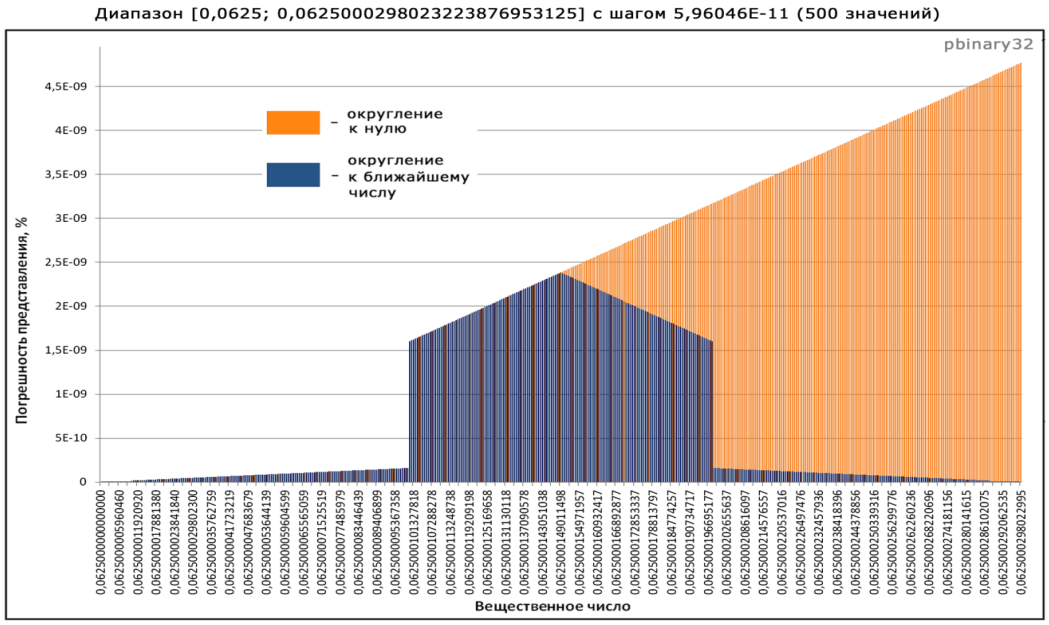


Рисунок 2. Гистограмма относительной ошибки точности представления чисел в постбинарном формате pbinary32 в диапазоне соседних базовых точек (числа порядка $10^{-2}$)

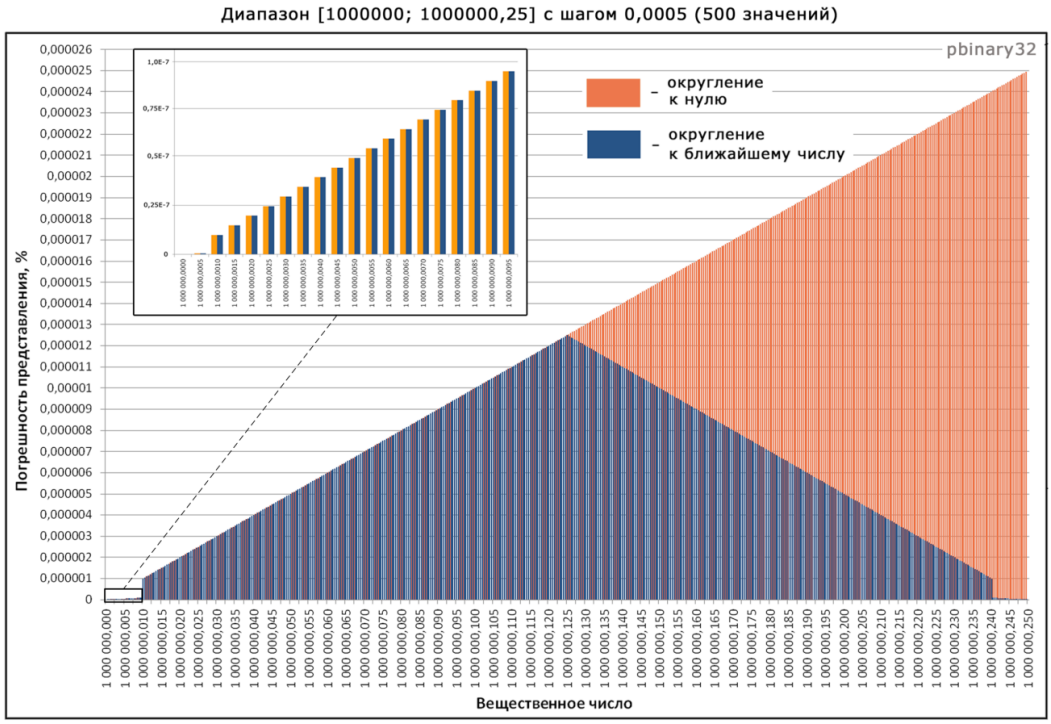


Рисунок 3. Гистограмма относительной ошибки точности представления чисел в постбинарном формате pbinary32 в диапазоне соседних базовых точек (числа порядка $10^{+6}$)

Разница в значениях погрешности при различных округлениях объясняется непосредственно самим принципом округления, т. е. выбором отображения исходного значения одной из приведенной пары базовых точек, между которыми и лежит данное число. При округлении к ближайшему числу, исходное значение представляется, как это следует из названия округления, той базовой точкой, которая находится ближе (отсюда тенденция симметричного изменения погрешности к середине между базовыми точками). При округлении к нулю, исходное значение представляется той базовой точкой, которая находится ближе к нулю (значение погрешности увеличивается пропорционально удалению от меньшей базовой точки). Поэтому в первой половине расстояния между соседними базовыми точками значения погрешностей в обоих вариантах округления совпадают.



## Выводы

В результате проведения эксперимента были подтверждены теоретические расчеты и были получены значения относительной погрешности для различных диапазонов вещественной числовой оси. При этом расчеты производились при различном шаге чисел, в результате изменения которого были получены значения погрешности при различных выборках. Кроме того, погрешность в указанных диапазонах рассчитывалась для чисел, представленных в постбинарных форматах различной точности: от 32-разрядных (одинарная точность) до 256-разрядных (восьмерная точность).

На основании проведенных экспериментов можно выделить следующие наблюдения:

1. При округлении к ближайшему числу погрешность в постбинарных форматах

чисел с плавающей запятой прямо пропорциональна расстоянию к ближайшей меньшей по модулю базовой точке и обратно пропорциональна расстоянию к ближайшей большей по модулю базовой точке.

2. В диапазоне соседних базовых точек наблюдаются границы скачкообразного увеличения и снижения значений погрешности. Экспериментально доказано, что с увеличением порядка числа граница скачка приближается к базовым точкам. При больших числах (рис. 4) границы скачков примыкают к базовым точкам.

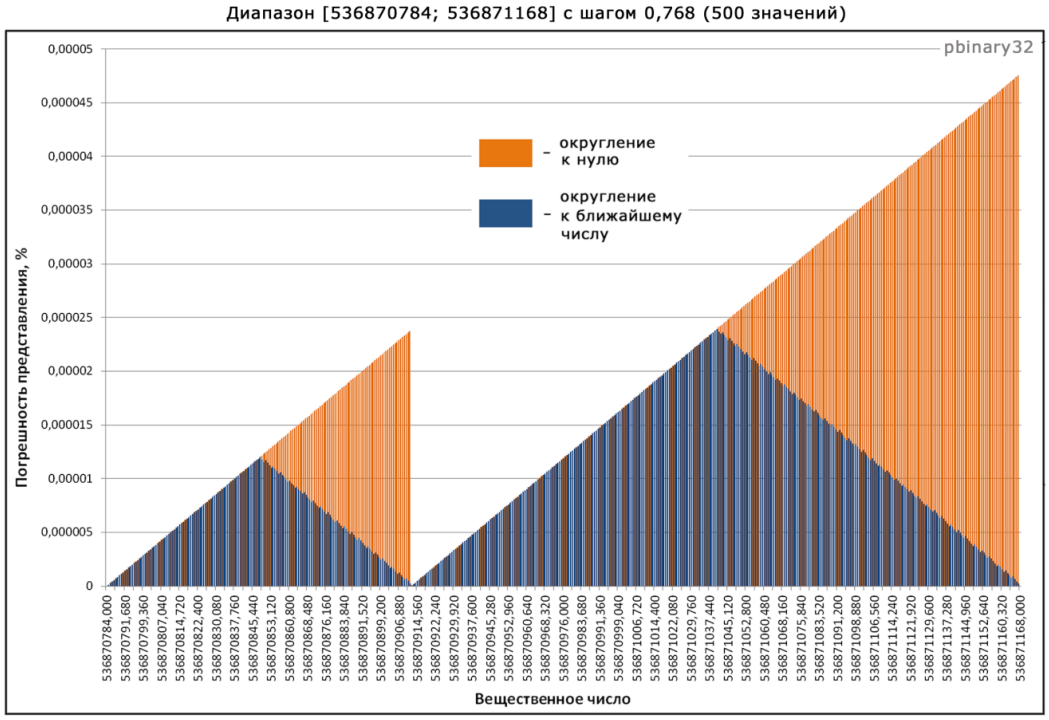


Рисунок 4. Гистограмма относительной ошибки точности представления чисел в постбинарном формате pbinary32 в диапазоне трех базовых точек на границе увеличения порядка

**Список источников**

[1] Число с плавающей запятой. Материал из Википедии – свободной энциклопедии. Электронный ресурс. Режим доступа: http://ru.wikipedia.org/wiki/Вещественный_тип_данных

[2] Аноприенко, А.Я., Гранковский, В.А., Иваница, С.В. Пример Румпа в контексте традиционных, интервальных и постбинарных вычислений. / Наукові праці Донецького національного технічного університету. Серія «Проблеми моделювання та автоматизації проектування» (МАП–2011). Випуск: 9 (179) — Донецьк: ДонНТУ. — 2011. Режим доступа: http://ea.donntu.edu.ua/handle/123456789/1287

[3] Аноприенко, А.Я., Иваница, С.В. Постбинарный компьютинг и интервальные вычисления в контексте кодо-логической эволюции. — Донецк, ДонНТУ, УНИТЕХ, 2011. — 248 с. Режим доступа: http://ea.donntu.edu.ua/handle/123456789/7544