Машинная арифметика

Технологии и языки программирования

Юдинцев В. В.

Кафедра теоретической механики
Самарский университет

28 апреля 2018 г.

САМАРСКИЙ УНИВЕРСИТЕТ
SAMARA UNIVERSITY

Содержание

Стандарт IEEE-754

- IEEE Standard for Binary Floating-Point Arithmetic (ANSI/IEEE Std 754-1985)
- Стандарт разработан в 1985 году ассоциацией IEEE (Institute of Electrical and Electronics Engineers) и используется для представления действительных чисел в двоичном коде.
- Используется в микропроцессорах и программных средствах.
- Особенности стандарта необходимо учитывать при программной реализации численных алгоритмов.

Двоичная запись числа

Двоичная запись целого числа

$$153_{10} = 10011001_2$$

Преобразование целых чисел

$$153/2 = 76 \cdot 2 + 1$$

$$\boxed{\qquad\qquad\qquad 1}$$

Преобразование целых чисел

$$153/2 = 76 \cdot 2 + 1$$
$$76/2 = 38 \cdot 2 + 0$$

$\boxed{0\ 1}$

Преобразование целых чисел

$$153/2 = 76 \cdot 2 + 1$$
$$76/2 = 38 \cdot 2 + 0$$
$$38/2 = 19 \cdot 2 + 0$$

$\boxed{0\;0\;1}$

Преобразование целых чисел

$$
\begin{aligned}
153/2 &= 76 \cdot 2 + 1 \\
76/2 &= 38 \cdot 2 + 0 \\
38/2 &= 19 \cdot 2 + 0 \\
19/2 &= 9 \cdot 2 + 1
\end{aligned}
$$

$\boxed{\quad 1\ 0\ 0\ 1}$

Преобразование целых чисел

$$153/2 = 76 \cdot 2 + 1$$
$$76/2 = 38 \cdot 2 + 0$$
$$38/2 = 19 \cdot 2 + 0$$
$$19/2 = 9 \cdot 2 + 1$$
$$9/2 = 4 \cdot 2 + 1$$

$$\boxed{1\,1\,0\,0\,1}$$

Преобразование целых чисел

$$\begin{aligned}
153/2 &= 76 \cdot 2 + 1 \\
76/2 &= 38 \cdot 2 + 0 \\
38/2 &= 19 \cdot 2 + 0 \\
19/2 &= 9 \cdot 2 + 1 \\
9/2 &= 4 \cdot 2 + 1 \\
4/2 &= 2 \cdot 2 + 0
\end{aligned}$$

$$\boxed{0\ 1\ 1\ 0\ 0\ 1}$$

Преобразование целых чисел

$$
\begin{aligned}
153/2 &= 76 \cdot 2 + 1 \\
76/2 &= 38 \cdot 2 + 0 \\
38/2 &= 19 \cdot 2 + 0 \\
19/2 &= 9 \cdot 2 + 1 \\
9/2 &= 4 \cdot 2 + 1 \\
4/2 &= 2 \cdot 2 + 0 \\
2/2 &= 1 \cdot 2 + 0 \\
1/2 &= 0 \cdot 2 + 1
\end{aligned}
$$

$$\boxed{0\ 0\ 1\ 1\ 0\ 0\ 1}$$

Преобразование целых чисел

$$\begin{aligned}
153/2 &= 76 \cdot 2 + 1\\
76/2 &= 38 \cdot 2 + 0\\
38/2 &= 19 \cdot 2 + 0\\
19/2 &= 9 \cdot 2 + 1\\
9/2 &= 4 \cdot 2 + 1\\
4/2 &= 2 \cdot 2 + 0\\
2/2 &= 1 \cdot 2 + 0\\
1/2 &= 0 \cdot 2 + 1
\end{aligned}$$

$\boxed{1\ 0\ 0\ 1\ 1\ 0\ 0\ 1}$

$$153_{10} = 10011001_2$$

Обратное преобразование

1	0	0	1	1	0	0	1
$\cdot 2^7$	$\cdot 2^6$	$\cdot 2^5$	$\cdot 2^4$	$\cdot 2^3$	$\cdot 2^2$	$\cdot 2^1$	$\cdot 2^0$

$$2^7 + 2^4 + 2^3 + 2^0 = 128 + 16 + 8 + 1 = 153$$

Преобразование дробных чисел

$$0.235 \cdot 2 = 0.470$$

$$\boxed{.0}$$

Преобразование дробных чисел

$$0.235 \cdot 2 = 0.470$$

$$0.470 \cdot 2 = 0.940$$

$$\boxed{.0\ 0}$$

Преобразование дробных чисел

$$0.235 \cdot 2 = 0.470$$

$$0.470 \cdot 2 = 0.940$$

$$0.940 \cdot 2 = 1.880$$

$$\boxed{.0\;0\;1}$$

Преобразование дробных чисел

$$0.235 \cdot 2 = 0.470$$

$$0.470 \cdot 2 = 0.940$$

$$0.940 \cdot 2 = 1.880$$

$$0.880 \cdot 2 = 1.760$$

$$\boxed{.0\;0\;1\;1}$$

Преобразование дробных чисел

$$0.235 \cdot 2 = 0.470$$
$$0.470 \cdot 2 = 0.940$$
$$0.940 \cdot 2 = 1.880$$
$$0.880 \cdot 2 = 1.760$$
$$0.760 \cdot 2 = 1.520$$

$$\boxed{.0\ 0\ 1\ 1\ 1}$$

Преобразование дробных чисел

$$0.235 \cdot 2 = 0.470$$
$$0.470 \cdot 2 = 0.940$$
$$0.940 \cdot 2 = 1.880$$
$$0.880 \cdot 2 = 1.760$$
$$0.760 \cdot 2 = 1.520$$
$$0.520 \cdot 2 = 1.040$$

$$\boxed{.0\,0\,1\,1\,1\,1}$$

Преобразование дробных чисел

$$0.235 \cdot 2 = 0.470$$
$$0.470 \cdot 2 = 0.940$$
$$0.940 \cdot 2 = 1.880$$
$$0.880 \cdot 2 = 1.760$$
$$0.760 \cdot 2 = 1.520$$
$$0.520 \cdot 2 = 1.040$$
$$0.040 \cdot 2 = 0.080$$

$$\boxed{.0\ 0\ 1\ 1\ 1\ 1\ 0}$$

Преобразование дробных чисел

$$0.235 \cdot 2 = 0.470$$
$$0.470 \cdot 2 = 0.940$$
$$0.940 \cdot 2 = 1.880$$
$$0.880 \cdot 2 = 1.760$$
$$0.760 \cdot 2 = 1.520$$
$$0.520 \cdot 2 = 1.040$$
$$0.040 \cdot 2 = 0.080$$
$$0.080 \cdot 2 = 0.160$$

$$\boxed{.0\,0\,1\,1\,1\,1\,0\,0}$$

Преобразование дробных чисел

$$0.235 \cdot 2 = 0.470$$
$$0.470 \cdot 2 = 0.940$$
$$0.940 \cdot 2 = 1.880$$
$$0.880 \cdot 2 = 1.760$$
$$0.760 \cdot 2 = 1.520$$
$$0.520 \cdot 2 = 1.040$$
$$0.040 \cdot 2 = 0.080$$
$$0.080 \cdot 2 = 0.160$$
$$0.160 \cdot 2 = 0.320$$

$$\ldots$$

$$\boxed{.0\ 0\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ \ldots}$$

$$0.235_{10} \approx 0.001111000_2$$

Обратное преобразование

.0	0	1	1	1	1	0	0	0
$\cdot 2^{-1}$	$\cdot 2^{-2}$	$\cdot 2^{-3}$	$\cdot 2^{-4}$	$\cdot 2^{-5}$	$\cdot 2^{-6}$	$\cdot 2^{-7}$	$\cdot 2^{-8}$	$\cdot 2^{-9}$

$$2^{-3} + 2^{-4} + 2^{-5} + 2^{-6} =$$

$$0.125 + 0.0625 + 0.03125 + 0.015625$$

$$= 0.234375 \approx 0.235$$

Нормализованная форма записи

Запись числа в формате с плавающей точкой

Варианты записи числа 1251 в форме с плавающей точкой

$$
\begin{aligned}
1251 &= 0.1251 \times 10^4 \\
&= 1.2510 \times 10^3 \\
&= 12.510 \times 10^2 \\
&= 125.10 \times 10^1
\end{aligned}
$$

Нормализованная форма b=10

Нормализованная десятичная форма числа 1251:

$$1251 = 1.251 \times 10^3$$

Число состоит из двух частей:

- мантисса: $\boxed{1.251}$
- показатель степени: $\boxed{+3}$

Модуль мантиссы нормализованного десятичного числа меньше 10

$$1 \leq |M| < 10$$

Денормализованная форма b=10

Денормализованная десятичная форма числа 1251

$$1251 = 0.1251 \times 10^{4}$$

- мантисса: $\boxed{0.1251}$
- показатель степени: $\boxed{+4}$

Модуль мантиссы денормализованного десятичного числа

меньше 1

$$0 \leq |M| < 1$$

Нормализованная форма b=2

Нормализованная двоичная форма числа 12.125:

$$12.125 = 1100.001_2 = 1.100001 \cdot 2^3$$

- мантисса: $\boxed{1.100001}$
- показатель степени: $\boxed{+3}$

Денормализованная форма b=2

Денормализованная двоичная форма числа 12.125:

$$12.125 = 1100.001_2 = 0.1100001 \cdot 2^4$$

- мантисса: $\boxed{0.1100001}$
- показатель степени: $\boxed{+4}$

Стандарт IEEE-754

Форматы чисел стандарта IEEE-754

- **числа одинарной точности**
  `single precision` – 32 бита
- **числа двойной точности**
  `double precision` – 64 бита
- числа расширенной одинарной точности
  `single extended precision` – $\geq$ 43 бита
- числа расширенной двойной точности
  `double extended precision` – $\geq$ 79 бит
- Для записи чисел используется форма с плавающей точкой.

Число одинарной точности (single precision)

S. Знак	*Смещ. показатель* E_s	*M. Мантисса*
1 бит	8 бит	23 бита

Для записи числа выделяется 4 байта:

1. 1 старший бит — знак числа (0 или 1)
2. 23 бита для мантиссы без первой (старшей) единицы
3. 8 бит для смещенного на $(2^8)/2 - 1 = 127$ показателя степени

$$E_s = E + 127 > 0 \Rightarrow E_{min} = -126.$$

Смещение позволяет не вводить знаковый бит показателя степени

Число одинарной точности

$$12.125 = \boxed{1.2125 \cdot 10^1} = 1100.001 = \boxed{1.100001 \cdot 2^3}$$

- Мантисса: 1.100001 (без первой единицы)
- Показатель степени: 3
- Смещённый показатель степени: 3 + 127 = 130 = 10000010

0	10000010	10000100000000000000000

Предельные нормализованные числа (single)

Максимальное нормализованное число

- смещённый показатель степени 1111 1110 = 255 - 1 = 254;
- показатель степени 254 - 127 = 127;
- мантисса $1.11111111111111111111111 = 2 - 2^{23}$

$$(2 - 2^{23}) \cdot 2^{127} = 3.402823466385286 \cdot 10^{38}$$

Минимальное нормализованное число

- смещённый показатель степени 0000 0001 = 1;
- показатель степени 1 - 127 = -126;
- мантисса $1.00000000000000000000000 = 1$

$$1 \cdot 2^{-126} = 1.175494350822875 \cdot 10^{-38}$$

Число двойной точности

- Для записи числа выделяется 8 байт
- Смещение: $2^{11}/2 - 1 = 1023$.

$$E_s = E + 1023$$

S. Знак	E_s. *Смещ. показатель*	*M. Мантисса*
1 бит	11 бит	52 бит

Числа, близкие к нулю

$$12.125 = \boxed{1.2125 \cdot 10^1} = 1100.001 = \boxed{1.100001 \cdot 2^3}$$

- Нормализованная форма записи не позволяет представить ноль, т.к. при восстановлении числа мантисса всегда дополняется единицей:

$$\boxed{0}\ \boxed{10000010}\ \boxed{10000100000000000000000}$$

$$1.100001 \cdot 2^{10000010-1111111}$$

- Для записи нуля и близких к нулю чисел используется **денормализованная форма**

Денормализованное число

Для отличия денормализованных чисел от нормализованных, биты показателя степени денормализованного числа заполняются нулями

Нормализованное число

0/1	$0 < E_s <$ 11111…11	любое число
Знак	Смещенный показатель	Мантисса

Денормализованное число

0/1	0000…0000	не ноль
Знак	Смещенный показатель	Мантисса

Предельные значения (single)

Максимальное денормализованное число

- показатель степени -126;
- мантисса $0.11111111111111111111111 = 1 - 2^{-23}$

$$(1 - 2^{23}) \cdot 2^{-126} = 1.1754942106924411 \cdot 10^{-38}$$

Минимальное денормализованное число

- показатель степени = -126;
- мантисса $0.00000000000000000000001 = 2^{-23}$

$$2^{-23} \cdot 2^{-126} = 1.401298464324817 \cdot 10^{-45}$$

Специальные значения

“Положительный” ноль

0	0000...0000	0000...0000
Знак	Смещенный показатель	Мантисса

“Отрицательный” ноль

1	0000...0000	0000...0000
Знак	Смещенный показатель	Мантисса

Специальные значения

Плюс бесконечность $+\infty$

Знак	Смещенный показатель	Мантисса
0	1111...1111	0000...0000

Минус бесконечность $-\infty$

Знак	Смещенный показатель	Мантисса
1	1111...1111	0000...0000

Не число (NaN)

Знак	Смещенный показатель	Мантисса
0/1	1111...1111	не ноль

Восстановление нормализованного числа

$$\boxed{F = (-1)^s 2^{\left(E_s - 2^{(b-1)} + 1\right)} \left(1 + M/2^n\right)}$$

- b – количество бит, отводимых под показатель степени;
- E_s – смещенный показатель степени;
- M – остаток мантиссы;
- n – количество бит, отводимых под мантиссу.

Восстановление нормализованного числа

single precision

$$\boxed{F = (-1)^s 2^{E_s - 127}(1 + M/2^{23})}$$

double precision

$$\boxed{F = (-1)^s 2^{E_s - 1023}(1 + M/2^{52})}$$

Восстановление денормализованного числа

$$\boxed{F = (-1)^s 2^{\left(E_s - 2^{(b-1)} + 2\right)} \left(M/2^n\right)}$$

- b – количество бит, отводимых под показатель степени;
- E_s – смещенный показатель степени;
- M – остаток мантиссы;
- n – количество бит, отводимых под мантиссу.

Погрешности представления чисел

Точность представления числа

- В ЭВМ представимы лишь конечный набор рациональных чисел.
- Эти числа образуют представимое множество вычислительной машины.
- Для всех остальных чисел возможно лишь их приближенное представление с ошибкой, которую принято называть **ошибкой представления (ошибкой округления)**.

Пример непредставимого числа

Число 1/10 невозможно точно представить в двоичной системе

$$0.1_2 = 0.00011001100110011001 1\ldots$$

В десятичной системе подобным числом является $1/3$

$$\left(\frac{1}{3}\right)_{10} = 0.333333\ldots$$

Точность числа в стандарте IEEE-754

- Абсолютная максимальная ошибка для числа в формате IEEE-754 равна в пределе половине шага чисел.
- Шаг чисел удваивается с увеличением показателя степени двоичного числа на единицу.
- Чем дальше от нуля, тем шире шаг чисел в формате IEEE754 по числовой оси.

Абсолютная погрешность

Предел максимальной абсолютной ошибки будет равен 1/2 шага числа:

- `single` : $A(x^*) = 2^{E_s-23-127}/2 = 2^{(E_s-151)}$
- `double` : $A(x^*) = 2^{E_s-52-1023}/2 = 2^{(E_s-1076)}$

Относительная погрешность

- Относительная погрешность нормализованного числа

$$\boxed{\Delta(x^*) = \frac{2^{E-151}}{2^{E-127}\left(1 + \frac{M}{2^{23}}\right)} = \frac{1}{2^{24} + 2M}}$$

- Относительная погрешность денормализованного числа

$$\boxed{\Delta(x^*) = \frac{2^{E-150}}{2^{E-126}\frac{M}{2^{23}}} = \frac{1}{2M}}$$

Погрешность чисел одинарной точности

x^*	Абсолютная погрешность
$2^{-149} \approx 1.401298 \times 10^{-45}$	$2^{-150} \approx 0.700649 \times 10^{-45}$
$2^{-148} \approx 2.802597 \times 10^{-45}$	$2^{-150} \approx 0.700649 \times 10^{-45}$
1.0	$2^{-23} \approx 1.192 \times 10^{-7}$
100	$2^{-17} \approx 7.6294 \times 10^{-6}$
1.0×10^{10}	$2^{10} \approx 1.024 \times 10^{3}$

Погрешность чисел двойной точности

x^*	Абсолютная погрешность
$2^{-1074} \approx 4.940656 \times 10^{-324}$	$2^{-1075} \approx 2.470328 \times 10^{-324}$
$2^{-1073} \approx 9.881313 \times 10^{-324}$	$2^{-1075} \approx 2.470328 \times 10^{-324}$
1.0	$2^{-52} \approx 2.220446 \times 10^{-16}$
100	$2^{-46} \approx 1.421085 \times 10^{-14}$
1.0×10^{10}	$2^{280} \approx 1.942669 \times 10^{84}$

Особенности машинной арифметики

Сложение и вычитание чисел

- Значащих цифр в мантиссе двоичного числа в формате `single` не более 24.
- Если числа отличаются более чем в 2^{23} (для `single`) и 2^{52} (для `double`), то операции сложения и вычитания между этими числами невозможны.

$300 =$

$\boxed{1.00101100000000000000000} \times 2^8$

$0.00001 =$

$\boxed{0.00000000000000000000000010100111110001011011} \times 2^8$

$300 + 0.00001 = 300$

Пример

```
a = 100.0

b = a + 1e-14

b
```

100.00000000000001

```
a = 100.0

b = a + 1e-15

b
```

100.0

“Бесконечный” цикл

Выполнится ли строка 5?

```
1 x = 1.0
2 while (x != x + 1):
3     x = 2 * x
4
5 print(x)
```

“Бесконечный” цикл

```python
x = 1.0
while (x != x + 1):
    x = 2 * x

print(x)
```

9 007 199 254 740 992.0

“Бесконечный” цикл

```
x = 1.0
while (x != x + 0.001):
    x = 2 * x

print(x)
```

17 592 186 044 416.0

Ассоциативность

$$(a + b) + c = a + (b + c)$$

Нарушение свойства ассоциативности операции сложения (вычитания)

$$(10^{20} + 1) - 10^{20} = 0 \neq (10^{20} - 10^{20}) + 1 = 1$$

```
(10e20 + 1) - 10e20
```

0.0

```
(10e20 - 10e20) + 1
```

1.0

Сравнение чисел

```
a = 0.1
b = 0.1
b = b + 10
b = b - 10

a==b

False

a
0.1
b
0.09999999999999964

a-b
3.608224830031759e-16
```

Сравнение чисел

Функция isclose

```
math.isclose(a,b,*,rel_tol=1e-09,abs_tol=0.0)
```

используется для проверки на равенство с заданной точностью двух вещественных чисел.

- Результат работы функции `True` или `False`
- Если относительная или абсолютная погрешность чисел меньше заданной величины, то числа считаются равными, т.е. результат работы функции `True`, если выполняется неравенство:

$$|a-b| \leq \mathsf{max}(rel\_tol \cdot \mathsf{max}(|a|,|b|), abs\_tol)$$

Источники

1. IEEE 754-2008
   `https://ru.wikipedia.org/wiki/IEEE_754-2008`
2. Волков Е. А. Численные методы: Учебное пособие для вузов. 2-е изд., испр. М.: Наука. 1987.
3. IEEE 754 - стандарт двоичной арифметики с плавающей точкой
   `http://www.softelectro.ru/ieee754.html`
4. Что нужно знать про арифметику с плавающей запятой
   `https://habrahabr.ru/post/112953`