Тема 4. *Формы представления двоичных чисел*

Двоичные числа могут быть представлены в двух формах:

в форме с **фиксированной точкой**

в форме с **плавающей точкой**.

Рассмотрим сначала представление двоичных чисел с *фиксированной точкой.*

- **<u>Представление целых чисел без знака</u>.** Формат целых чисел без знака выглядит следующим образом:

разряд: (n-1) 1 0

b_{n-1}	. . .	b_1	b_0

• десятичная точка

Старшим является разряд (n-1), а младшим разряд 0. Десятичная точка неявно фиксирована справа (заметим, что эта точка физически не представляется). Некоторое число N представленное в данном формате находится в интервале

$$1 \le N \le 2^n-1.$$

Например, 8-разрядное двоичное число может иметь значения между 1 и 255.

Примеры:

		128	64	32	16	8	4	2	1	
		7	**6**	**5**	**4**	**3**	**2**	**1**	**0**	.
	38	0	0	1	0	0	1	1	0	.
max	**255**	1	1	1	1	1	1	1	1	.
min	**1**	0	0	0	0	0	0	0	1	.

- **<u>Представление дробных (меньше единицы) чисел без знака</u>.** В этом случае, формат чисел аналогичен формату целых чисел без знака. Отличие состоит в том, что десятичная точка неявно фиксирована слева.

разряд: -1 -2 -m

b_{-1}	b_{-2}	. . .	b_{-m}

• десятичная точка

Двоичное дробное число без знака находится в интервале

$$2^{-m} \le N \le 1-2^{-m}.$$

Представление двоичных чисел со знаком

Представление двоичных чисел со знаком предполагает использование некоторого соглашения относительно цифры знака. Согласно принятой конвенции знак “+” представляется цифрой 0 а знак “-” цифрой 1. Двоичные числа со знаком могут быть представлены в прямом, дополнительном и обратном коде.

- **<u>Представление двоичных чисел в прямом коде</u>.** Целые числа со знаком имеют следующий формат:

разряд: n (n-1) 1 0

Sg	b_{n-1}	. . .	b_1	b_0

• десятичная точка

где Sg изображает знак числа. Прямой код числа со знаком можно записать в виде:

$$N_{cd} = \begin{cases} 0\, b_{n-1} b_{n-2} \ldots b_1 b_0; & N \geq 0 \\ 1\, b_{n-1} b_{n-2} \ldots b_1 b_0; & N < 0 \end{cases}$$

Приведем пример представления в прямом коде двоичных чисел со знаком, длина которых 8 бит (n=7):

		Знак	64	32	16	8	4	2	1
		7	**6**	**5**	**4**	**3**	**2**	**1**	**0**
	38	0	0	1	0	0	1	1	0
	-38	1	0	1	0	0	1	1	0
max	**127**	0	1	1	1	1	1	1	1
min	-127	1	1	1	1	1	1	1	1
	0+	0	0	0	0	0	0	0	0
	0-	1	0	0	0	0	0	0	0

Диапазон значений целого числа в прямом коде определяется как

$$-(2^n-1) \leq N_{cd} \leq 2^n-1.$$

Двоичные дробные числа со знаком имеют формат:

0	-1	-2		-m
Sg	b_{-1}	b_{-2}	. . .	b_{-m}

Значение дробного числа в прямом коде удовлетворяет неравенству:

$$-(1-2^{-m}) \leq N_{cd} \leq 1-2^{-m}.$$

<u>**Представление двоичных чисел в обратном коде:**</u> Обратный код положительного числа совпадает с прямым кодом этого числа. Обратный код отрицательного числа получается путем инвертирования цифровых разрядов и проставляя 1 в знаковый разряд.

$$N_{cc} = \begin{cases} 0\, b_{n-1} b_{n-2} \ldots b_1 b_0; & N \geq 0 \\ 1\, \overline{b}_{n-1} \overline{b}_{n-2} \ldots \overline{b}_1 \overline{b}_0; & N < 0 \end{cases}$$

Диапазон значений целых чисел в обратном коде задается в виде

$$-(2^n-1) \leq N_{CI} \leq 2^n-1.$$

Например, значения 8-разрядного целого числа будут находиться в интервале [-127; +127].

			64	32	16	8	4	2	1
		7	**6**	**5**	**4**	**3**	**2**	**1**	**0**
	38	0	0	1	0	0	1	1	0
	-38	1	1	0	1	1	0	0	1
max	**127**	0	1	1	1	1	1	1	1
min	-127	1	0	0	0	0	0	0	0
	0+	0	0	0	0	0	0	0	0
	0-	1	1	1	1	1	1	1	1

В случае дробных чисел, диапазон значений задается неравенством:

$$-(1-2^{-m}) \le N_{CI} \le 1-2^{-m}.$$

* **<u>Представление двоичных чисел в дополнительном коде</u>.** Дополнительный код положительного числа совпадает с прямым кодом этого числа. В случае отрицательных чисел правило получения дополнительного кода следующее: инвертируются все разряды с последующим прибавлением 1 в младший разряд, а в поле цифры знака проставляется 1. Следовательно, можно записать

$$N_{cc} = \begin{cases} 0\, b_{n-1}\, b_{n-2} \ldots b_1\, b_0\, ; \ \ N \ge 0 \\ 1\, \bar{b}_{n-1}\, \bar{b}_{n-2} \ldots \bar{b}_1\, (\bar{b}_0 + 1)\, ; \ \ N < 0 \end{cases}$$

Проиллюстрируем как записываются в дополнительном коде некоторые целые 8-разрядные числа (n=7):

			64	32	16	8	4	2	1	
		7	**6**	**5**	**4**	**3**	**2**	**1**	**0**	.
	38	0	0	1	0	0	1	1	0	.
	-38	1	1	0	1	1	0	0	1	.
									1	
		1	**1**	**0**	**1**	**1**	**0**	**1**	**0**	
max	**127**	**0**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	.
	-127	1	0	0	0	0	0	0	1	.
min	**-128**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	.
	0+	0	0	0	0	0	0	0	0	.
	0-	0	0	0	0	0	0	0	0	.

Отсюда следует, что диапазон значений целых чисел в дополнительном коде задается неравенством:

$$-2^{n} \le N_{CC} \le 2^{n}-1.$$

Например, значения 8-разрядного целого числа будут находиться в интервале [-128; +127].

32-разрядное целое число будет принимать значения в диапазоне $[-2^{31}; (2^{31}-1)]$.

В случае дробных чисел, диапазон значений задается неравенством:

$$-1 \le N_{CC} \le 1-2^{-m}.$$

Примеры

Определите значение числа, представленного в 8 битах (n = 7) для разных форматов

7	6	5	4	3	2	1	0
1	0	0	1	0	0	1	1

1. Целое число без знака

7	6	5	4	3	2	1	0	
1	0	0	1	0	0	1	1	.
128+			16+			2+	1=147	

2. Дробное число без знака

	-1	-2	-3	-4	-5	-6	-7	-8
.	1	0	0	1	0	0	1	1

$2^{-1}+$ $2^{-4}+$ $2^{-7}+$ 2^{-8} =

= 0,5+ 0,0625+ 0,0078125+ 0,00392= 0,57352

3. Целое число со знаком в прямом коде

7(Sg)	6	5	4	3	2	1	0	
1	0	0	1	0	0	1	1	.

\- 2^4+ 2^1+ 2^0 = - 16+2+1= - 19

4. Целое число со знаком в обратном коде

7(Sg)	6	5	4	3	2	1	0	
1	0	0	1	0	0	1	1	.

N_{CI}=1.0010011

N_{CD}=1.1101100= $-(2^6+2^5+2^3+2^2)$ = - (64+32+8 +4)= -108

5. Целое число со знаком в дополнительном коде

7(Sg)	6	5	4	3	2	1	0	
1	0	0	1	0	0	1	1	.

N_{CC}=1.0010011

	1	1	1	0	1	1	0	0
								1
N_{CD}=	1	1	1	0	1	1	0	1

= - 2^6+ 2^5+ 2^3+ 2^2 + 2^0 = - (64+32+8 +4+1)= -109

Представление с плавающей точкой

Другой формой представления двоичных чисел является *представление с плавающей точкой*. Представление некоторого числа N в форме с плавающей точкой определяется выражением

$$N = m_N 2^{e_N},$$

где m_N и e_N соответствуют мантиссе и порядку двоичного числа N.

Следовательно, формат числа с плавающей точкой имеет два поля: одно для мантиссы и другое для порядка. В соответствии с существующими соглашениями мантисса двоичного числа, представленного в форме с плавающей точкой должна удовлетворять неравенству

$$\frac{1}{2} \le |m_N| < 1.$$

Один из стандартных форматов для чисел с плавающей точкой включает 32 разряда и имеет следующий вид:

разряд:	31	30 . . . 23	22 . . . 0
	Sg_m	e_N	m_N

В этом стандартном формате мантисса m_N представлена в прямом коде, причем цифра знака Sg_m находится в старшем разряде. Порядок e_N представляется в дополнительном коде. Таким образом, порядок может принимать значения в интервале $-128 \le e_N \le 127$. Минимальное значение числа в вышеуказанном формате равно

$$|N_{min}| = (0.100...0)\cdot 2^{-128} = 2^{-1}\cdot 2^{-128} = 2^{-129},$$

а максимальное

$$|N_{max}| = (0.111...1)\cdot 2^{127} = (1-2^{-23})\cdot 2^{127} \approx 2^{127}.$$

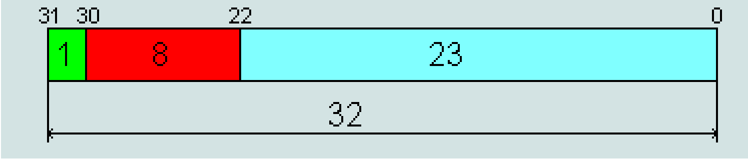
Стандарт IEEE 754-1985 определяет:

- как представлять нормализованные положительные и отрицательные числа с плавающей точкой
- как представлять денормализованные положительные и отрицательные числа с плавающей точкой
- как представлять нулевые числа
- как представлять специальную величину бесконечность (Infinity)
- как представлять специальную величину "Не число" (NaN или NaNs)
- четыре режима округления

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

- с одинарной точностью (single-precision) 32 бита
- с двойной точностью (double-precision) 64 бита
- с одинарной расширенной точностью (single-extended precision) >=43 бит (редко используемый)
- с двойной расширенной точностью (double-extended precision) >= 79 бит (обычно используют 80 бит)

Описание преобразования в 32 битный формат IEEE 754:



1. Число может быть + или - .
   Поэтому отводится 1 бит для обозначения знака числа:
   0-положительное
   1-отрицательное
   Этот самый старший бит в 32 битной последовательности .

2. Далее пойдут биты экспоненты, для этого выделяют 1 байт (8 бит).
   Экспонента может быть, как и число, со знаком + или -.
   Для определения знака экспоненты, чтобы не вводить ещё один бит знака, добавляют смещение к экспоненте в половину байта +127(0111 1111). То есть, если наша экспоната = $+7_{10}$ ($+111_2$), то смещенная экспонента = 7+127=134.

А если бы, наша экспонента была -7 , то смещенная экспонета=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. Оставшиеся 23 бита отводят для мантиссы.
   Но, у нормализованной двоичной мантиссы первый бит всегда равен 1, так как число лежит в диапазоне 1<=M<2.
   Нет смыла, записывать единицу в отведенные 23 бита, поэтому в отведенные 23 бита записывают остаток от мантиссы.

Пример.
Преобразовать 639.6875

$$639.6875 = 1001111111.1011_2 = 1.0011111111011 \times 2^9$$

$$s = 0$$
$$e = 9 + 127 = 136 = 10001000$$
$$f = 0011111111011$$

Знак	Порядок	Мантисса
0	1 0 0 0 1 0 0 0	0 0 1 1 1 1 1 1 1 1 0 1 1 0 0 0 0 0 0 0 0 0 0
1	8	23

Пример.
Преобразовать

1 01111100 11000000000000000000000

1. Знак=1
2. Порядок $01111100_2=124_{10}$ 124-127=-3 e=-3
3. Мантисса 1,110000……..=1,75

$$= (-1.75 * 2^{-3}) = -0.21875.$$