Микропроцессорные устройства обработки сигналов

Лекция L04
«Представление дробных чисел»

http://vykhovanets.ru/course67/

Форматы представления чисел

- Натуральные числа: N8, N16, N32, N64.
- Целые числа: Z8, Z16, Z32, Z64.
- Числа с фиксированной запятой: Q4.12, Q1.15, Q1.31.
- Числа с плавающей запятой: F32, F64, F80.
- Рациональные числа $\frac{\text{числитель}}{\text{знаменатель}}$: R8, R16, R32, R64.
- Логарифмические форматы (логарифм значения и знак).
- Символьные форматы.

Целочисленные форматы

N16 [0, 65 535], ε = 1

Вес	2^{15}	2^{14}	2^{13}	2^{12}	2^{11}	2^{10}	2^{9}	2^{8}	2^{7}	2^{6}	2^{5}	2^{4}	2^{3}	2^{2}	2^{1}	2^{0}
Разряд	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
Бит	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Z16 [-32 768, 32 767], ε = 1

	-2^{15}	2^{14}	2^{13}	2^{12}	2^{11}	2^{10}	2^{9}	2^{8}	2^{7}	2^{6}	2^{5}	2^{4}	2^{3}	2^{2}	2^{1}	2^{0}
	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
	s	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

N32 [0, 4 929 967 295], ε = 1

2^{31}	2^{30}	2^{29}	2^{28}	2^{27}	2^{26}	2^{25}	2^{24}	2^{23}	2^{22}	2^{21}	2^{20}	2^{19}	2^{18}	2^{17}	2^{16}	2^{15}	2^{14}	2^{13}	2^{12}	2^{11}	2^{10}	2^{9}	2^{8}	2^{7}	2^{6}	2^{5}	2^{4}	2^{3}	2^{2}	2^{1}	2^{0}
31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Z32 [-2 147 483 648, 2 147 483 647], ε = 1

-2^{31}	2^{30}	2^{29}	2^{28}	2^{27}	2^{26}	2^{25}	2^{24}	2^{23}	2^{22}	2^{21}	2^{20}	2^{19}	2^{18}	2^{17}	2^{16}	2^{15}	2^{14}	2^{13}	2^{12}	2^{11}	2^{10}	2^{9}	2^{8}	2^{7}	2^{6}	2^{5}	2^{4}	2^{3}	2^{2}	2^{1}	2^{0}
31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
s	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

$$X_{N16} = \sum_{i=0}^{15} x_i \cdot 2^i$$

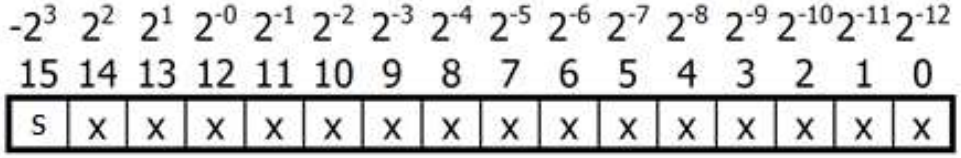
$$X_{Z16} = -s \cdot 2^{15} + \sum_{i=0}^{14} x_i \cdot 2^i$$

$$X_{N32} = \sum_{i=0}^{31} x_i \cdot 2^i$$

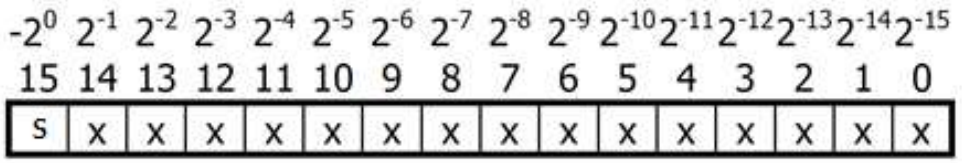
$$X_{Z32} = -s \cdot 2^{31} + \sum_{i=0}^{30} x_i \cdot 2^i$$

Дробные форматы

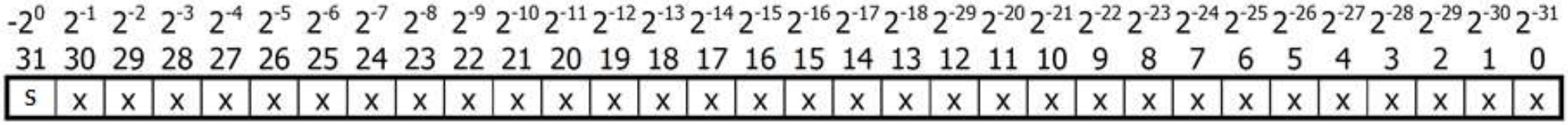
Q4.12 [-8, 8), $\varepsilon = 2^{-12} = 2{,}44\times10^{-4}$



Q1.15 [-1, 1), $\varepsilon = 2^{-15} = 3{,}05\times10^{-5}$



Q1.31 [-1, 1), $\varepsilon = 2^{-31} = 4{,}66\times10^{-10}$



$$X_{Q4.12} = -s\cdot 2^3 + \sum_{i=0}^{14} x_i \cdot 2^{-12+i} = 2^{-12}\left(-s\cdot 2^{15} + \sum_{i=0}^{14} x_i \cdot 2^{i}\right) = 2^{-12}\cdot X_{Z16}$$

$$X_{Q1.15} = -s + \sum_{i=0}^{14} x_i \cdot 2^{-15+i} = 2^{-15}\left(-s\cdot 2^{15} + \sum_{i=0}^{14} x_i \cdot 2^{i}\right) = 2^{-16}\cdot X_{Z16}$$

$$X_{Q1.31} = -s + \sum_{i=0}^{30} x_i \cdot 2^{-31+i} = 2^{-31}\left(-s\cdot 2^{31} + \sum_{i=0}^{30} x_i \cdot 2^{i}\right) = 2^{-31}\cdot X_{Z32}$$

Фиксированная запятая

$$0000{,}0001_2 = \frac{1}{16} = 0{,}625$$

$$0001{,}1000_2 = \frac{24}{16} = 1{,}5$$

$$\frac{a}{c} + \frac{b}{c} = \frac{a+b}{c} \qquad \frac{a}{c} * \frac{b}{c} = \frac{(a*b)/c}{c}$$

$$1{,}5 * 2{,}5 = 3{,}75$$

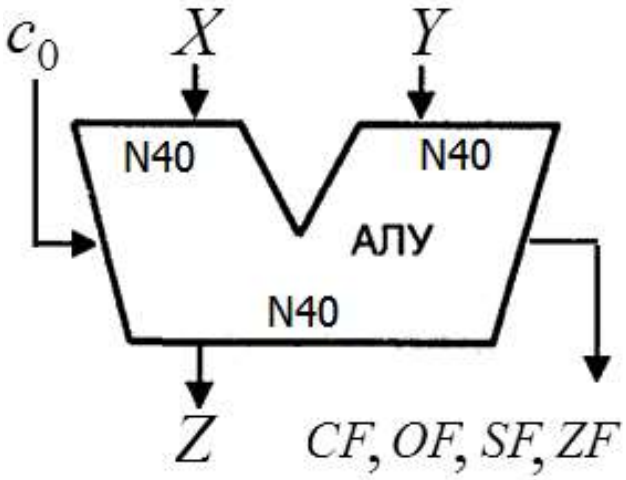
$$0001{,}1000_2 * 0010{,}1000_2 =$$

$$= 0011{,}1100 \mid 0000_2$$

Внутренний формат Q9.31

Q9.31 [-256, 256), ε = 2^{-31} = 4,66×10^{-10}

-2^8	2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^{-0}	2^{-1}	2^{-2}	2^{-3}		2^{-13}	2^{-14}	2^{-15}	2^{-16}	2^{-17}	2^{-18}	2^{-29}	2^{-20}	2^{-21}	2^{-22}	2^{-23}	2^{-24}	2^{-25}	2^{-26}	2^{-27}	2^{-28}	2^{-29}	2^{-30}	2^{-31}
39	38	37	36	35	34	33	32	31	30	29	28	…	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
s	X	X	X	X	X	X	X	X	X	X	X	…	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X



$$X_{N40} = \sum_{i=0}^{39} x_i \cdot 2^i$$

$$X_{Z40} = -s \cdot 2^{39} + \sum_{i=0}^{38} x_i \cdot 2^i$$

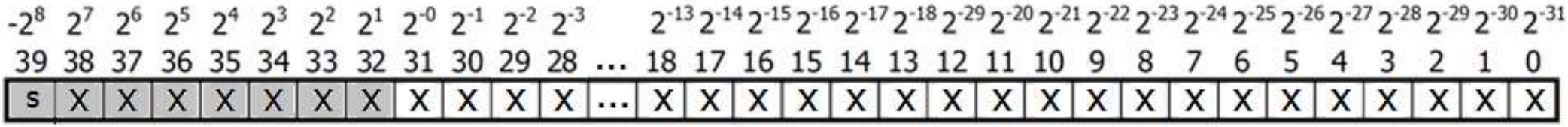
$$X_{Q9.31} = -s \cdot 2^8 + 2^{-31} \sum_{i=0}^{38} x_i \cdot 2^i$$

$$X_{Q9.31} = 2^{-31}\left(-s \cdot 2^{39} + \sum_{i=0}^{38} x_i \cdot 2^i\right) = 2^{-31} \cdot X_{Z40}$$

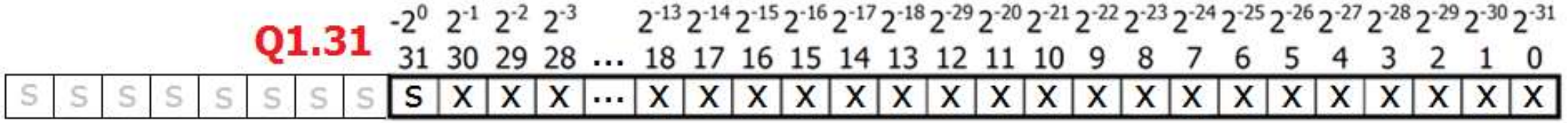
$$X_{Q1.31} \pm Y_{Q1.31} = 2^{-31}\left(X_{Z40} \pm Y_{Z40}\right)$$

Преобразование дробных

Q9.31



Q1.31



$$X_{Q9.31} = -s \cdot 2^8 + 2^{-31} \sum_{i=0}^{38} x_i \cdot 2^i$$

$$X_{Q1.31} = -s + 2^{-31} \sum_{i=0}^{30} x_i \cdot 2^i$$

$$\sum_{i=0}^{n-1} aq^i = a \frac{q^n - 1}{q - 1}$$

$$2^8 - \sum_{i=0}^{7} 2^i = 2^8 - \frac{2^8 - 1}{2 - 1} = 2^0$$

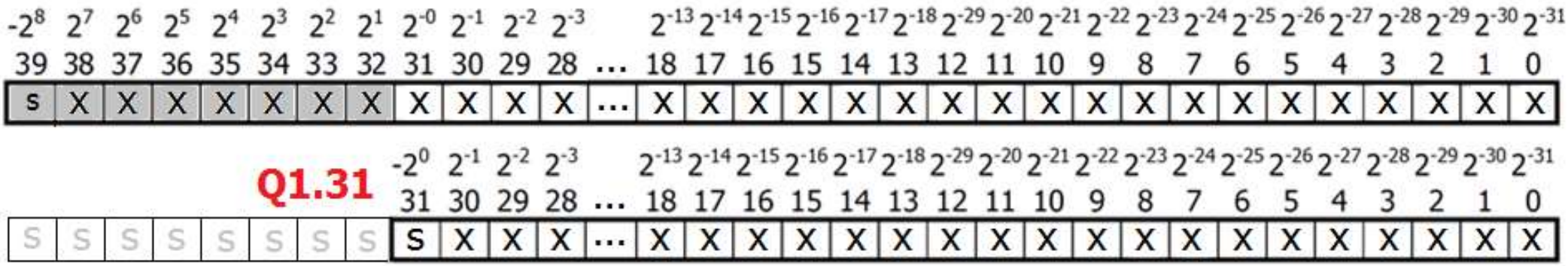
$$X_{Q1.31} = -s \cdot 2^0 + 2^{-31} \sum_{i=0}^{30} x_i \cdot 2^i =$$

$$= -s \left(2^8 - \sum_{i=0}^{7} 2^i \right) + 2^{-31} \sum_{i=0}^{30} x_i \cdot 2^i =$$

$$= -s \cdot 2^8 + 2^{-31} \sum_{i=0}^{38} x_i \cdot 2^i \Bigg| \begin{array}{l} x_{38} = s \\ x_{37} = s \\ \cdots \\ x_{31} = s \end{array}$$

Насыщение

Q9.31



Q9.31		Q1.31
FE80000009h	≠	80000009h
0180000009h	≠	80000009h

Q9.31		Q9.31		Q1.31	
FE80000009h	≈	FF80000000h	=	80000000h	(-1)
0180000009h	≈	007FFFFFFFh	=	7FFFFFFFh	$(+1-2^{-31})$

FF80000000h – минимальное число в Q1.31
007FFFFFFFh – максимальное число в Q1.31

Насыщение (saturation)

Умножение дробных

$$X_{Q1.15} = -s_x + 2^{-15}\sum_{i=0}^{14} x_i \cdot 2^i, \quad Y_{Q1.15} = -s_y + 2^{-15}\sum_{i=0}^{14} y_i \cdot 2^i.$$

$$Z = 2^{-30}\left(-s_x \cdot 2^{15} + \sum_{i=0}^{14} x_i \cdot 2^i\right)\left(-s_y \cdot 2^{15} + \sum_{i=0}^{14} y_i \cdot 2^i\right)$$

$$Z = 2^{-30}\left(X_{Z16} \cdot Y_{Z16}\right) = 2^{-30} Z_{Z32}, \quad Z_{Q1.31} = 2 \cdot Z.$$

Z32

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
s	s	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Q1.31

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
s	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	0

$$Z_{Q1.31} = -s_z + 2^{-31}\sum_{i=0}^{30} z_i \cdot 2^i$$

Q1.15

15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
s	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

$$Z_{Q1.31} = Z_{Z32} << 1$$

```
int x, y; long z;
x = 0xFDE3, y = 0x71A4;
z = ((long)x*y)<<1;
```

$$Z_{Q1.15} = Z_{Z32} >> 15$$

```
int x, y, z;
x = 0xFDE3, y = 0x71A4;
z = ((long)x*y)>>15;
```

Деление дробных

$$f(x) = \frac{1}{x} - c \qquad f(x) = 0 \rightarrow x = \frac{1}{c}$$

$$f(x) = 0 \qquad x_{m+1} = x_m - \frac{f(x_m)}{f'(x_m)}$$

$$x_{m+1} = 2x_m - x_m^2 c \ \ (|x_{m+1} - x_m| < \varepsilon)$$

$$c = M \cdot 2^E, \qquad |M| \in [\frac{1}{2}, 1)$$

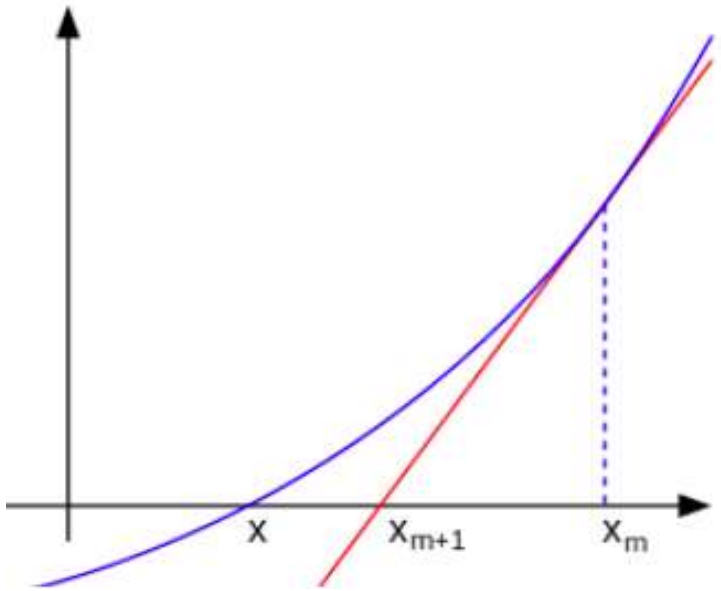
$$M = 01XX \ldots Xb, \quad E = 0YYY \ldots Yb$$

$$x_{m+1} = 2x_m - x_m^2 M \ (m = 0, 1, 2, 3)$$

$$x_0 = (M \ll 1)\text{^}0\text{x}1\text{FFF}$$

$$E_0 = E - 1$$

Метод Ньютона
(метод касательных)

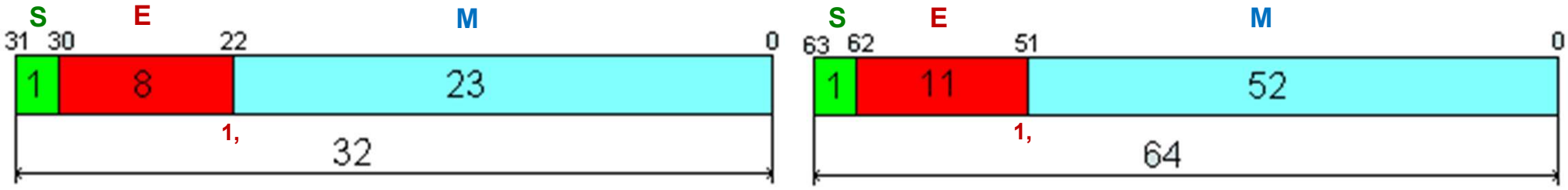


void ldiv16(LDATA *x, DATA *y, DATA *z, DATA *zexp, ushort nx);

-2^0	2^{-1}	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}	2^{-9}	2^{-10}	2^{-11}	2^{-12}	2^{-13}	2^{-14}	2^{-15}	2^{-16}	2^{-17}	2^{-18}	2^{-29}	2^{-20}	2^{-21}	2^{-22}	2^{-23}	2^{-24}	2^{-25}	2^{-26}	2^{-27}	2^{-28}	2^{-29}	2^{-30}	2^{-31}
31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
s	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Форматы с плавающей запятой

IEEE 754



F32 $(-1)^S 2^{(E-127)} (1+M/2^{23})$ F64 $(-1)^S 2^{(E-1023)} (1+M/2^{52})$

S – **S**ign (знак числа)

E – **E**xponent (8 или 11 бит смещенного на 127 или 1023 порядка числа)

M – **M**antissa (23 или 52 бита мантиссы, дробная часть числа)

0100 0011 0001 1011 1010 0000 0000 0000b 431BA000h

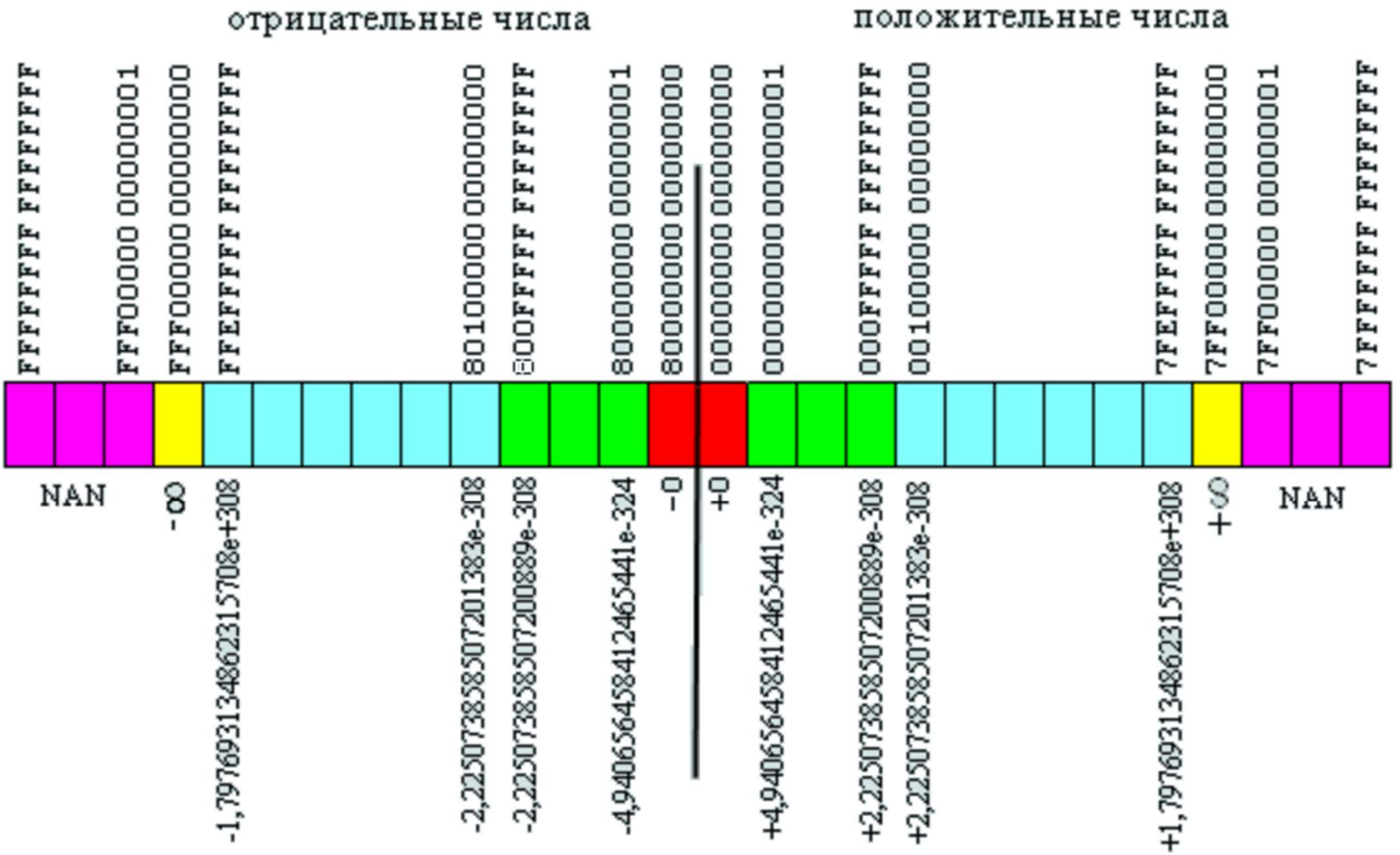
S **E** **M**

0 10000110 00110111010000000000000b

0 134 1810432

$(-1)^0 \cdot 2^{(134-127)} \cdot (1+1810432/2^{23}) = 155{,}625$

Числа с плавающей запятой



- нулевые числа
- денормализованные числа
- нормализованные числа
- бесконечность
- не числа (NANs)

Нормализованные числа

F32 $(-1)^S 2^{(E-127)} (1+M/2^{23})$

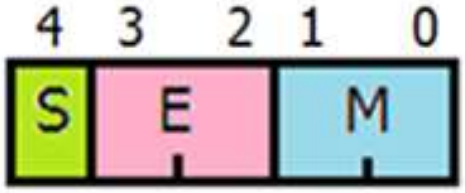
F64 $(-1)^S 2^{(E-1023)} (1+M/2^{52})$

Денормализованные числа

F32 $(-1)^S 2^{-127} M/2^{23}$

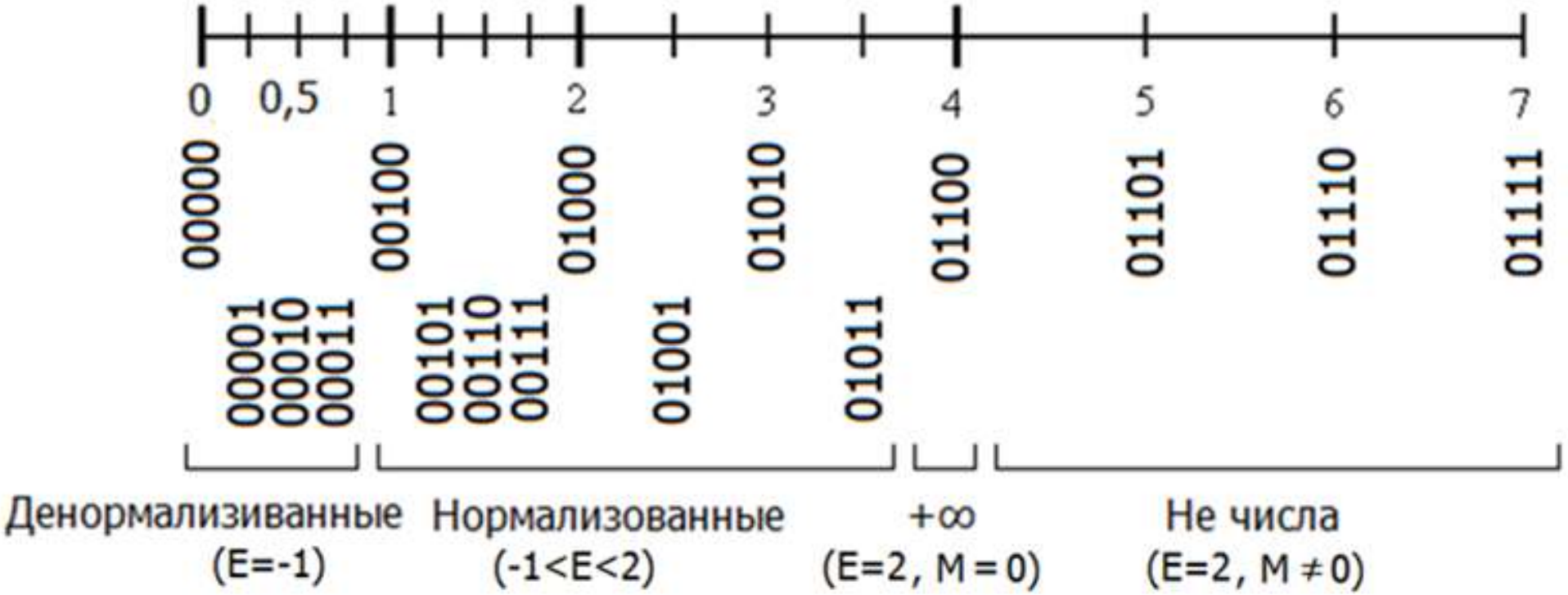
F64 $(-1)^S 2^{-1023} M/2^{52}$

Денормализованные числа

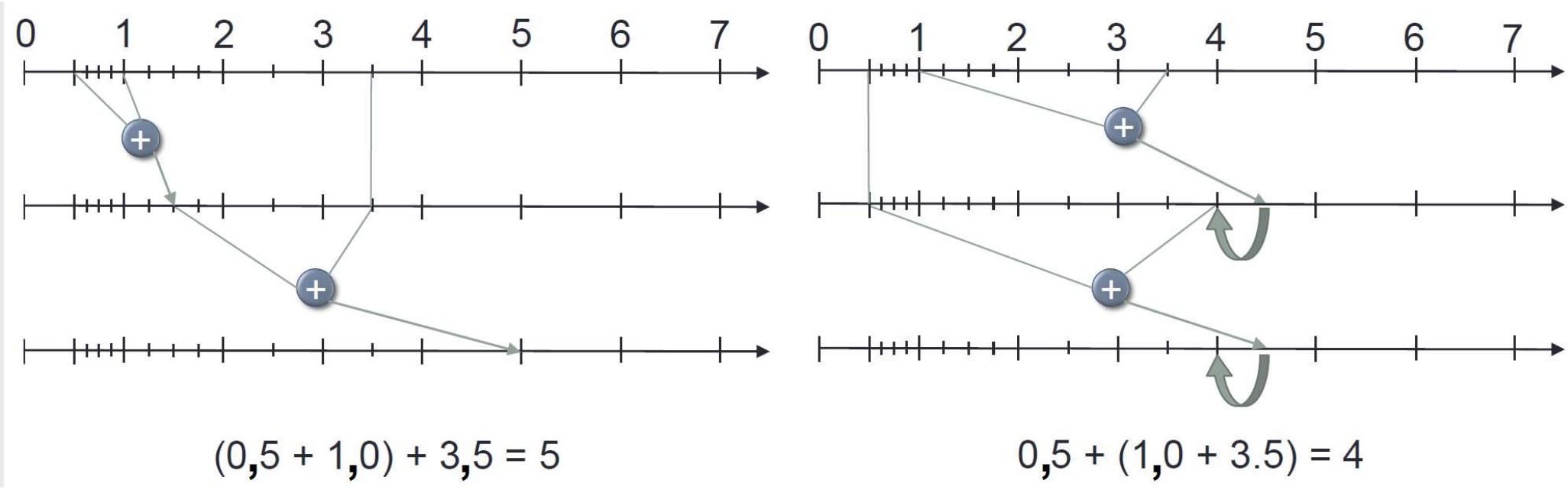


E: -1, 0, 1, 2 (-1⩽E⩽2).

M: 1,00, 1,01, 1,10, 1,11 (1,00⩽M⩽1,11).



Пример вычислений



(0,5 + 1,0) + 3,5 = 5

0,5 + (1,0 + 3.5) = 4

Методы округления

- Round To Nearest, Ties to Even – округление в сторону ближайшего, к чётному в случае конфликтов.
  - Режим по умолчанию в IEEE 754.

$01{,}101_2 \to 01{,}1_2, 01{,}110_2 \to 10{,}0_2, 01{,}111_2 \to 10{,}0_2, -01{,}101_2 \to -01{,}1_2$

- Round Up - округление в сторону $+\infty$;

$01{,}101_2 \to 10{,}0_2, 01{,}110_2 \to 10{,}0_2, 01{,}111_2 \to 10{,}0_2, -01{,}101_2 \to -01{,}1_2$

- Round Down – округление в сторону $-\infty$;

$01{,}101_2 \to 01{,}1_2, 01{,}110_2 \to 01{,}1_2, 01{,}111_2 \to 01{,}1_2, -01{,}101_2 \to -10{,}0_2$

- Round Towards Zero – округление в сторону 0;
  - Для реализации достаточно отбросить «лишние» биты.

$01{,}101_2 \to 01{,}1_2, 01{,}110_2 \to 01{,}1_2, 01{,}111_2 \to 01{,}1_2, -01{,}101_2 \to -01{,}1_2$

Арифметические проблемы

- Не все числа имеют представление.
- Преобразование в целые: 63,0/9,0→7, 0,63/0,09→6.
- Преобразование в символьные форматы и обратно, многие числа нельзя ввести или вывести точно: 0,2→ 0,200000000003.
- Порядок вычисления может влиять на результат и его точность: не выполняются законы ассоциативности и дистрибутивности.
- Проблемы сравнения: x==y.

Рекомендации

- Выполнять вычисления в одном формате чисел.
- Избегать лишних преобразований форматов.
- При сравнении чисел F32, F64 использовать $abs(x-y) \leq \epsilon$.
- Избегать сложений чисел, экспоненты которых сильно отличаются.
- Избегать вычитания близких чисел.
- Функции exp, log, sin и т.д. не вычисляются точно.

Goldberg D. *What Every Computer Scientist Should Know About Floating-Point Arithmetic //* Computing Surveys*. 1991.*