

# Арифметика с плавающей запятой

Лисицын Сергей  
МФТИ 2019г.

# История

1976

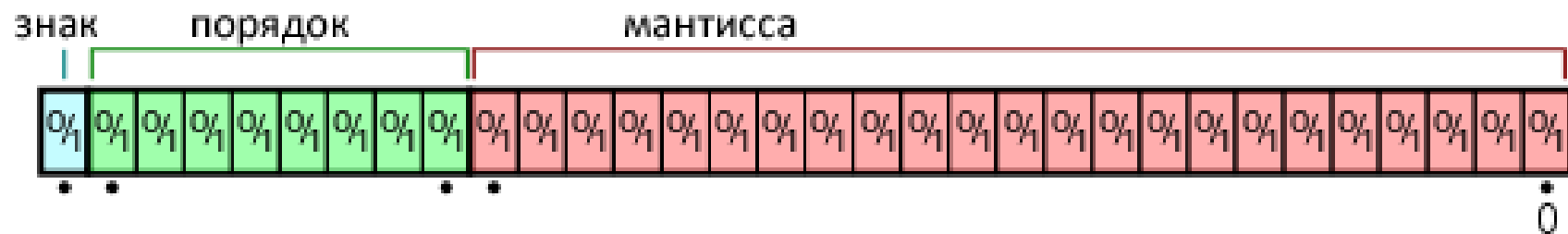
DEC, National Superconductor, Zilog, Motorola, Intel

VAX (DEC) VS K-C-S (Уильям Кэхэн, Джероми Кунен и Гарольд Стоун)

Стандарт IEEE 754:

- формат чисел с плавающей точкой;
- представление положительного и отрицательного нуля, положительной и отрицательной бесконечностей, а также нечисла;
- методы, используемые для преобразования числа при выполнении математических операций;
- исключительные ситуации: деление на ноль, переполнение, потеря значимости, работа с денормализованными числами и другие;
- операции: арифметические и другие.

# ОСНОВЫ



$$(-1)^s \times M \times B^E$$

$s$  — знак

$B$  — основание

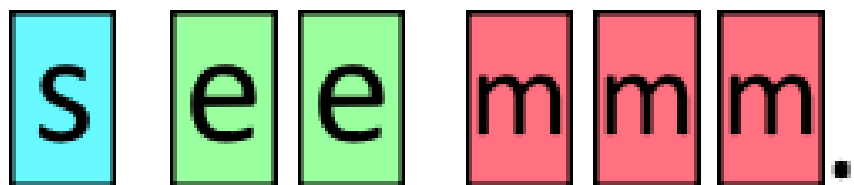
$E$  — порядок

$M$  — мантисса

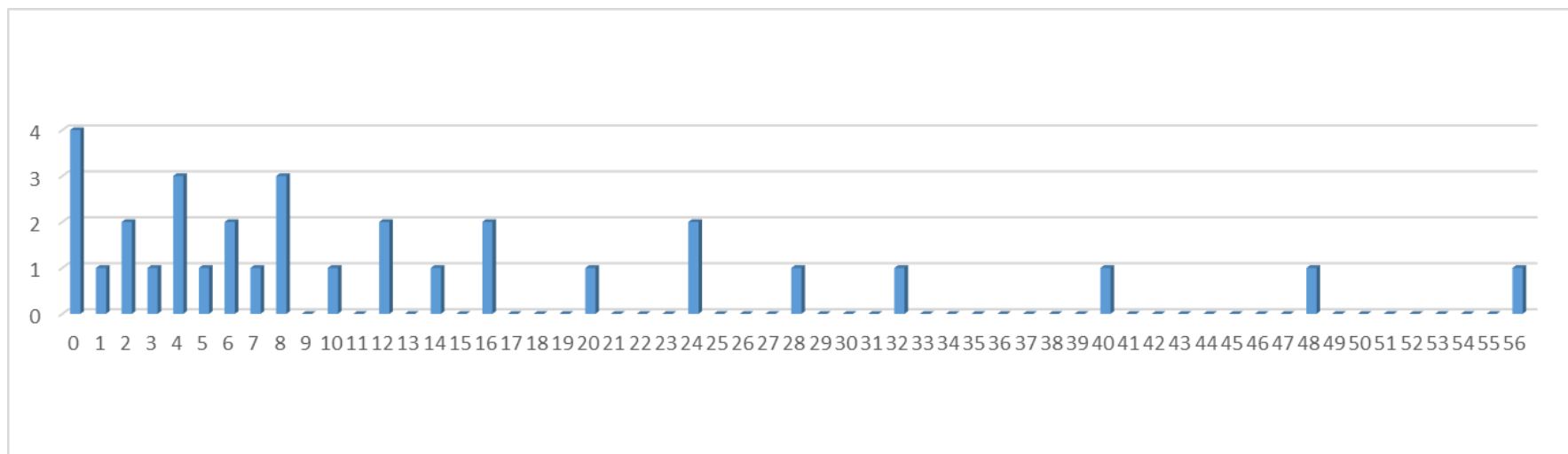
$B = 2$  наиболее устойчиво к ошибкам округления

$$1,010e+1 = 1 \times 2^1 + 0 \times 2^0 + 1 \times 2^{-1} + 0 \times 2^{-2} = 2 + 0,5 = 2,5$$

# Тип float64

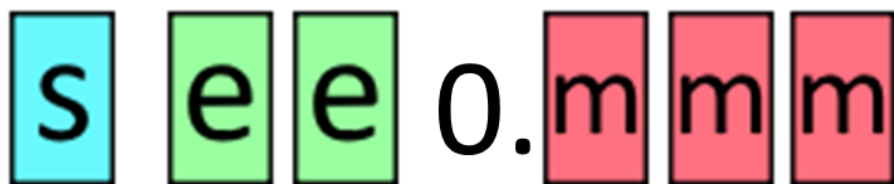


$$(-1)^s \times M \times 2^E$$

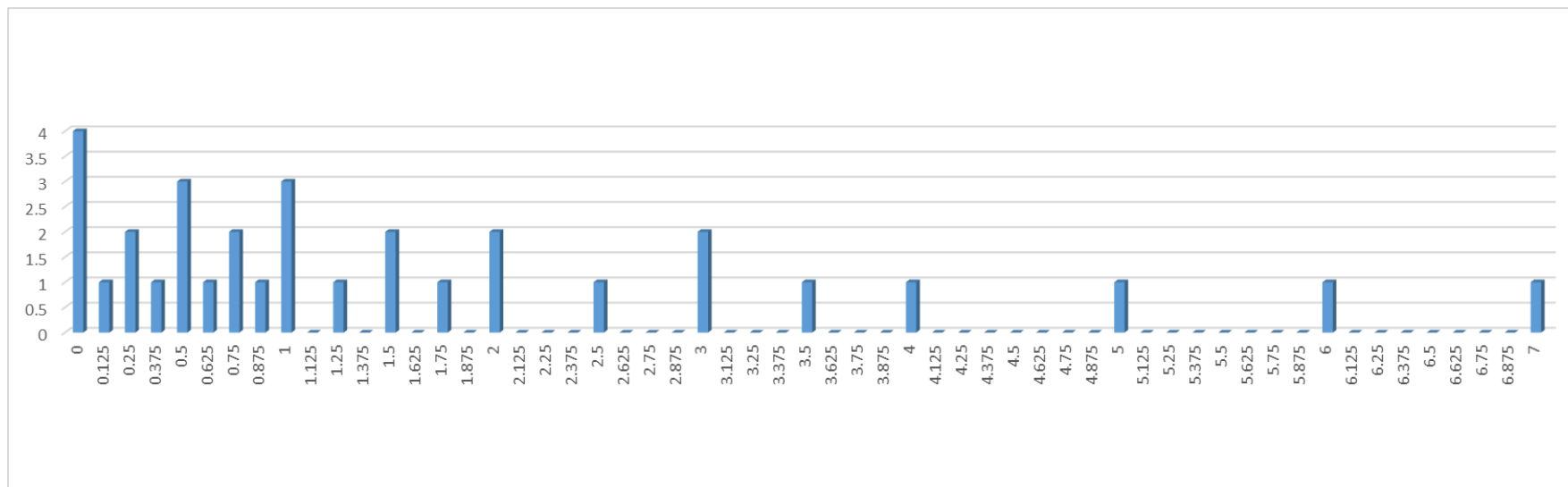


exp	mantissa	res
0 0	0 0 0	= 0
0 0	0 0 1	= 1
0 0	0 1 0	= 2
0 0	0 1 1	= 3
0 0	1 0 0	= 4
0 0	1 0 1	= 5
0 0	1 1 0	= 6
0 0	1 1 1	= 7
0 1	0 0 0	= 0
0 1	0 0 1	= 2
0 1	0 1 0	= 4
0 1	0 1 1	= 6
0 1	1 0 0	= 8
0 1	1 0 1	= 10
0 1	1 1 0	= 12
0 1	1 1 1	= 14
1 0	0 0 0	= 0
1 0	0 0 1	= 4
1 0	0 1 0	= 8
1 0	0 1 1	= 12
1 0	1 0 0	= 16
1 0	1 0 1	= 20
1 0	1 1 0	= 24
1 0	1 1 1	= 28
1 1	0 0 0	= 0
1 1	0 0 1	= 8
1 1	0 1 0	= 16
1 1	0 1 1	= 24
1 1	1 0 0	= 32
1 1	1 0 1	= 40
1 1	1 1 0	= 48
1 1	1 1 1	= 56

# Эффективный порядок

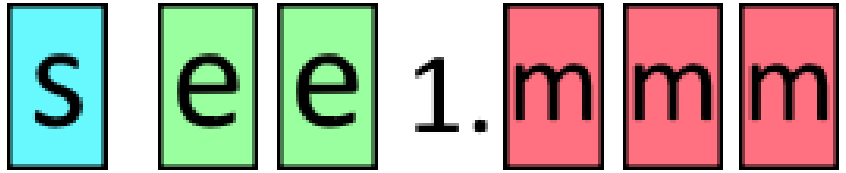


$$(-1)^s \times 0.M \times 2^E$$

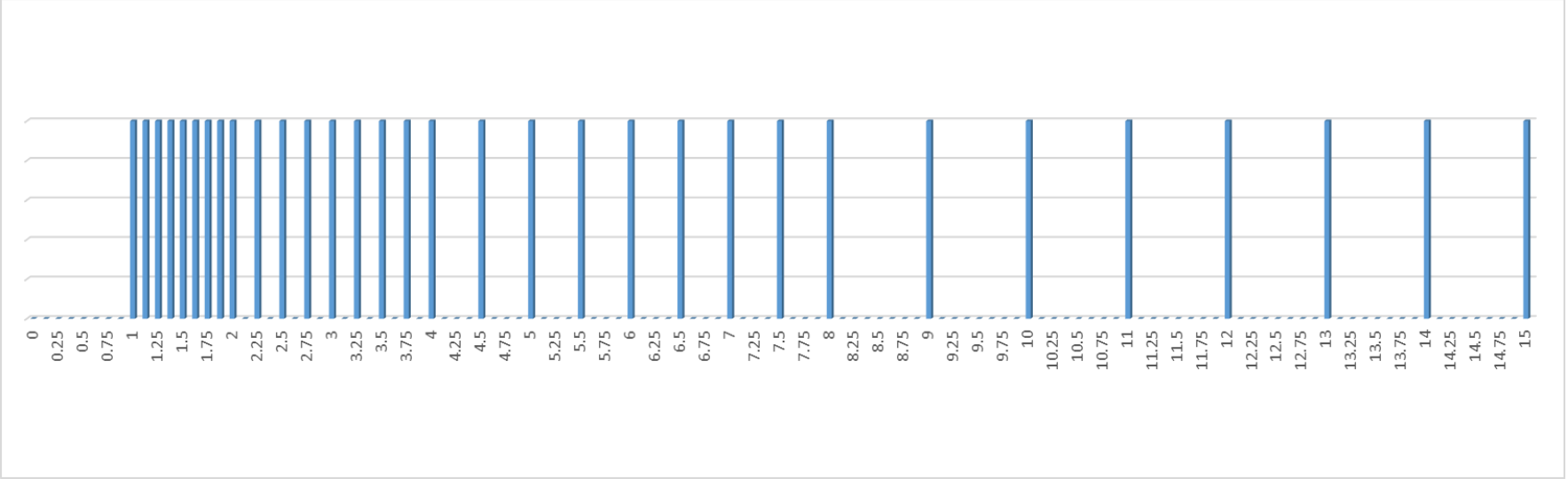


exp	mantissa			res
0	0	0	0	= 0
0	0	0	1	= 0.125
0	0	0	1	= 0.25
0	0	0	1	= 0.375
0	0	1	0	= 0.5
0	0	1	0	= 0.625
0	0	1	1	= 0.75
0	0	1	1	= 0.875
0	1	0	0	= 0
0	1	0	0	= 0.25
0	1	0	1	= 0.5
0	1	0	1	= 0.75
0	1	1	0	= 1
0	1	1	0	= 1.25
0	1	1	1	= 1.5
0	1	1	1	= 1.75
1	0	0	0	= 0
1	0	0	0	= 0.5
1	0	0	1	= 1
1	0	0	1	= 1.5
1	0	1	0	= 2
1	0	1	0	= 2.5
1	0	1	1	= 3
1	0	1	1	= 3.5
1	1	0	0	= 0
1	1	0	0	= 1
1	1	0	1	= 2
1	1	0	1	= 3
1	1	1	0	= 4
1	1	1	0	= 5
1	1	1	1	= 6
1	1	1	1	= 7

# Нормализованные числа

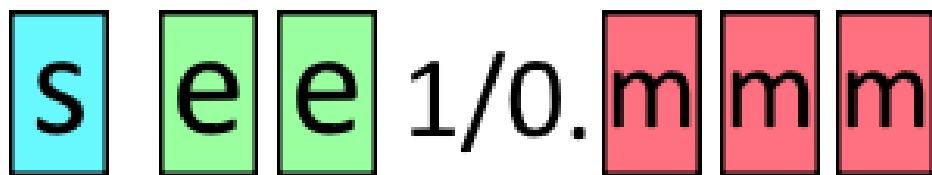


$$(-1)^s \times 1.M \times 2^E$$



exp		mantissa				res
0	0	0	0	0	=	1
0	0	0	0	1	=	1.125
0	0	0	1	0	=	1.25
0	0	0	1	1	=	1.375
0	0	1	0	0	=	1.5
0	0	1	0	1	=	1.625
0	0	1	1	0	=	1.75
0	0	1	1	1	=	1.875
0	1	0	0	0	=	2
0	1	0	0	1	=	2.25
0	1	0	1	0	=	2.5
0	1	0	1	1	=	2.75
0	1	1	0	0	=	3
0	1	1	0	1	=	3.25
0	1	1	1	0	=	3.5
0	1	1	1	1	=	3.75
1	0	0	0	0	=	4
1	0	0	0	1	=	4.5
1	0	0	1	0	=	5
1	0	0	1	1	=	5.5
1	0	1	0	0	=	6
1	0	1	0	1	=	6.5
1	0	1	1	0	=	7
1	0	1	1	1	=	7.5
1	1	0	0	0	=	8
1	1	0	0	1	=	9
1	1	0	1	0	=	10
1	1	0	1	1	=	11
1	1	1	0	0	=	12
1	1	1	0	1	=	13
1	1	1	1	0	=	14
1	1	1	1	1	=	15

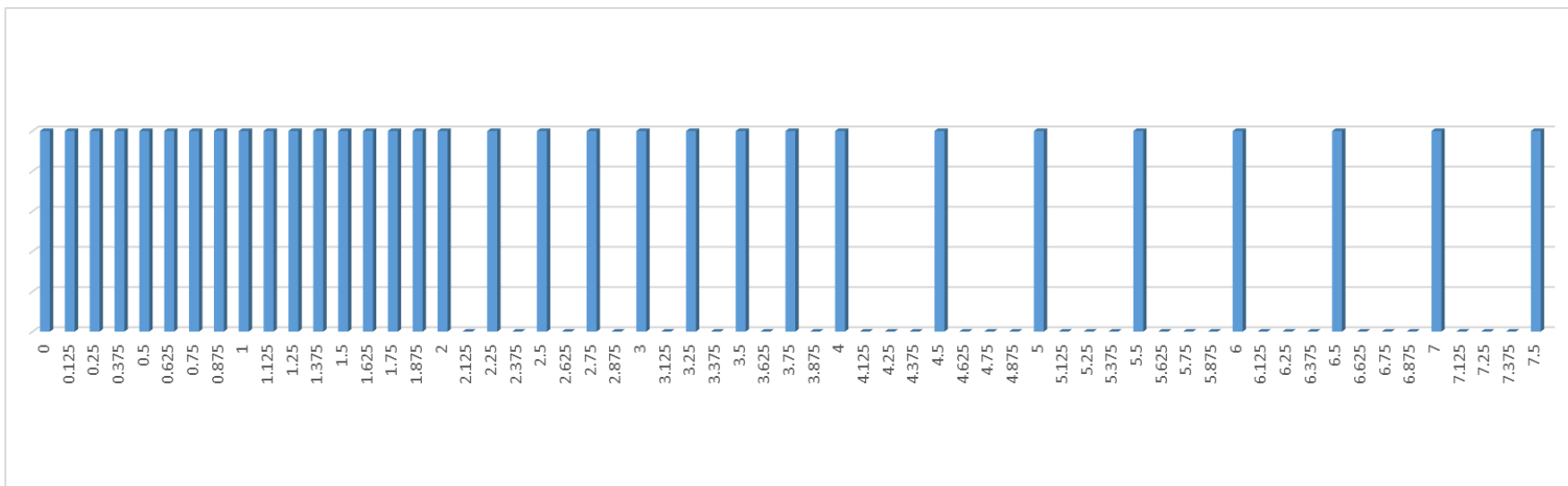
# Денормализованные числа



$(-1)^s \times 1.M \times 2^E$ , если  $E_{\min} \leq E \leq E_{\max}$

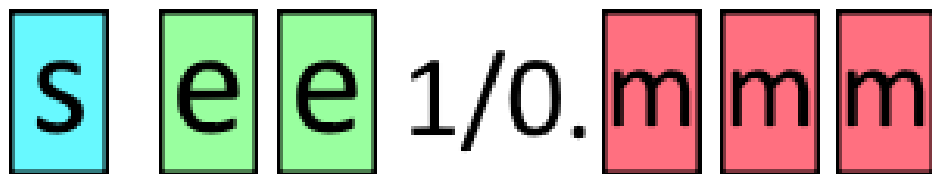
$(-1)^s \times 0.M \times 2^{E_{\min}}$ , если  $E = E_{\min} - 1$

$E_{\max} = 3$ ,  $E_{\min} = 1$ ,  $E = 0$  – денормализованные числа



exp		mantissa				res
0	0	0	0	0	=	0
0	0	0	0	1	=	0.125
0	0	0	1	0	=	0.25
0	0	0	1	1	=	0.375
0	0	1	0	0	=	0.5
0	0	1	0	1	=	0.625
0	0	1	1	0	=	0.75
0	0	1	1	1	=	0.875
0	1	0	0	0	=	1
0	1	0	0	1	=	1.125
0	1	0	1	0	=	1.25
0	1	0	1	1	=	1.375
0	1	1	0	0	=	1.5
0	1	1	0	1	=	1.625
0	1	1	1	0	=	1.75
0	1	1	1	1	=	1.875
1	0	0	0	0	=	2
1	0	0	0	1	=	2.25
1	0	0	1	0	=	2.5
1	0	0	1	1	=	2.75
1	0	1	0	0	=	3
1	0	1	0	1	=	3.25
1	0	1	1	0	=	3.5
1	0	1	1	1	=	3.75
1	1	0	0	0	=	4
1	1	0	0	1	=	4.5
1	1	0	1	0	=	5
1	1	0	1	1	=	5.5
1	1	1	0	0	=	6
1	1	1	0	1	=	6.5
1	1	1	1	0	=	7
1	1	1	1	1	=	7.5

# Специальные числа



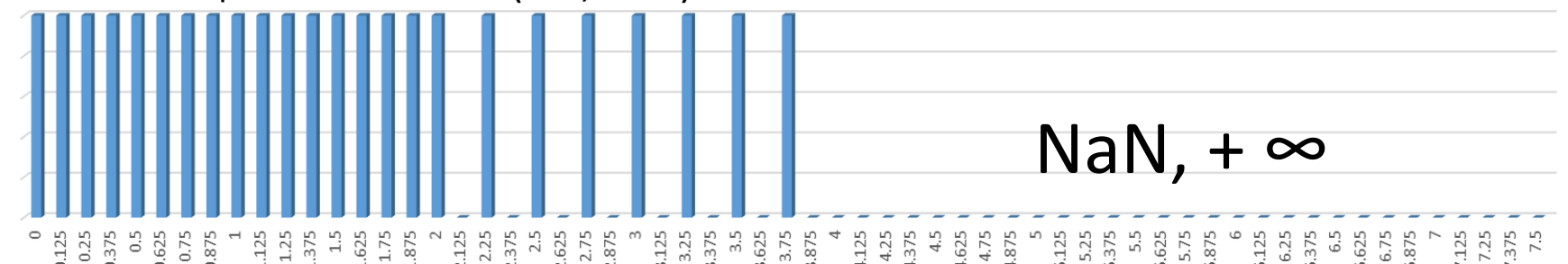
$(-1)^s \times 1.M \times 2^E$ , если  $E_{min} \leq E \leq E_{max}$

$(-1)^s \times 0.M \times 2^{E_{min}}$ , если  $E = E_{min} - 1$

$E_{max} = 2$ ,  $E_{min} = 1$

$E = 0$  – денормализованные числа

$E = 3$  – специальные числа ( $\pm\infty$ , NaN)



NaN:

$\infty + (-\infty)$

$0 \times \infty$

$0/0, \infty/\infty$

$\sqrt{x}$ , где  $x < 0$

$\pm 0$ :

$(+\infty/0) + \infty = +\infty$

$(+\infty/-0) + \infty = \text{NaN}$

exp	mantissa					res
0	0	0	0	0	0	0
0	0	0	0	1	0	0.125
0	0	0	1	0	0	0.25
0	0	0	1	1	0	0.375
0	0	1	0	0	0	0.5
0	0	1	0	1	0	0.625
0	0	1	1	0	0	0.75
0	0	1	1	1	0	0.875
0	1	0	0	0	0	1
0	1	0	0	1	0	1.125
0	1	0	1	0	0	1.25
0	1	0	1	1	0	1.375
0	1	1	0	0	0	1.5
0	1	1	0	1	0	1.625
0	1	1	1	0	0	1.75
0	1	1	1	1	0	1.875
1	0	0	0	0	0	2
1	0	0	0	1	0	2.25
1	0	0	1	0	0	2.5
1	0	0	1	1	0	2.75
1	0	1	0	0	0	3
1	0	1	0	1	0	3.25
1	0	1	1	0	0	3.5
1	0	1	1	1	0	3.75
1	1	0	0	0	0	$+\infty$
1	1	0	0	1	0	NaN
1	1	0	1	0	0	NaN
1	1	0	1	1	0	NaN
1	1	1	0	0	0	NaN
1	1	1	0	1	0	NaN
1	1	1	1	0	0	NaN
1	1	1	1	1	0	NaN
1	1	1	1	1	1	NaN



# Неассоциативность арифметики

1	0	0	0	0	=	2
+						
0	0	0	1	1	=	0.375
+						
0	0	0	1	1	=	0.375
=						
1	0	0	1	0	=	2.5

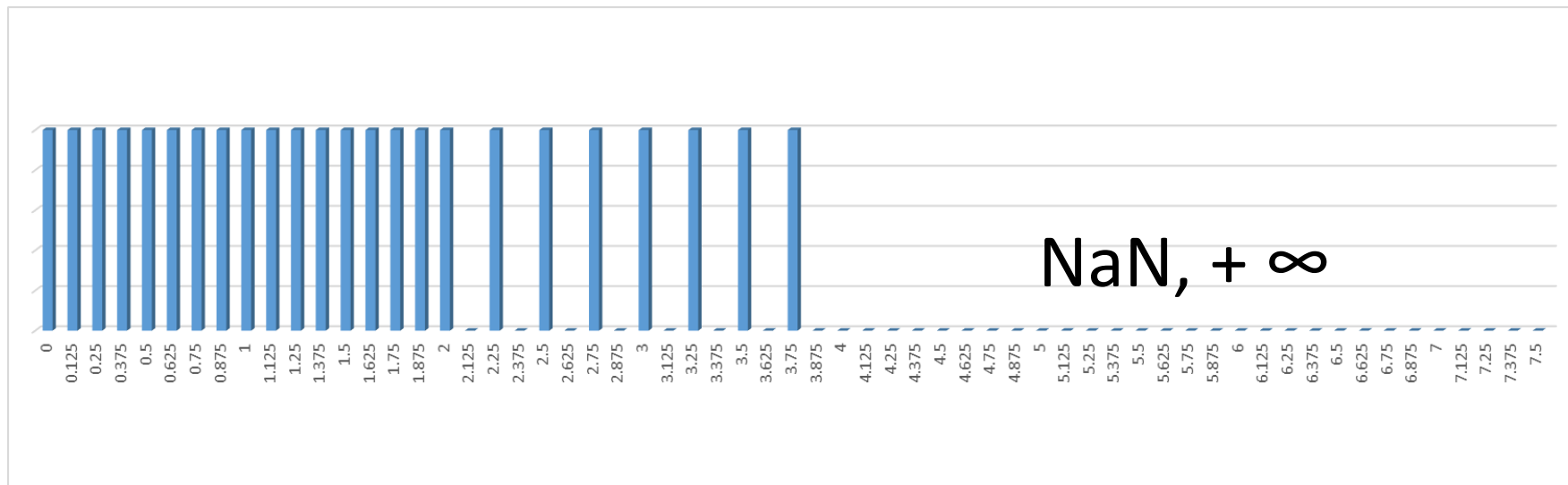
0	0	0	1	1	=	0.375
+						
0	0	0	1	1	=	0.375
+						
1	0	0	0	0	=	2
=						
1	0	0	1	1	=	2.75

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

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

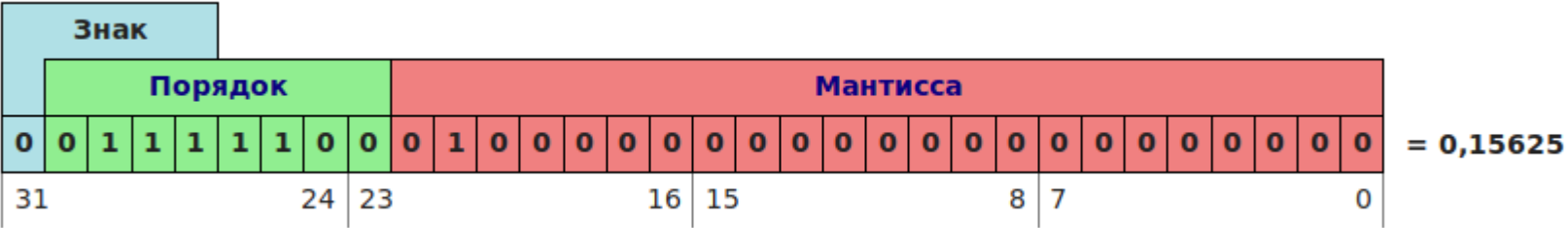
$$x^2 - y^2 \text{ vs } (x - y)(x + y)$$

$$x == y \text{ vs } (x - y) < \epsilon$$



exp		mantissa				res
0	0	0	0	0	=	0
0	0	0	0	1	=	0.125
0	0	0	1	0	=	0.25
0	0	0	1	1	=	0.375
0	0	1	0	0	=	0.5
0	0	1	0	1	=	0.625
0	0	1	1	0	=	0.75
0	0	1	1	1	=	0.875
0	1	0	0	0	=	1
0	1	0	0	1	=	1.125
0	1	0	1	0	=	1.25
0	1	0	1	1	=	1.375
0	1	1	0	0	=	1.5
0	1	1	0	1	=	1.625
0	1	1	1	0	=	1.75
0	1	1	1	1	=	1.875
1	0	0	0	0	=	2
1	0	0	0	1	=	2.25
1	0	0	1	0	=	2.5
1	0	0	1	1	=	2.75
1	0	1	0	0	=	3
1	0	1	0	1	=	3.25
1	0	1	1	0	=	3.5
1	0	1	1	1	=	3.75
1	1	0	0	0	=	$+\infty$
1	1	0	0	1	=	NaN
1	1	0	1	0	=	NaN
1	1	0	1	1	=	NaN
1	1	1	0	0	=	NaN
1	1	1	0	1	=	NaN
1	1	1	1	0	=	NaN
1	1	1	1	1	=	NaN

# IEEE 754



```
3f80 0000 = 1
c000 0000 = -2

7f7f ffff ≈ 3.4028234 × 1038 (максимальное одинарной точности)
0000 0001 = 2-149 ≈ 1.401298464 × 10-45 (Минимальное положительное число одинарной точности – денормализованное)
0080 0000 = 2-126 ≈ 1.175494351 × 10-38 (Минимальное нормализованное положительное число одинарной точности)

0000 0000 = 0
8000 0000 = -0

7f80 0000 = infinity
ff80 0000 = -infinity

3eaa aaab ≈ 1/3
```

$(-1)^s \times 1.M \times 2^{E-127}$ , если  $E_{min} \leq E-127 \leq E_{max}$   
 $(-1)^s \times 0.M \times 2^{E_{min}}$ , если  $E=E_{min}-1$   
 $E_{max} = 127$ ,  $E_{min} = -126$   
 $E-127 = -127$  – денормализованные числа  
 $E-127 = 128$  – специальные числа ( $\pm\infty$ , NaN)

Округление до четного

Название	Полное название	Основание	Кол-во двоичных разрядов мантиссы	Число десятичных разрядов	Экспонента (бит)	Десятичный E max	Смещение экспоненты <sup>[1]</sup>	E min	E max	Примечания
binary16	Половинная точность	2	11	3.31	5	4.51	2 <sup>4</sup> –1 = 15	–14	+15	Не основной
binary32	Одинарная точность	2	24	7.22	8	38.23	2 <sup>7</sup> –1 = 127	–126	+127	
binary64	Двойная точность	2	53	15.95	11	307.95	2 <sup>10</sup> –1 = 1023	–1022	+1023	
binary128	Четырёхкратная точность	2	113	34.02	15	4931.77	2 <sup>14</sup> –1 = 16383	–16382	+16383	
binary256	Восьмикратная точность	2	237	71.34	19	78913.2	2 <sup>18</sup> –1 = 262143	–262142	+262143	Не основной
decimal32		10	7	7	7.58	96	101	–95	+96	Не основной
decimal64		10	16	16	9.58	384	398	–383	+384	
decimal128		10	34	34	13.58	6144	6176	–6143	+6144	