IEEE 754

31 30   23 22   0

← Float (32 bits)

знак   порядок   мантисса

63 62   52 51   0

← Double (64 bits)

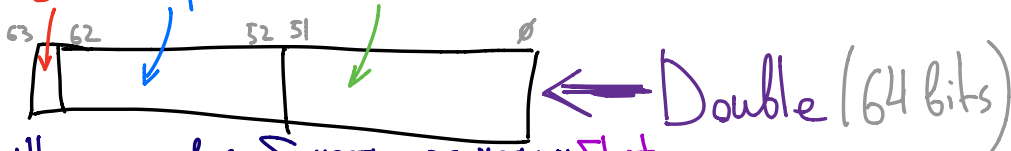*Не ограничивая общности, рассмотрим Float

#sign → 1
#exp → 8
#mant → 23

$bias = 2^{\#exp-1} - 1$
(отклонение)

0111 1111

"Виды" значений

Normalized
(остальные числа)

$0 < exp < 2^{\#exp} - 1$

$(-1)^{sign} \cdot \left(1 + \dfrac{M}{2^{\#mant}}\right) \cdot 2^{(exp-bias)}$

Denormalized
(малые числа)

0000000

$exp = 0$

$(-1)^{sign} \cdot \left(\dfrac{M}{2^{\#mant}}\right) \cdot 2^{(1-bias)}$

Special
(специальные числа)

11111111

$exp = 2^{\#exp} - 1$

$mant = 0 \Rightarrow (-1)^{sign} \cdot \infty$

$mant \ne 0 \Rightarrow NaN$

ЧТО ХРАНЯТ

ПРИЗНАК

Перевод в dec

Demo: 10^{10}
10^{-40}
-15
NaN, Inf
whatever

Примеры

#sign = 1
#exp = 4 ⟶ bias = 7
#mant = 5

* Min abs normalized? (>∅)

| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

$$(1+0)\cdot 2^{1-7} = 2^{-6}$$

* Max normalized?

| 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |

$$\left(1+\frac{31}{32}\right)\cdot 2^{14-7} = 2^7 + 31\cdot 2^2$$

* Min representable?
abs > ∅

| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

$$\frac{1}{32}\cdot 2^{-6} = 2^{-11}$$

* Max representable? (den.)

| 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |

$$2^{-11}\cdot 31$$

1475

Norm. **out of bounds**

* convert 147
↓ bin
X 0010011

$$7 = x - 7$$
+7

| ∅ | 1 | 1 | 1 | 0 | 0 | 1 | 0 | ? | ∅ |

...... .1
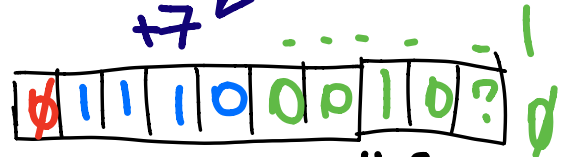
144 vs 148

* convert -14.75

X 110.11 $3 = x - 7$
10

| 1 | 1 | ∅ | 1 | 0 | 1 | 1 | 0 | 1 | 1 |

* convert 0.0117

*2¹¹

23.9616 ~ 24

0.00000 011000

*2⁶ (компенсация bias'a)

| ∅ | ∅ | ∅ | ∅ | ∅ | 1 | 1 | 0 | 0 | 0 |

("съедание")

✶ ассоциативность

$$(42 + 10^{10}) - 10^{10} = \emptyset$$
$$42 + (10^{10} - 10^{10}) = 42$$

Сломана

✶ дистрибутивность

$$(10^{20} - 10^{20}) \cdot 10^{20} = \emptyset$$
$$\underbrace{10^{20} \cdot 10^{20}}_{\infty} - \underbrace{10^{20} \cdot 10^{20}}_{\infty} = NaN$$

Сломана

Больше примеров и пояснений:

Computer Systems
A Programmer's Perspective
(Ch. 2)