

Convert Integer to Floating-Point

int 12,345 移位

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

11 0000 0011 1001 (14位)

$$12,345 = 1.1\ 0000\ 0011\ 1001 \times 2^{13} \quad \text{bias}_{float} = 127$$

0	1000	1100	1	0000	0011	1001	0000	0000	00
---	------	------	---	------	------	------	------	------	----

exp

(e)

frac

$$E = e - \text{bias}$$

$$e = 140$$

Rounding

For value

x

x'

$$x = 1.5$$

$$x' = 1 \text{ or } 2 ?$$

1.Round-to-even (向偶数舍入)

2.Round-toward-zero (向零舍入)

3.Round-down (向下舍入)

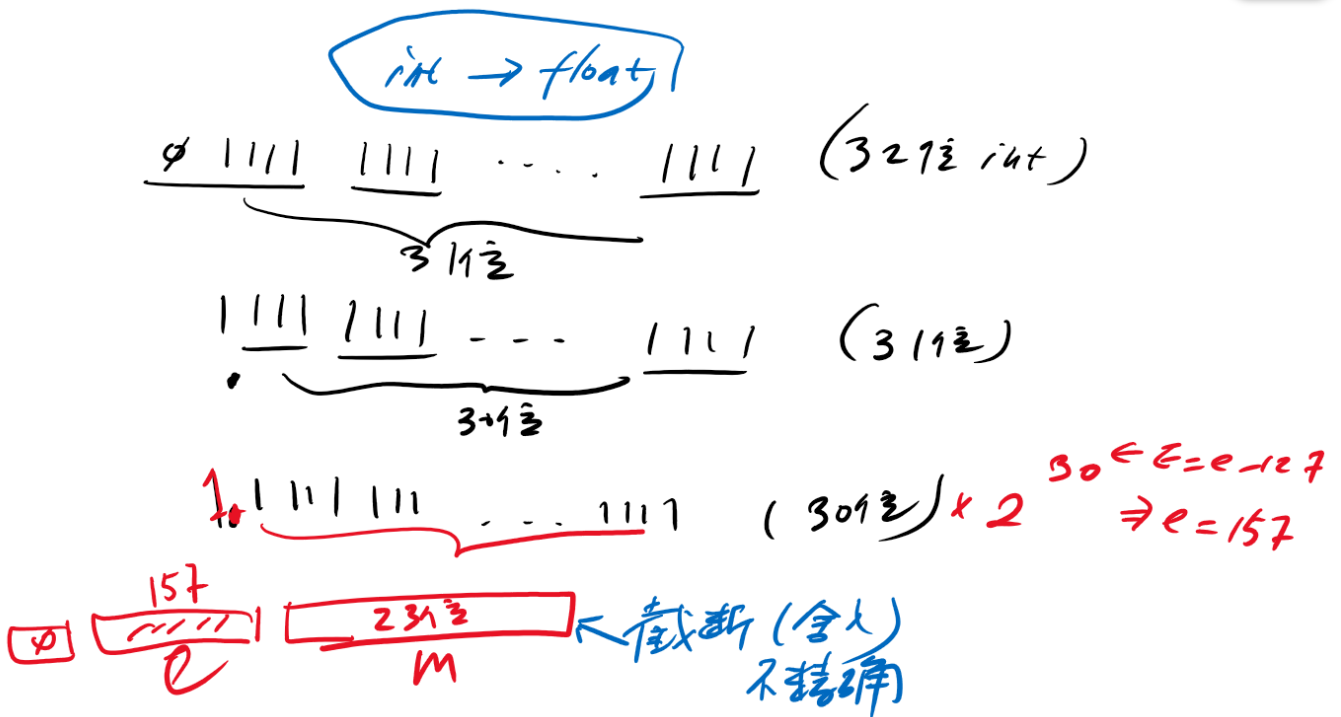
4.Round-up (向上舍入)

Rounding

Mode	1.40	1.60	1.50	2.50	-1.50
Round-down	1	1	1	2	-2
Round-up	2	2	2	3	-1
Round-toward-zero					



int => float转换，会存在精度问题，不存在溢出问题。



Floating Point in C

九曲園十

int \longleftrightarrow float \longleftrightarrow double

1. int \longrightarrow float (精度舍入)
2. int/float \longrightarrow double (OK)
3. double \longrightarrow float (溢出/精度舍入)
4. float/double \longrightarrow int (精度舍入/溢出
向0舍入)