$$\langle \cdots \cdots \cdots \cdots \cdots \rangle$$

$$-2^{31} \quad 0 \quad 2^{31}-1$$

Float — Single Precision

1999 754

32-bits

1-bit sign

23-bits

8-bits exponent

1.256 e23

$$1.256 \times 10^{23}$$

mantissa

$$-128 \cdots 127$$

1.25

32-bit   int

32-bit   float

1.25

half

-1 0)

float a, b , c

$a = v_1$

e

$b = v_2$

Π

$c = v_3$ Commutative

associative

$(a+b) + c$ Not necessarily

Not associative be equal to

$a + (b+c)$

float f = 5/4

1.25

a / b

1 ⟶ 1.000

5.0/4

5/4.0

$2.25/0.15$

0.9×0.9

15

14.9999

$0 \cdot 1$

0.0   0.81

0.1   0.8

0.2

0.9×0.5

0.45

0.9

0.4

0.5

arbitrary precision libraries

GMP

- Round to nearest

- Round to inf

- Rond to -inf

- Round to zero

fenv.h

$0 \cdot 81$

$0 \cdot 8$

$0 \cdot 81$
↓
$0 \cdot 9$