

CE 100 Accuracy of Computation

BASE 2

$$\begin{array}{r} \text{SIGN} \\ 1 \quad 0 \quad 1 \quad 1 \\ \hline \end{array}$$

-3

$$\begin{array}{r} \\ \\ + \\ \hline 1 \end{array}$$

2's complement

-2 110

+3 011

~~1001~~ +1

FLIP 1001

$$\begin{array}{r} \\ \\ + 1 \\ \hline 1 \end{array}$$

-6

BIAS

SHIFT OF RANGE
TO ALLOW EASY
COMPARISON

REAL

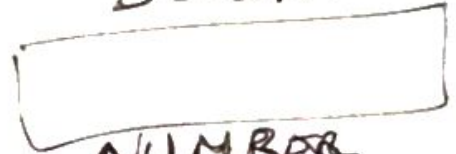
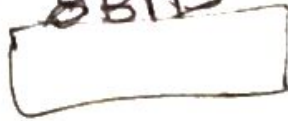
FLOAT

32 BITS

8 BITS EXP

23 BITS

1 SIGN



NUMBER

23.25

23.2

1 0 1 1 1 . 0 1

16 8 4 2 1 1/2 1/4

exp

1. _____ x 2⁴

1. 011101 x 2⁴

0 | 0 0 0 0 0 1 0 0
 4

011101

23 BITS

1. 0000.....

37994642.37

↑ .0000000000000001

Decimal value	Binary (two's-complement representation)
0	000
1	001
2	010
3	011
-4	100
-3	101
-2	110
-1	111

Type	Size in Bytes	Range
byte	1 byte	-128 to 127
short	2 bytes	-32,768 to 32,767
int	4 bytes	-2,147,483,648 to 2,147,483,647
long	8 bytes	-9,223,372,036,854,775,808 to 9,223,372,036,854,775,807
float	4 bytes	approximately ±3.40282347E+38F (6-7 significant decimal digits) <i>Java implements IEEE 754 standard</i>
double	8 bytes	approximately ±1.79769313486231570E+308 (15 significant decimal digits)
char	2 byte	0 to 65,536 (unsigned)