Number Representations

Primitive C++ Datatypes

Integers

1 Byte	char
2 Byte	short
4 Byte	int

Floats

4 Byte	float
8 Byte	double

4 Byte float format

S m s C

↑ ↑ ↑ ↖

1bit 23bits 1bit 7bits = 32 bits

Use Binary Base 2 scaled as

$\pm 0.mm\ldots m \times 2^{\pm cccccccc}$

↑ ↑ ↑ ↑

1bit 23bits 1bit 7bits

Example

$$8_{10} = 1000_2$$

has to be scaled to fit in to 4 byte float

$$+0.1 \times 2^{+4}$$

Note:

$$0.1 \times 2^4 = 1.0 \times 2^3$$
$$= 10.0 \times 2^2$$
$$= 100.0 \times 2^1$$
$$= 1000 \times 2^0$$

But when inserted into 4 Byte float

$$0.1 \times 2^4$$

All numbers for 4 Byte or 8 Byte need to be scaled so that mantissa has maximum number of bits free to be set.