

Number Representations

Primitive C++ Datatypes

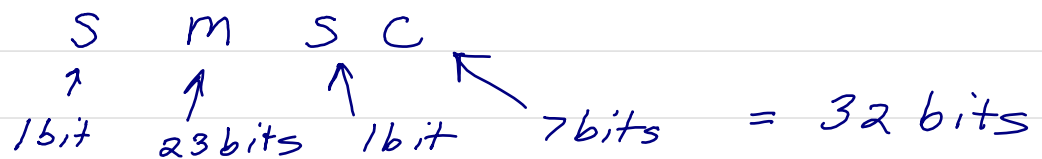
Integers

1 Byte	char
2 Byte	short
4 Byte	int

Floats

4 Byte	float
8 Byte	double

4 Byte float format



Use Binary Base 2 scaled as

$$\begin{array}{ccccccc} \pm & 0. & m & m, \dots m & \times 2^{\pm} & c & c & c & c & c & c & c \\ \uparrow & & \uparrow & & & \uparrow & & \uparrow & & & & \\ 1\text{bit} & & 23\text{bits} & & & 1\text{bit} & & 7\text{bits} & & & & \end{array}$$

MER

Example

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

has to be scaled to fit in to
4 byte float

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

Note:

$$\begin{aligned} 0.1 \times 2^4 &= 1.0 \times 2^3 \\ &= 10.0 \times 2^2 \\ &= 100.0 \times 2^1 \\ &= 1000 \times 2^0 \end{aligned}$$

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.