IEEE Floating Point Format

S	Exponent	Mantissa (fraction)

S – *Sign (0 for positive, 1 for negative)*

Exponent – *Power, base 2 – 8 bits for single, 11 for double*

Range – $\pm 2^{\pm 127}$, approx 5.87 x 10^{-39} to 1.7 x 10^{+38}, single

– $\pm 2^{\pm 1024}$, approx 4.78 x 10^{-308} to 8.98 x 10^{+307}, double

Mantissa – *Fraction, base 2, radix pt assumed to be to left of all the digits, 23 bits for single, 52 bits for double*

Accuracy, single – one part in 2^{24}, one part in 16 million, 5.96 x 10^{-8}, or a little better than 7 digits

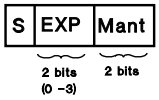
Accuracy, double – one part in 2^{53}, better than 16 digits

FLTPT020

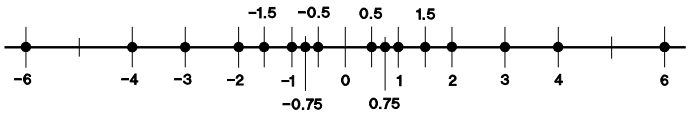
University of Idaho

Example Floating Point Value

For the following (very limited!) floating point representation:



The following values can represented:



FLTPT001

University of Idaho

Some Important Binary Values

2^{10} – 1024 or 1k, binary approximation of 1000
2^{20} – 1048576, or 1M, 1 meg, binary approx of 1 million
2^{30} – 1G, 1 gig, binary approx of 1 billion

(2^{7} – 1), 127, largest value that can be stored in a byte (signed)
(2^{8} – 1), 255, largest value that can be stored in a byte (unsigned)
(2^{15} – 1), 32767, largest value that can be stored in 16 bits (signed)
(2^{16} – 1), 65535, largest value that can be stored in 16 bits (unsigned)
(2^{31} – 1), 2 gig, largest value that can be stored in 32 bits (signed)
(2^{32} – 1), 4 gig, largest value that can be stored in 32 bits (unsigned)

C Simple Types

The C standard doesn't specify the internal representation that must be used for the simple data types. It says that an int should be the "usual" size of values used on a particular system, that short int should be no larger than ints, and that long int should be no shorter than int. The actual limits are included in <limits.h> (or <climits>) and <float.h> (or <cfloat>). Some typical values:

char – either signed or unsigned
signed char – 8 bit 2's complement, range –128 – +127
unsigned char – 8 bit unsigned, range 0 – 255

int – 32 bit 2's complement, range –2147483648 – +2147483647
unsigned int – 32 bit unsigned, range 0 – 4294967295
short int – 16 bit 2's complement, range –32768 – +32767
short unsigned int – 16 bit unsigned, range 0 – 65535
long int – 32 bit 2's complement, same as int
long signed int – 32 bit unsigned, same as unsigned int

float – IEEE floating point, range approx $\pm 1.17549 \times 10^{-38}$ – $\pm 3.40282 \times 10^{+38}$
double – IEEE double, range approx $\pm 2.22044 \times 10^{-308}$ – $\pm 1.79769 \times 10^{+308}$