IEEE Floating Point Format

<table>
<thead>
<tr>
<th>S</th>
<th>Exponent</th>
<th>Mantissa (fraction)</th>
</tr>
</thead>
</table>

S - Sign 0 for positive, 1 for negative

Exponent - Power, base 2 - 8 bits for single, 11 for double

- $2^{8}$, approx $0.87 \times 10^{38}$ to $1.7 \times 10^{38}$, single
- $2^{11}$, approx $4.79 \times 10^{38}$ to $9.98 \times 10^{38}$, double

Mantissa - Fraction, base 2, radix point 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, $0.96 \times 10^{-2}$, or a little better than 7 digits

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

Example Floating Point Value

For the following very limited floating point representation:

<table>
<thead>
<tr>
<th>S</th>
<th>Exp</th>
<th>Mant</th>
</tr>
</thead>
</table>

The following values can be represented:

- $-1.6$, $-0.8$, $0.2$, $1.6$
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

$(2^{31} - 1) = 2147483647$, largest value that can be stored in a byte (signed)  
$(2^{32} - 1) = 4294967295$, largest value that can be stored in 16 bits (signed)  
$(2^{32} - 1) = 65535$, largest value that can be stored in 16 bits (unsigned)  
$(2^{31} - 1) = 2G$, largest value that can be stored in 32 bits (signed)  
$(2^{32} - 1) = 4G$, 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 int, and that long int should be no shorter than int. The actual limits are included in `<limits.h>` (or `<limits>`) and `<float.h>` (for `<float.h>`). Some typical values:

- `char` - either signed or unsigned  
  `signed char` - 8 bit 2's complement, range $-128$ to $+127$  
  `unsigned char` - 8 bit unsigned, range $0$ to $255$

- `int` - 32 bit 2's complement, range $-2147483648$ to $2147483647$  
  `unsigned int` - 32 bit unsigned, range $0$ to $4294967295$

- `short int` - 16 bit 2's complement, range $-32768$ to $+32767$  
- `short unsigned int` - 16 bit unsigned, range $0$ to $65535$

- `long int` - 32 bit 2's complement, same as int  
- `long unsigned int` - 32 bit unsigned, same as unsigned int

- `float` - IEEE floating point, range approx $1.17549 \times 10^{-38}$ to $3.40282 \times 10^{38}$

- `double` - IEEE double, range approx $4.94067 \times 10^{-324}$ to $1.79769 \times 10^{308}$