Single precision floating point

Sign	Exponent (not 2's complement)	Mantissa (decimals)
1	8	23

Sign : 0 = positive
1 = negative

$2^{(8-1)} - 1$

Exponent : exp + 127

Mantissa : $1,\underbrace{101010}_{\text{only this}}$

Convert decimal to single precision float

ex: 45.45

1. Convert decimal to binary

$45 = 22 \cdot 2 + 1$
$22 = 11 \cdot 2 + 0$
$11 = 5 \cdot 2 + 1$
$5 = 2 \cdot 2 + 1$
$2 = 1 \cdot 2 + 0$
$1 = 0 \cdot 2 + 1$

$0,45 \times 2 = 0,9$
$0,9 \times 2 = 1,8$
$0,8 \times 2 = 1,6$
$0,6 \times 2 = 1,2$
$0,2 \times 2 = 0,4$
$0,4 \times 2 = 0,8$
$0,8 \times 2 = 1,6$

repeat

$101101.01\overline{1100}$

2. Normalize

$101101,01\overline{1100} = 1,\underbrace{0110101\overline{1100}}_{\text{Mantissa}} \times 2^5$

S = 0 (positive)

$E = 5 + \left(2^{(8-1)} - 1\right) = 5 + 127 = 135 = 1000\ 0100$

$M = 0110\ 101\ \overline{1100}$

3. Put numbers in

$\overset{S}{0}\ \underbrace{1000\ 0100}^{E}\ \overbrace{0110\ 1011\ 1001\ 1001\ 1001\ 100}^{M}$

Convert single precision float to decimal

ex: 1000 1000 1000 1000 1000 0000 0000 0000

1. Put numbers in their field

S = 1 (negative)

E = 0001 0001

M = 0001 0001 0000 0000 0000 0000

2. Convert exponent and mantissa

$E = 17 - (2^{(8-1)} - 1) = 17 - 127 = -110$

$M = 2^{-4} + 2^{-8} = 0{,}06640625$

3. Write the number

$-1{,}06640625 \times 10^{-110}$

Double Precision

Sign: 1 bit
Exponent: 11 bits
Mantissa: 52 bits

The conversion process is same as single precision but instead of adding 127, you add 1023 since $(2^{(11-1)} - 1) = 1023$