Floating Point

s	exp k	frac n

Normalized

$(-1)^s \times M = (1.\text{frac}) \times 2^{E}$, where $E = \text{Exp} - \text{bias}$ and $\text{bias} = 2^{k-1} - 1$

Denormalized

$(-1)^s \times M = (0.\text{frac}) \times 2^{E}$, where $E = 1 - \text{bias}$ and $\text{bias} = 2^{k-1} - 1$

★ overflow handled as $\pm\infty$

Special

↳ exp → all 1s

frac → 0s → ∞

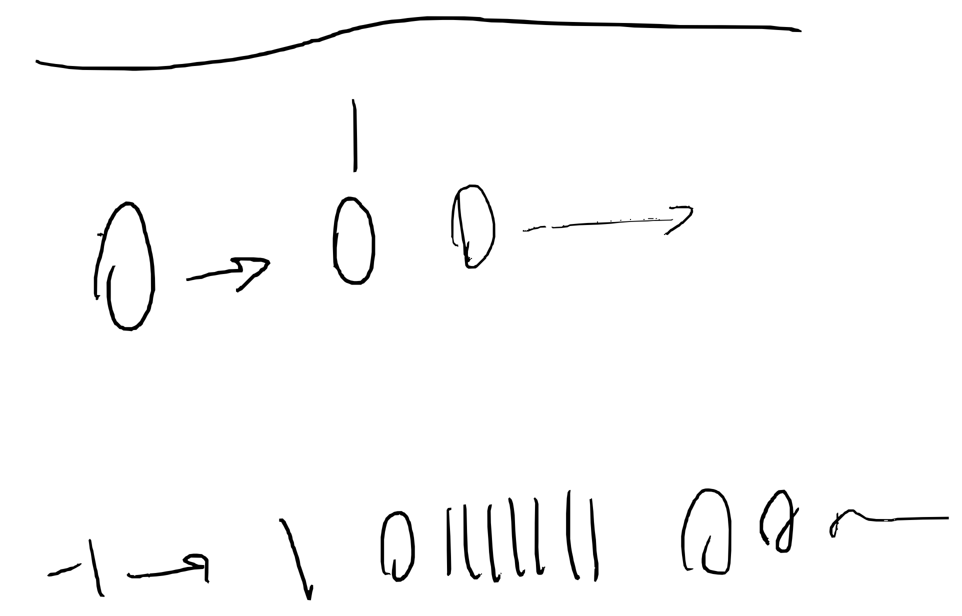
frac → non 0s → NaN

Max factionary → $1 - \frac{1}{2^N}$

Rang →

$-(2 - 2^{-N}) \times 2^{\text{bias}}$

$+(2 - 2^{N}) \times 2^{\text{bias}}$



362.1875

362 = 161 · 2 + 0
161 = 80 · 2 + 1
80 = 40 · 2 + 0
40 = 20 · 2 + 0
20 = 10 · 2 + 0
10 = 5 · 2 + 0
5 = 2 · 2 + 1
2 = 1 · 2 + 0
1 = 0 · 2 + 1

101000010

0.1875 · 2 = 0.375 → 0
0.375 · 2 = 0.75 → 0
0.75 · 2 = 1.5 → 1
0.5 · 2 = 1.0 → 1

0011

101000010.0011

1010000100011 · 2^{-8}

S = 0 119 127 + (-8)
64 + 32 13 (4) 8 E - exp - bias

0 0110 1101

01000010 → 0

0117.34

117 = 58 · 2 + 1
58 = 29 · 2 + 0
29 = 14 · 2 + 1
14 = 7 · 2 + 0
7 = 3 · 2 + 1
3 = 1 · 2 + 1
1 = 0 · 2 + 1

1110101 1

0 1 0 1 0 1 1 1 1 0 0 0

127 + 6
133
15

1110101 . 0101 011 0000 10100

8
0 | 1000 1111
1101010101

$(-1)^S \times M(1.\text{frac}) \times 2^{E(\text{exp} - \text{bias})}$
$(-1)^S \times M(0.\text{frac}) \times 2^{E(1 - \text{bias})}$

Addition

1. match exponent
   * manage in a way to always add (not subtract) and if necessary change the sign
   * or subtraction if not 0> result =
2. Add
3. Normalize
4. to fp rep if asked

−230.625

1

230 = 115·2 + 0
115 = 57·2 + 1
57 = 28·2 + 1
28 = 14·2 + 0
14 = 7·2 + 0
7 = 3·2 + 1
3 = 1·2 + 1
1 = 0·2 + 1

0.625·2 = 1
0.25 0
1

101

11100110.101

−1.1100110101·2^7

300.250

0

300 = 150·2 + 0
150 = 75·2 + 0
75 = 37·2 + 1
37 = 18·2 + 1
18 = 9·2 + 0
9 = 4·2 + 1
4 = 2·2 + 0
2 = 1·2 + 0
1 = 0·2 + 1

0
1

100101100.01

−1.1100110101·2^7

− 1.1100110101·2^7
1.0010110001·2^8

1.0010110010·2^8
− 0.11100110 101 ·2^8

1.1001 00101·2^7

24

7+128 134 128
134

0 1

400.625

400 = 200·2 + 0
200 = 100·2 + 0
100 = 50·2 + 0
50 = 25·2 + 0
25 = 12·2 + 1
12 = 6·2 + 0
6 = 3·2 + 0
3 = 1·2 + 1
1 = 0·2 + 1

0.625 × 2 = 1.250 1
0.250 ·2 = 0
1

110010000.101

127
8
135

+1.10010000101·2^8

200.250

135
128
2 7
5

11001000.01

+1.1001000001·2^7

0 10000111 (1) 1001000000101
− 0 10000111 (0) 11001000010

(10) 01011100111

101 27

0 10000110 (10) 001 0110001 11

Multiplication

1, add E's
2, multiply significands
3, convert to fp rep after!

Division

1, subtract E's
2, divide significands
3, convert to fp after!

Denormalized fp's

are just like normalized, in Arithmetic Point of view, except

1, E can't go lower that 1-bias
2, M < 1, if after arithmetic
M >= 1, reinterpret result as normalized!

if this violated
set E = 1-bias
and then shift M appropriately.

★ if E > -bias
then use denormalized.

● associative
int ✓
float X

1.10010000101×2^8

1.100100001×2^7 101

```
   1.1001 0000 101
 × 1.1001 0000 100
 -----------------------
          11001000010100
       1100100001010000000
      11001000010100000000
    1100100001010000000000000
```