5 - 6 (5 bit)

```
  00101   (5)
+ 11010   (-6)
-------
  11111   (-1)
```

6

```
  00110
  11001
     +1
  11010   (-6)
```

```
  11111
  00101   5
+ 11111   -1
-------
 100100   4
```

carry ~~1~~

```
  11
  01100   12
+ 00110   +6
-------
  10010   -14
```

overflow

(-14)

```
  10010
  01101
     +1
  01110
```

negate

14

Real numbers

5 bits fixed point

3 bits integer
2 bits fraction

$2^2\ 2^1\ 2^0\ 2^{-1}\ 2^{-2}$

1 0 1 . 0 1

5.25

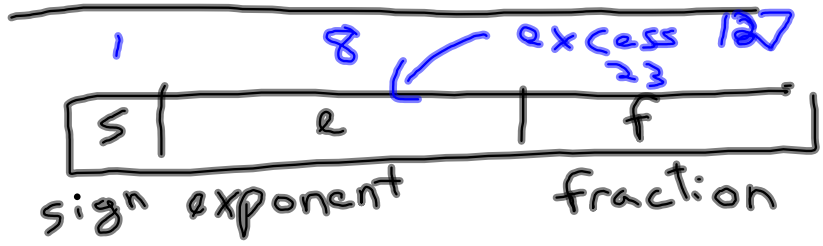
$1 \times 2^2 = 4$
0×2^1
$1 \times 2^0 = 1$
0×2^{-1}
$1 \times 2^{-2} = \frac{1}{2^2} = \frac{1}{4}$

(0, .25, .5, .75)

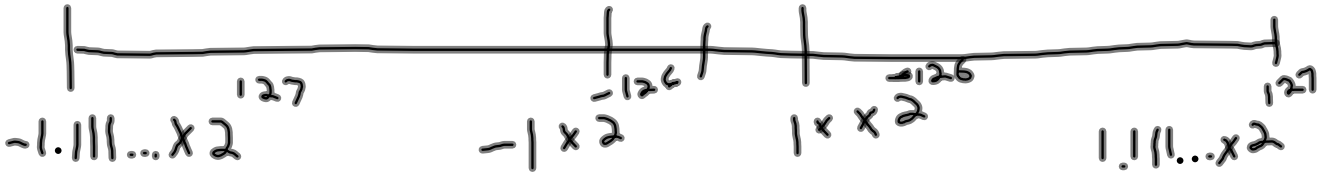
Floating Point

$0.003 \qquad 3.0 \times 10^{-3}$

$50 \qquad 5.0 \times 10^{1}$



$$v = (-1)^s \times 1.f \times 2^e$$

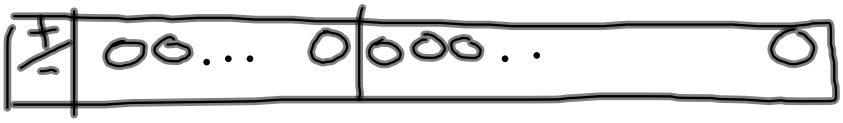


format

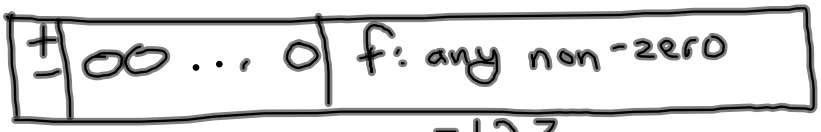
Normalized

$\pm 1.f \times 2^{e}$

Zero

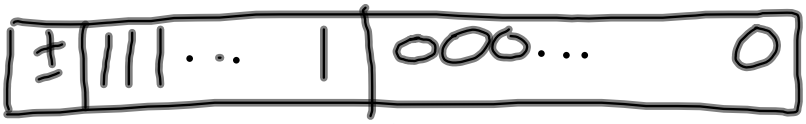


Denormalized



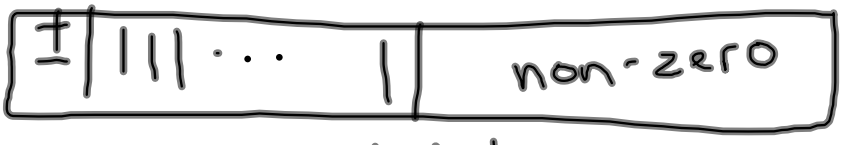
$0.f \times 2^{-127}$

Infinity



$\pm\infty$

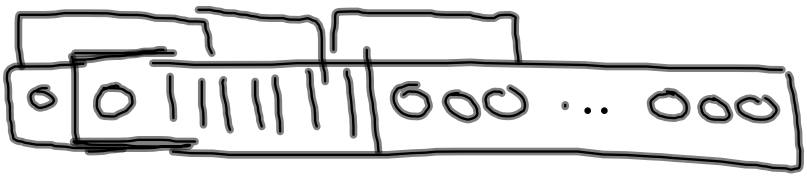
Not a number



NaN

	float	double
sign	1	1
exp	8	11
frac	23	52
exp system	Excess 127	Excess 1023
Decimal Range	$\sim 10^{-38}$ to 10^{38}	$\sim 10^{-308}$ to 10^{308}
Smallest denorm	10^{-45}	10^{-324}

Decimal 1



$1_{10} = +1.0 \times 2^{0}$

0x3F800000

0x42E48000

0100 0010 1110 0100 1000 0000.

s e f

e = 133 - 127
= 6

$+1.11001001 \times 2^{6}$

1110010.01000000..

1
64
+32
+16
+ 2

114

+114.25

0x80400000

1000 0000 0100 0000 0000 0000...

- | -127 | base 2

-0.1×2^{-127}

$= -0.5 \times 2^{-127}$

Denormalized

0x7F8E1000

0111 0101 1000 1110 0001 0000...

X (nevermind)

0x7F8E1000

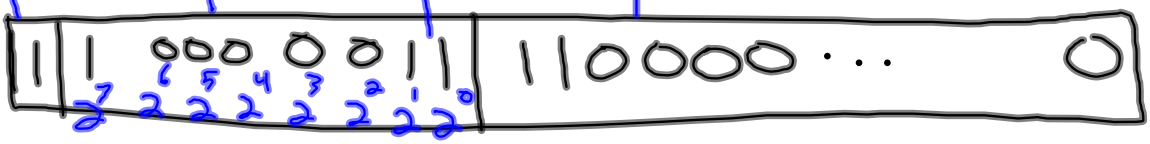
0111 1111 1000 1110 0001 0000...

+ | e | f

NAN

-3.5×2^3

C 1 E 0 0 ... 0



exp $4 + 127 = 131$ convert unsigned

$f - 3.5_{10} = 11.1_2$

$(1.11 \times 2^1) \times 2^3$

1.11×2^4

0xC1E00000