2-1

$$UZ(B) = \sum_{i=0}^{n} 2^i \times B_i$$

$$\underbrace{xxxx}_{m} \, . \, \underbrace{yyyy}_{-n}$$

$$FZ(B) = \sum_{i=-n}^{m} 2^i \times B_i$$

0010.1011

$2^1 + 2^{-1} + 2^{-3} + 2^{-4}$

$2 + 1/2 + 1/8 + 1/16$

$= 2.6875$

fixed-point binary number

2.5

2.0

10111 $\overset{mod}{\Rightarrow}$ 0111   01111 $\overset{mod}{\Rightarrow}$ 1000

10111 $\overset{sat}{\Rightarrow}$ 1111   01111 $\overset{trunc}{\Rightarrow}$ 0111

N = sign * magnitude, exponent

s   m   e

$(-1)^s \times m \times 2^e$

$(s_1, m_1, e_1) * (s_2, m_2, e_2)$ — $s_1 \in (0,1)$, m_1: a, e_1: b

$= (s_1 \times s_2, \; m_1 \times m_2, \; e_1 + e_2)$

$(-1)^{s_1} (m_1)(2^{e_1}) \times (-1)^{s_2} (m_2)(2^{e_2})$

$m_1 \times m_2 \to 2a$, $e_1 + e_2 \to b+1$

32-bit ~ 18 trillion

~9 trillion positive

4.5 tril $\in [0,1]$

2-2

MSB LSB

sign	exponent	1 .	Mantissa
1 bit	8 bits		23 bits

float32

$(-1)^s \times (1 . M_0 \ldots M_{22}) \times 2^{U(E) - 127}$

0 1000 0001 11 00 0 ——

+ 129 − 127 = 2 1 + 1/2 + 1/4 = 1.75 = ~~3~~

$2^2 = 4$ = 7

$2^7 + 2^0 = 128 + 1 = 129$

exp ≠ 0 & ≠ 255 ⇒ normalized

exp = 0 ⇒ de-normalized

no hidden bit, E = 1 − field value

exp = 255 (all 1s) → mantissa = 0 ⇒ infinity

→ ≠ 0 ⇒ NaN

+ 88.00 1, 2, 4, 8, 16, 32, 64 88 − 64 = 24 − 16 = 8

~~64 0 1 0 1 000~~ $88 = (-1)^s \times (1+m) \times 2^E$ = 6

01011000 1. . . . $\times 2^E$ 88 / 64

1.375

= 1.011

1.375
64 ⟌ 88
− 64
240
~~192~~
480
448
320

+ 6 0.375

0 1000 0101 0110 ——

---

0 1000 0111 001 1000 0001 0 ——

304.125