Number systems

Decimal → base 10
↳ digits {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}

Example of computation of decimal numbers:

524.503

$= 5\times10^{2} + 2\times10^{1} + 4\times10^{0} + 5\times10^{-1} + 0\times10^{-2} + 3\times10^{-3}$

base 10

Binary → base 2
↳ digits {0, 1}

Example of binary computation:

101.011

$= 1\times2^{2} + 0\times2^{1} + 1\times2^{0} + 0\times2^{-1} + 1\times2^{-2} + 1\times2^{-3}$

base 2

Repeating numbers

decimal system examples: $\frac{1}{3} = 0.3333 = 0.\overline{3}$

$\frac{1}{7} = 0.142857142857$

$= 0.\overline{142857}$

binary system examples: $\frac{1}{3} = 0.\overline{01}$

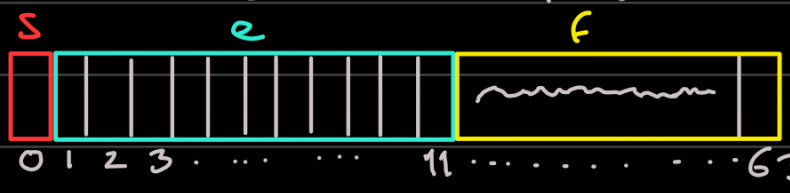
$\frac{1}{7} = 0.001001\ldots$

$= 0.\overline{001}$

These repeating numbers repeat indefinitely and are shortened with a bar over their heads

Double precision

The IEEE arithmatic standard:



formula to calculate #s

$= (-1)^{S} \times 2^{e-1023} \times f$

S: sign bit (either a 0 or 1)

e: biased exponent of 2 (contains 11 bits; first position bit, i.e bit 1 gives e = 2048)

f: significand (52 bits)

IEE standard also has some special reserved numbers that are represented as:

(i) e is all 1's and
f = 0 we say the number is Inf (infinity)
or, f ≠ 0 we say the number is NaN (not a number)

(ii) e is all 0's, then 1.f becomes 0.f (graceful underflow)

Calculating realmax/realmin

For realmax e is almost all 1's (last e = 0 in bit position 11 to avoid infinity).

$e \Longrightarrow 11111111110$

$\therefore e = 1024+512+256+128+64+32+16+8+4+2+0$

or, $e = 2^{11} - 1 - 1$ [gives the sum of first 10 bits
subtracts the 1 from the last bit position since the bit is a 0]

and all the f's (all the 52 bits) are 1 ⟹

$$f = \overset{1}{\frac{1}{2}} + \overset{2}{\frac{1}{4}} \cdots + \overset{51}{\frac{1}{2^{51}}} + \overset{52}{\frac{1}{2^{52}}} + 1 \; ?$$

or, $f = 1 + \left(1 - \frac{1}{2^{52}}\right)$ sums upto bit position 52

f starts with bit position 0 as well?

the s = 0 (+ve number)

$$\text{realmax} = (-1)^0 \times 2^{2046-1023} \times (2 - 2^{-52})$$

$$= 1.7977 e^{308}$$

Important

To understand why $f = \frac{1}{2} + \frac{1}{4} \cdots + \frac{1}{2^{51}} + \frac{1}{2^{52}}$ is equal to $1 - \frac{1}{2^{52}}$

assume we want first 3 decimal places of the binary system:

$f = \frac{1}{2} + \frac{1}{4} + \frac{1}{8}$

or, $f = 2^{-3}(4+2+1)$

or $f = 2^{-3}(2^3 - 1)$

or $f = 1 - \frac{1}{8} = 1 - \frac{1}{2^3}$

Double precision number formats

$(1)_{10} = (-1)^0 \times 2e^{1023-1023} \times 1.0$ (both sides are in decimal and should equal)

binary for $e \to 1023 = 2^{10} - 1$

or $(1023)_{10} = (1000000000)_2 - (0000000000 1)_2$

$= (01111111111)_2$

∴ The binary/computer format of 1 is 0011111111110000...0 (52 f's all 0)

$\left(\frac{1}{2}\right)_{10} = (-1)^0 \times 2e^{1022-1023} \times 1.0$

binary for $e \to 1022 = 2^9 + 2^8 + \ldots 2^1$

$\therefore e = (01111111110)_2$

computer representation of $\frac{1}{2} = (0011111111100\cdots0)_2$ (52 f's all 0)

$\frac{1}{3} = \frac{1}{4}\left(1 + \frac{1}{3}\right)$

$\left(\frac{1}{3}\right)_{10} = (-1)^0 \times 2e^{1021-1023} \times 1.\frac{1}{3}$

$\therefore e \to 1021 = 2^9 + 2^8 + \ldots 2^0 - 2 = (01111111101)_2$

$f \to \frac{1}{3} = 0.\overline{01}$

$\therefore \left(\frac{1}{3}\right)_{10} = (0011111111101010101010\ldots01)_{10}$

Signed vs. Unsigned

Unsigned ⟶ +ve numbers

binary example : $(1000\ 1101)_2 = (141)_{10}$

signed ⟶ accounting for -ve numbers

(0 for +ve / 1 for -ve)

example : 0100 0110 $= (70)_{10}$ +ve

1000 1101

$= -1\times2^7 + 0\times2^6 + 0\times2^5 + 0\times2^4 + 1\times2^3 + 1\times2^2 + 0\times2^1 + 1\times1$

$= (-115)$ -ve

Word ⟶ 32 bits long

half word ⟶ 16 bits long

Binary to unsigned

$(-100)_{10}$

i) Convert 100 to binary : $(01100100)_2$

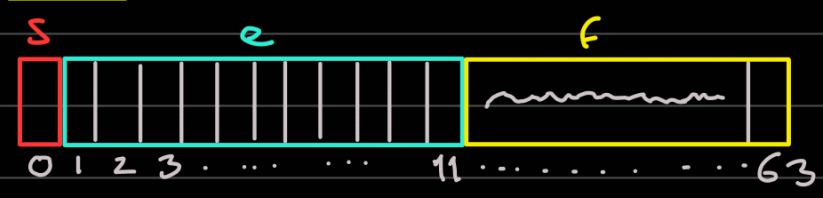
recall → 1+0 = 1, → 1+1 = 10

ii) Take 2's complement : 1001 1011 flipping bits

\+ 1 adding 1

1001 1100 ⟶ -100 in binary

Eps



1 is represented as : $(-1)^0 \times 2^0 \times 1.0$

The next largest number is : $(-1)^0 \times 2^0 \times (1 + 2^{-52})$ [last bit of f is a 1]

$= (1 + 2^{-52})$

machine eps $= (1 + 2^{-52}) - 1 = 2^{-52}$

In a similar fashion :

next largest number after 2 : $(-1)^0 \times 2^1 \times (1 + 2^{-52}) = 2(1 + 2^{-52})$

∴ difference between 2 and next largest number

$= 2(1 + 2^{-52}) - 2 = 2 \times 2^{-52}$