CS 314 - Lecture 6

(Section 5.3 in book)

Computers store bits

$\boxed{1\ 0\ 0\ 1\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 1}$

$5 \xrightarrow{\text{binary}} \overset{2^2}{\boxed{1}}\overset{2^1}{\boxed{0}}\overset{2^0}{\boxed{1}}$

$+1 \cdot 2^0$

$0 \cdot 2^1$

$+$

$1 \cdot 2^2$

??

$1/3 \to$??

Floats are like Scientific notation

1.5×10^3

-2.7135×10^2

6.67×10^{-11}

2.998×10^8

A float consists of

- a sign
- Exponent
- "fraction" / "number"
  Mantissa

A simple system of floats

- 1 bit for the sign
- 2 bits for the exponent
- 2 bits for the mantissa

$\approx 2^5$ floats.

$\boxed{S}$ $\boxed{\text{mantissa}}$

$(-1)^S \cdot (1.b_1 b_2)_2$

1 → 0 00

$(-1)^0 \cdot 1.00 \cdot$

2 → 0 00

$(-1)^0 \cdot (1.00)_2 \cdot$

1.5 → 0 10

$(-1)^0 \cdot (1.10)_2 \cdot$

Aside we'll store 4 ex

0, 1, -1, ∅

Exp

$\cdot 2^{e}$

0

$2^{0} = \underline{1}$

$\underline{1}$

$2^{1} = 2$

$~~1~~0$

$2^{0} =$

Base 10

$$512.23$$

$$= 5 \cdot 10^{2} + 1 \cdot 10^{1} + 2 \cdot 10^{0} + 2 \cdot 10^{-1} + 3 \cdot 10^{-2}$$

Base 2

$$(101.01)_2$$

$$1 \cdot 2^{2} + 0 \cdot 2^{1} + 1 \cdot 2^{0} + 0 \cdot 2^{-1} + 1 \cdot 2^{-2} = 5.25$$

$$(1.10)_2 = 1 \cdot 2^{0} + 1 \cdot 2^{-1} + 0 \cdot 2^{-2}$$

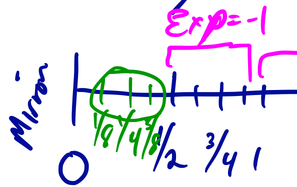
$$= 1.5$$

	S	Mantissa	Exp
0.75 →	0	10	-1
		$(1.10)_2$	$\cdot 2^{-1}$

Quiz write out float for $\boxed{-3}$

S	Mant	Exp
1	10	1
$(-1)^1$	$(1.10)_2$	2^1

Floating pt System



1/2 → 0 00

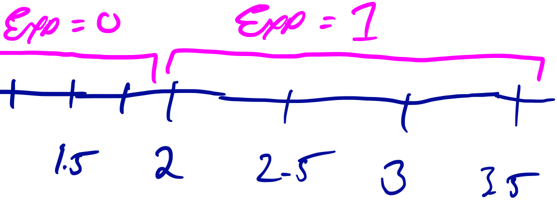
0 → 0 00

if exp = 0 and

then the # is

$$\boxed{(-1)^2 \cdot 0.b_1b_2}$$

Subnormal #s



-1

0

mantissa $\neq 0$

$)_2 \cdot 2^{-1} \longrightarrow$

IEEE floating

- a precise way to repr floats

Formats

Float64 (double)

Float32 (single)

pt.

, Mathematic

al + Manipulat

64-bits
1 sign bit
11 bits for Exponent
52 bits for Mantissa

32-bits
1 bit for Sign
8 bits for Exp
23 bits for Mantissa