IEEE754 Floating point representation

Bias.

Biased Exponent.

Decimal to Floating point.

Floating point to Decimal.

Floating point Addition, sub.

Floating point mul.

Overflow - Underflow.

Chapter 3 (video 1)

$12 \rightarrow X_{21}.$

addi X_{21}, X_0, 12

$12.5 \times 10^{3} \rightarrow f_{21}$ How?

$\rightarrow f_0 - f_{31}$

$\underbrace{4.5}_{\text{co-eff}} \times \underbrace{10}_{\text{Base}}^{6}$ → Exponent

4.5×10^{-3}

IEEE 754

single precision (32-bit)

double " (64 bit)

Die

Normalized

Only one digit before the decimal / bin point

Digit must be a non-zero ~~m~~ number.

(1) ~~1.2~~ 123. 456 ✗

↳ 3 digit

(2) 1. 2345 ✓

↳ 1 digit, != 0

(3) 0. 123 ✗

↳ 1 digit, it is zero.

(4) 9.123 ~~✗~~ ✓

↳ 1 digit, it's not zero

(I) 1.0111 ✓

(II) 1101. 1000 ✗

↳ not one.

(III) 0.11101 ✗

↳ It's zero

Steps to normalize.

To normalize a number you need to shift the decimal point left or right, until there is only one single digit

~~123.45~~

~~123~~

12 34.5678

1.2345678

11010.00110

1.101000110

if you left-shift, the number of times you left shift will be added to the exponent

112.54×10^{35}

$\hookrightarrow 1.1254 \times 10^{(35+2)} \Rightarrow 1.125 \times 10^{37}$

456.123.

$\hookrightarrow 4.56123 \times 10^{2}$.

If you shift right, the number of times you right shift will be substracted from the exponent

0.0065

$\Rightarrow 6.5 \times 10^{-3}$

* 110110.1101_2

$\rightarrow 110110.1101 \times 2^0$

$\rightarrow 1.10110 1101 \times 2^5$.

$(0.000001 0101)_2$

$\Rightarrow 1.0101 \times 2^{-5}$

IEEE Floating point format.

	Sig. Bit	Expon.	Fraction.
(32) single p.	1bit	8 bit	23 bit
Double P. (64)	1bit.	11 bit	52 bit.

Single precision. (32bit)

Example :

numbe...

01011×2^{30} → 6bit

$1.1010111101011 \times 2^{30}$ → 12 bit.

Exponent bit will be unsigned → Using Bias.

→ 0 - 255.

but ~~0 an 25~~ 0 and 255 is reserved.

∴ range for bias ⇒ 1 - 254.

Double presion.

11 bit unsigned.

$0 - 2^{11} - 1$

⇒ 0 - 2048

but 0 and 2047 are reserved.

So, the range.

1 to 2047

Bias ⇒ $(2^{8-1}) - 1$

⇒ 127

Actual (1-127) to (254 - 127)

(-126) to ~~(126)~~ (127)

Bias & Bias exponent.

$$\text{Bias} = 2^{n-1} - 1$$

n → size of the exponent field.

Biased exponent = Actual expo + Bias.

Ex:

1.1011×2^{34} → find the biased expo in IEEE 754 single precision format.

34 → Actual exponent

Single precision format → Expo → 8 bits.

$\text{Bias} = 2^{8-1} - 1$
$= 127$

① Check if the num is normalized.

$\therefore$ Biased exponent $= (34 + 127)$
$= 161.$

Ex: S.P.F → 11.1011×2^{-8}.

① The number is not normalized.

→ 11.1011×2^{-8}.

→ $1.11011 \times 2^{-8+1}$ ⇒

⇒ 1.11011×2^{-7}.

$\therefore$ Biased exponent $= (-7 + 127) = 120$

11.1011×2^{-8}, Find the biased exponent of given number in 15-bit IEEE754 format where size of fraction is 8 bits.

fraction = 8 bit

sign-bit = 1 bit

9 bit $\therefore$ Exponent = $(15-9) = 6$ bits.

$\therefore$ Bias = $2^{6-1} - 1 = 31$

$\therefore$ ~~Biased e~~

number = 11.1011×2^{-8}

$\hookrightarrow 1.11011 \times 2^{-7}$

$\therefore$ Biased exponent = $(31 - 7) = 24$ ~~bits~~.

Decimal to Floating point.

(I) Convert to binary
(II) Normalize
(III) Biased exponent ✓
(IV) sign - bit.
(V) Find the fraction.

Ex

Convert 50.6749_{10} to 32 bit IEEE-754 floating point representation.

6486

$(50.6749)_{10} = 110010.1010110011$

Normalize.

$1.\underline{100101010110011} \times 2^5$

Bias = 127.

Biased exponent = $127 + 5 = (132)_{10} = 1000\ 0100$

sign-bit = 0 ✓

~~fan~~ fraction = $\underbrace{100101010110011}_{15\text{ bits}}\ \underbrace{0000\ 0000}_{8\text{ bits.}}$

0	1000 0100	10010101011001100000000

nvert -0.0232 to 12 bit IEEE-754 loating point where biased exp is 4 bits.

14848

$\therefore$ sign = 1
bias = 4

$$\frac{\text{bias} = 4}{5}$$ $\therefore$ fraction = 7 bit.

$$(0.0232)_{10} = \frac{1(110111)_2}{2^{10}}$$

$$= (-0.00\ 00010)_2$$

normalize, 1.0×2^{-6}.

$$\text{bias} = 2^3 - 1$$
$$= 7$$

$\therefore$ biased exponent $= 7 - 6 = 1, = 0001$.

sign = 1.

$\therefore$

1	0000110	0000 000

Floating point to Decimal.

(I) Hex to Binary
(II) Arrange the binary acc. to format.
(III) Determine the sign.
(IV) Find out actual exponent.
(V) Conv. fraction to decimal.
(VI) Decimal number $= (-1)^{\text{sign bit}} \times (1+\text{Fraction}) \times 2^{(\text{Exponent})}$

0xF2400120 → Single Pre. Float point.

$\boxed{1}$ $\boxed{111 \quad 0010 \quad 0}$100 0000
0000 0001 0010 0000.

Exp

fraction.

4194592

sign bit = 1 = negative number.

b. exponent = 111000100. = 228.

bias = 127.

∴ actual exponent = 228 − 127
= 101.

fraction = 100 0000 0000 0001 0010 0000.

= 0.100 0000 0000 0001 0010 0000

= 0.5000343323.

$\text{Dec} = (-1)^1 \times (1+0.5000343323) \times (2)^{101}$

$\Rightarrow -3.803038843 \times 10^{30}$

nd it upto 6 dec. point.

$-3.80303884 \times 10^{30}$

$\Rightarrow -3.80304 \times 10^{30}$

Floating point addition.

- → Numbers are in binary.
- → Normalize A & B.
- → Align the bin point so that lower exp match the hig
- → Now add / sub. 102389
- → Normalize the result.
- → Round if necessary.
- → Overflow / underflow.

$(9.999 \times 10^{1}) + (1.610 \times 10^{-1})$.

$\Rightarrow (99.99) + (0.1610)$

$\Rightarrow 1100011111111\ 0101 + 0.00\ 1010\ 0100$ (कम) exp

$\Rightarrow (1.1000\ 1111\ 1111\ 0101 \times 2^{6}) + (1.0100100 \times 2^{-3})$

$\Rightarrow (1.1000\ 1111\ 1111\ 0101 \times 2^{6}) + 0.0000\ 0000\ 1010\ 100 \times 2^{6}$.

$\Rightarrow 2^{6}(\quad + \quad)$.

1.1000 1111 1111 0101
0.0000 0000 1010 0100

1.1001 0000 1001 1001

1.1000 1111 1111 0101
0.0000 0000 1010 0100

1.1001 0000 1001 1001

Floating point mul.

→ Make sure that the values are in binary.

→ Normalize. → Add the expo.

→ Mutiply

→ Normalize. → If it was negative

→ same.

$(1.110 \times 2^5) \times (11.111 \times 2^{-7})$

$= (1.110 \times 2^5) \times (1.1111 \times 2^{-5})$

$= (1.\underline{110} \times 1.\underline{1111}) \times (2^{5-5})$

→ 1110×11111 (cacu)

$= 11.011\ 0010$

→ $11.011\ 0010 \times 2^0$

$= 1.1011\ 0010 \times 2^1$

→ $-1.1011\ 0010 \times 2^1$.

Overflow / underflow

nge = (x to y).

lower upper.

a is a number.

if ($x <= a <= y$):

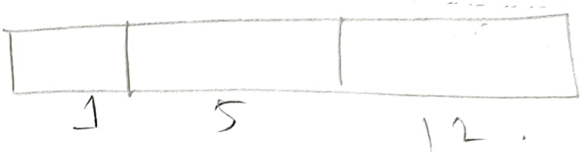
perfect

if ($a < x$):

under.

if ($y < a$):

overflow.



$\boxed{\text{Exponent field.}}$

$1.01 \times 2^{-89} \times 1.01 \times 2^{-89}$

$= (1.01 \times 1.01) \times 2^{-178}$

$= 1.1001 \times 2^{-178}$ → actual.

bias → $2^{5-1} - 1 = 15$.

biased exponent = $-178 + 15 = -163$

range of biased exponent,

$0 - 2^5 - 1$

$\Rightarrow 0 - 31$

range → $(1 - 30)$

∴ ~~overflow~~ underflow.

Long Multiplication

$$
\begin{array}{r}
1000 \\
1001 \\
\hline
1000 \\
0000\text{X} \\
0000\text{XX} \\
1000\text{XXX} \\
\hline
1001000
\end{array}
$$

Optimize multiplier.

iteration =< multiplier bits,

1 multiplier last bit 1

→ Multiplicand + product half MSB. A

→ 1 bit right shift (Product) B

or. multiplier last bit = 0.

→ 1 bit right shift (product) C

Product + multiplier.

1000 X ~~1001~~

Ite	Multiplicand	Product
0	1000	0000 1001
1	1000	1000 1001 A 0100 0100 B
2	1000	0010 0010 C
3	1000	0001 0001 C
	1000.	1001 0001 A 0100 1000. B