Representation of Floating-Point Number

Dr. Md. Nawab Yousuf Ali
Professor, Dept. of CSE
East West University

Floating Point Number: Contains two parts:

(i) Mantissa: Contains signed fixed-point number

(ii) Exponent: Specify the position of decimal (binary) point

$$\pm m \times r^{\pm e}$$

where, $\pm$ = Sign

m = mantissa

r = radix (base)

e = exponent

Examples:

$+215.37 = + 0.21537 \times 10^{+3}$

Where + = Sign

21537 = Mantissa

10 = Base (radix)

3 = exponent

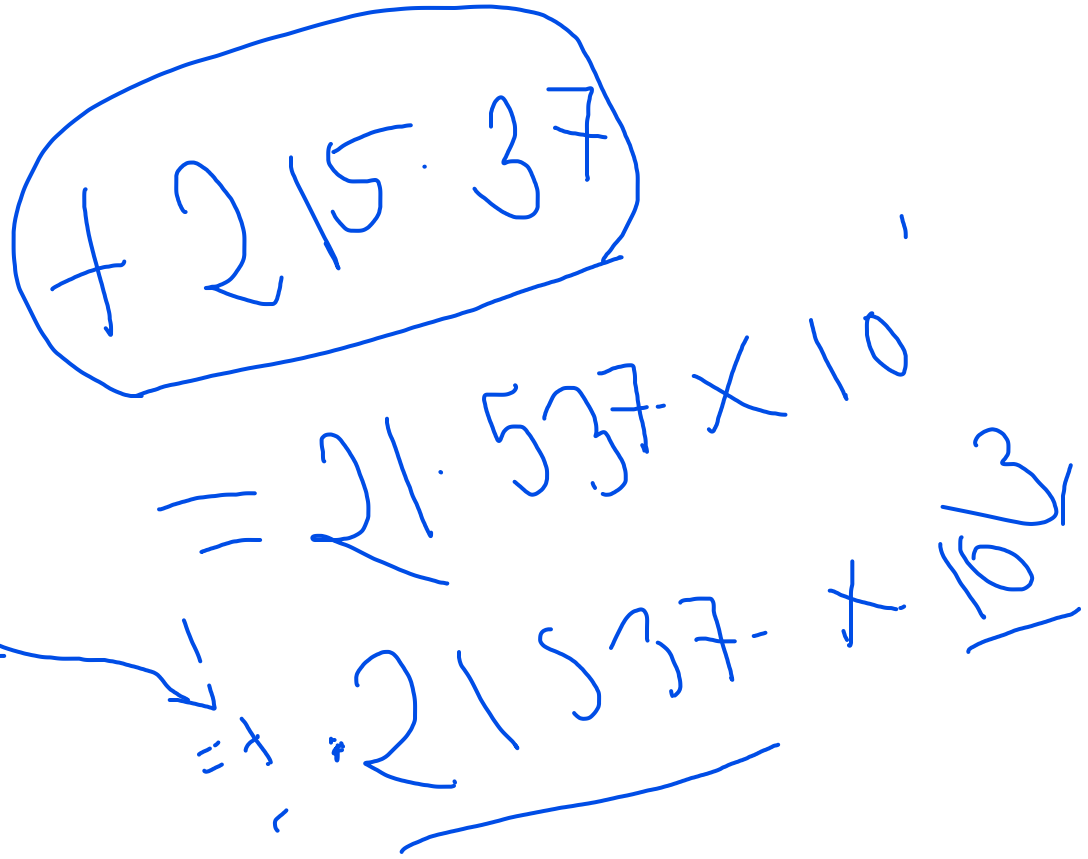
$+1000.110 = + 0.1000110 \times 2^{+4}$

+ = sign

1000110 = mantissa

2 = base (radix)

4 = exponent



$-101011 = -0.101011 \times 2^{+6}$

Here - = sign

101011 = mantissa

2 = base

6 = exponent

These numbers are called floating point numbers because the position of the decimal (binary) point is fluctuating (floating).

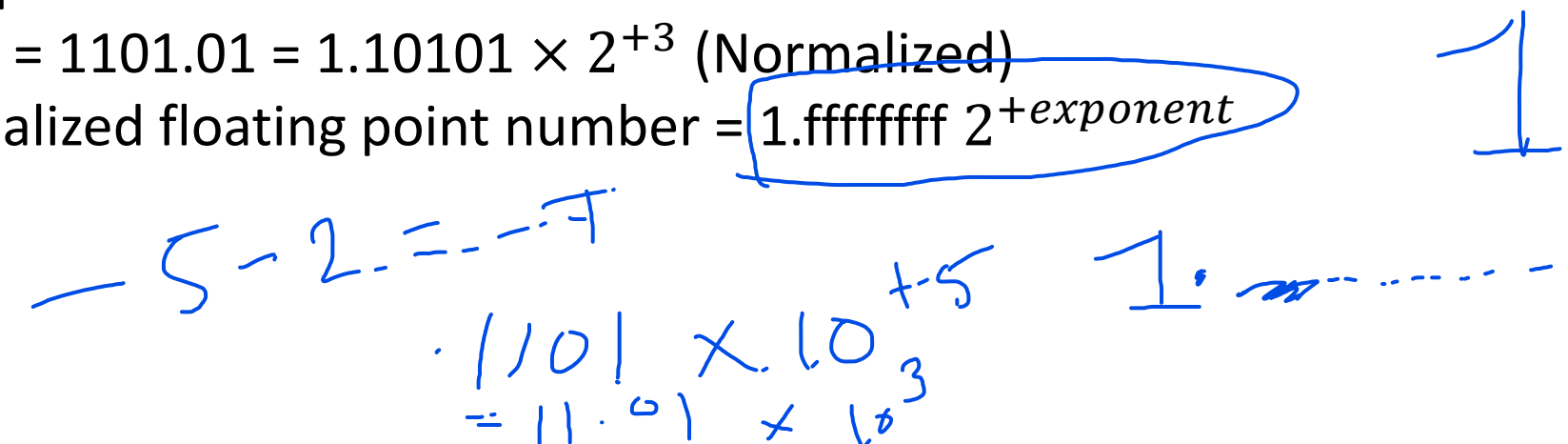
Normalization:

A floating-point number is said to be normalized, when we force the integer part of the mantissa to be 1 and allow its fractional part to anything.

Example:

$13.25 = 1101.01 = 1.10101 \times 2^{+3}$ (Normalized)

Normalized floating point number = $1.fffffff\ 2^{+exponent}$



IEEE 754 floating point representation

Mantissa positive → 0 Mantissa negative → 1

1. Single Precision floating point standard:

1 bit	8 bit	23 bit
Sign bit (mantissa)	Biased exponent bits (e+127)	Mantissa (Normalized)

Total 32 bits

(2) Double Precision floating point standard:

1 bit	11 bit	52 bit
Sign bit	Biased exponent bits (e+1023)	Mantissa (Normalized)

Total 64 bits

Example: Represent 85.125 in IEEE 754 single precision floating point representation

85.125 = 1010101.001

= $1.010101001 \times 2^{+6}$

Number is positive (+ve), so sign bit = 0

Exponent (e) = 6

Biased exponent = 6 + $(2^{8-1} - 1)$ = 6+127 = 133 = 10000101

Mantissa (Normalized) = 010101001 (Make of 23 bits)

= 01010100100000000000000 (23 bit)

1 bit	**8 bit**	**23 bit**
0	10000101	01010100100000000000000

Overflow/ Underflow in floating point numbers

(1) Significand Overflow:

When two similar sign significand (mantissa) are added, the sum may result in a carry out of the MSB. This is called significand overflow.

Ex:

$1.101 \times 2^{+3}$

$1.110 \times 2^{+3}$

$11.011 \times 2^{+3}$ → Corrected by shifting significand one position right

$1.1011 \times 2^{+4}$

(2) **Significand underflow:**
If the resulting value has 0 in the most significant position of the significand.
This is called significand underflow.

Ex:

$1.101 \times 2^{+5}$

$-1.100 \times 2^{+5}$

$0.001 \times 2^{+5}$

$=> 1.0 \times 2^{+2}$ → Corrected by shifting significand to the left

(3) Exponent Overflow:
An exponent overflow occurs, when a resulting position exponent exceeds the maximum possible exponent value.

(4) Exponent Underflow:
An exponent underflow occurs, when a resulting negative exponent is less than the minimum possible value.

***Floating point addition/ subtraction

** Algorithm for floating point addition/ subtraction:

(1) Check for zeros: if either no. is zero, result will be other numbers with appropriate sign.

(1) Align the mantissas: if exponent are equal, perform the arithmetic operation, else shift the mantissa with Smaller exponent to the right until its exponent equal to the larger exponent.

(3**) Perform the addition**/ **subtraction** depending on the operation and sign of the two mantissas.

Let X = $X_A \times 2^{+E_A}$

Y = $X_B \times 2^{+E_B}$

Z = ?

Example 1.

During addition: if Y=0, Z=X = $X_A \times 2^{+E_A}$

if X=0, Z=Y = $X_B \times 2^{+E_B}$

During Subtraction: if Y=0, Z=X = $X_A \times 2^{+E_A}$

if X=0, Z=-Y=- $X_B \times 2^{+E_A}$

Example 2.

Let X = $1.10 \times 2^{+3}$ (Larger exponent)

Y = $1.01 \times 2^{+1}$ (Smaller exponent)

Here, 1.10 and 1.01 are mantissa.

Shift the mantissa of Y to the right, until its exponent is equal to the exponent of X

Y = $0.101 \times 2^{+2}$(one time shift)

Y = $0.0101 \times 2^{+3}$(two-time shift)

Now

X = $1.10 \times 2^{+3}$

Y = $0.0101 \times 2^{+3}$

- During addition, if two significand are equal with opposite sign, the result will be zero.
- There is also a possibility of significand(mantissa) overflow by 1 digit, then the mantissa of the result is shifted right, and exponent is incremented.
- Significand overflow occurs, when the addition of two significand (mantissa) of same sign may result in a carry at MSB of mantissa.
- As we increment exponent, there is also a possibility of exponent overflow. Then result will show “error”.
- Normalize the result: If result is not normalized, then we shift mantissa left and decrement the exponent until the value “Left of binary point is 1”. As we are decreasing the exponent, so exponent can also be underflow, then again error is reported.

Example:

$X = 1.01 \times 2^{+3}$

$Y = -1.01 \times 2^{+3}$

Then $X+Y = 1.01 \times 2^{+3}$

$-1.01 \times 2^{+3}$

$Z = X+Y = 0 \times 2^{+3} = 0$

If $X = 1.01 \times 2^{+3}$

$Y = 1.10 \times 2^{+3}$

$Z = X+Y = 10.01 \times 2^{+3}$

(carry at MSB)

Correction = $1.011 \times 2^{+4}$

If $X = 1.01 \times 2^{+3}$

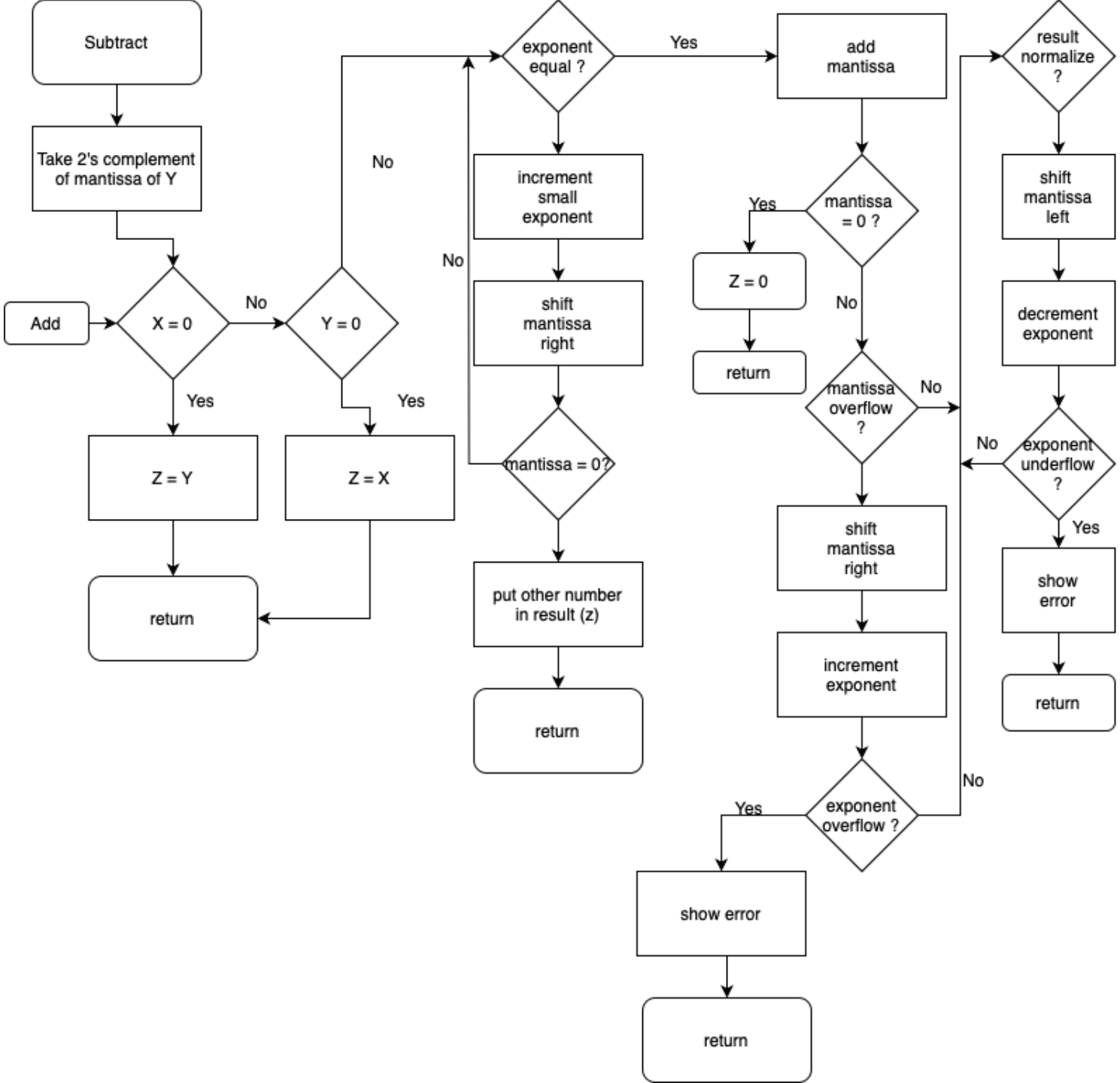
$Y = -1.00 \times 2^{+3}$

$Z = X+Y = 0.01 \times 2^{+3}$ (not normalized)

$= 1.0 \times 2^{+1}$ (normalized)

Flowchart

For Floating point Addition/ Subtraction

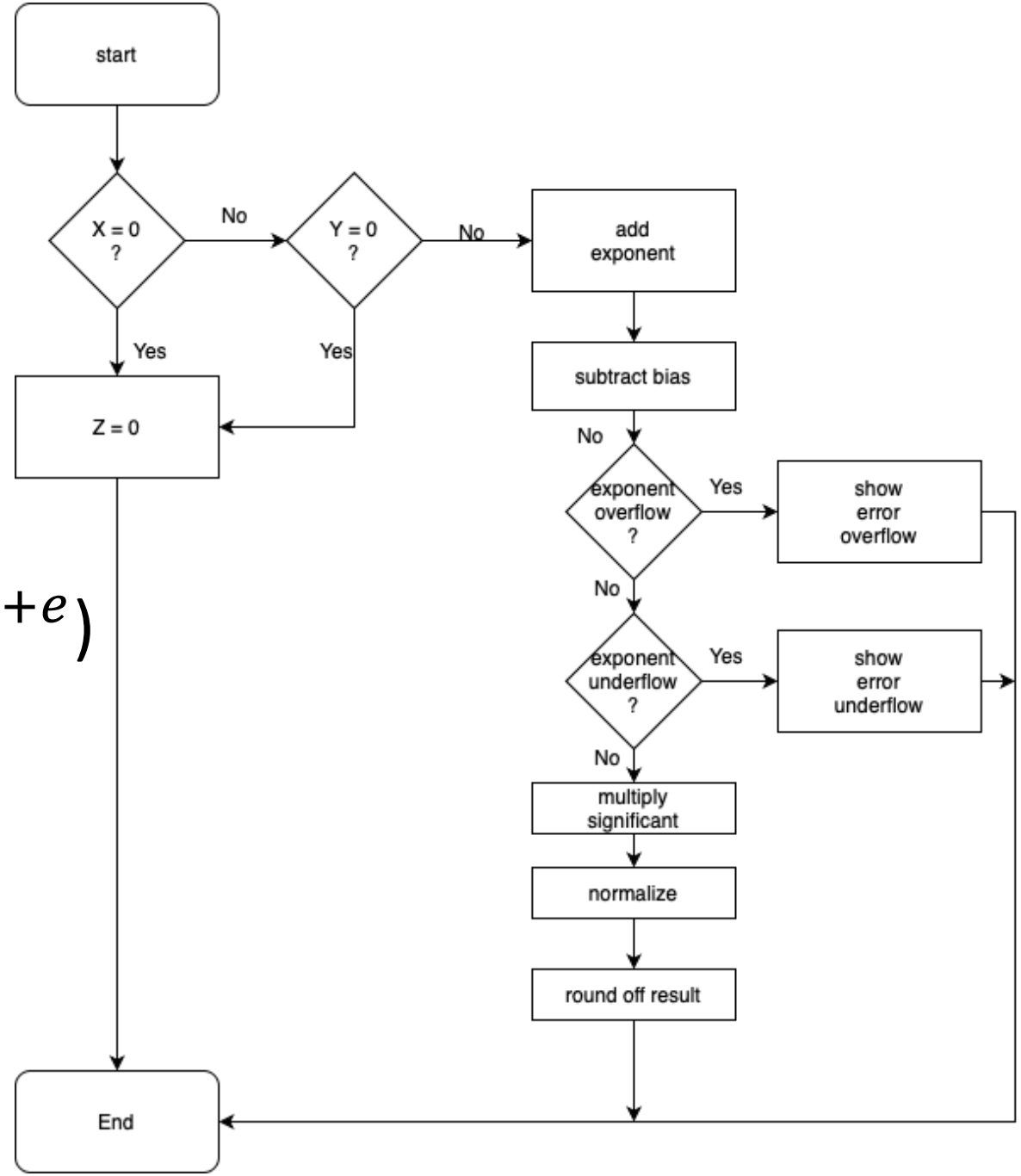


Flowchart: Floating Point Multiplication

Let $X = X_A \times 2^{E_A+e}$

$Y = X_B \times 2^{E_B+e}$

$$X \times Y = (X_A \times 2^{E_A+e}) \times (X_B \times 2^{E_B+e})$$
$$= (X_A \times X_B) \times 2^{E_A+E_B+2e}$$
$$= (X_A \times X_B) \times 2^{E_A+E_B+2e-e}$$
$$Z = (X_A \times X_B) \times 2^{E_A+E_B+e}$$



Flowchart Division of two Floating Point

Let $X = X_A \times 2^{E_A+e}$

$Y = X_B \times 2^{E_B+e}$

$$\frac{X}{Y} = \left(\frac{X_A}{X_B}\right) \times 2^{(E_A+e-E_B-e)}$$

$$\frac{X}{Y} = \left(\frac{X_A}{X_B}\right) \times 2^{(E_A-E_B)}$$

$$\frac{X}{Y} = \left(\frac{X_A}{X_B}\right) \times 2^{(E_A-E_B)+e}$$