Floating Point

A real number consists of infinity of numbers, as we can easily find the result of any arithmetic operation between any two decimal numbers. However, computers of varying sizes and accuracy can only express a specific number of these numbers, meaning they do not use decimal arithmetic, but rather use binary operations and their multiples. In this case, the real number is represented by an approximate number, and although the deleted part is very small, it results in a large error in numerical analysis operations, and thus the numbers are transformed into the case of a loose number, so this type of calculation is called a loose number arithmetic, as the numbers in it are in the following formula:

We assume that x is a real number that is not equal to zero, so x is a number with base β and is written in the form

$$f(x) = (-1)^s \times (d_1\, d_2\, d_3\, \dots d_n)_\beta \times \beta^k \qquad \dots\dots\dots(1-1)$$

Where $(d_1\, d_2\, d_3\, \dots d_n)_\beta = \frac{d_1}{\beta} + \frac{d_2}{\beta} + \cdots + \frac{d_n}{\beta}$ $\dots\dots\dots(1-2)$

Where s=1 or 0

d_1 *is called radix point*

k = integer

If β=2 is called Binary floating point

If β=10 is called decimal floating point

***Example3*:** Like a real number x=5.127 in the form of a floating point?

Solution

$$5.127 = (-1)^0 \times 0.5127 \times 10^1$$

In this case it is:

S=0, β =10 , k=1, $d_1 = 5$, $d_2 = 1, d_3 = 2, d_4 = 7$

Example 4 : Like a real number x=−0.0015 in the form of a floating point?

Solution

$$-0.0015 = (-1)^1 \times 0.0015 \times 10^0$$

In this case it is:

S=1, β =10 , k=1, $d_1 = 0$, $d_2 = 0, d_3 = 1, d_4 = 5$

Definition of Normalization

When floating point is said normalized if

$d_1 = d_2 = \cdots = d_n = 0 \; or \; d_1 \neq 0$

***Example5*:**

Such as the following real number 0.00069 in the form of **normalized** or **not-normalized?**

Solution:

$0.00069 = (-1)^0 \times 0.00069 \times 10^0$, $d_1 = 0$ **not normalized**

$0.00069 = (-1)^0 \times 0.0069 \times 10^{-1}$, $d_1 = 0$ **not normalized**

$0.00069 = (-1)^0 \times 0.069 \times 10^{-2}$, $d_1 = 0$ **not normalized**

$0.00069 = (-1)^0 \times 0.69 \times 10^{-3}$, $d_1 \neq 0$ **normalized**

Any real number such as x except zero can be converted to floating point using one of two methods: Rounding and chopping. For example, the following decimal numbers are converted to floating point decimals with a length of three units as follows:

X=293956, y=0.000318, z=483.62

By Rounding

$\hat{x} = 0.294 \times 10^{6}, \quad \hat{y} = 0.318 \times 10^{-3}, \quad \hat{z} = 0.484 \times 10^{3}$

By chopping

$\hat{x} = 0.293 \times 10^{6}, \quad \hat{y} = 0.318 \times 10^{-3}, \quad \hat{z} = 0.483 \times 10^{3}$

Performing arithmetic operations on numbers in floating point formulas

a) Addition process: We unify the exponents to the largest exponent and perform the addition operation on the decimal fractions.

Example

Let $\hat{x} = 0.0291 \times 10^{2}, \quad \hat{y} = 0.601 \times 10^{2}$ represents the approximate number of the two numbers x and y. find the sum of the two numbers: cutting and rounding.

$\hat{x} + \hat{y} = 0.0291 \times 10^{2} + 0.601 \times 10^{2} = 0.6301 \times 10^{2}$

$cutting \rightarrow 0.6301 \times 10^{2}$

$rounding \rightarrow 0.6301 \times 10^{2}$

b) Subtraction process: We unify the exponents to the largest exponent and perform the subtraction operation on the decimal fractions.

Example

Let $\hat{x} = 0.0291 \times 10^{2}, \quad \hat{y} = 0.601 \times 10^{2}$ represents the approximate number of the two numbers x and y. find the subtraction of the two numbers: cutting and rounding.

$\hat{x} - \hat{y} = 0.0291 \times 10^{2} - 0.601 \times 10^{2} = -0.5719 \times 10^{2}$

$cutting \rightarrow -0.571 \times 10^{2}$

$rounding \rightarrow -0.572 \times 10^{2}$

c) Multiplication process: We collect the exponents, multiply the fractions together, and adjust the result to the floating point case.

Example

Find the product of the following two numbers after rounding the result to three places.

$x \times y = 0.732 \times 10^{3} \times 0.225 \times 10^{-2} = 0.1647 \times 10^{1}$

$cutting \rightarrow 0.164 \times 10^{1}$

$rounding \rightarrow 0.165 \times 10^{1}$

d) Division is the opposite of multiplication, where we subtract the exponents x

Example:

Find the quotient of the following two numbers after rounding the result to three places.

$x \div y = 0.628 \times 10^{3} \div 0.244 \times 10^{-2} = 2.5737 \times 10^{5} = 0.25737 \times 10^{6}$

$cutting \rightarrow 0.257 \times 10^{6}$

$rounding \rightarrow 0.257 \times 10^{6}$

Exercise:

1) Write the following numbers in Floating Point form.

25149, -0.0125, 78.439, 0.733, 0.0039

2) $If\ x = 22.159,\ y = 0.03,\ z = 111,\quad find\ k = 2x + \frac{y+z}{x}$ by using Floating Point.