aEm
$Em = 10^m$

Unit - 1

Normalized floating point representation of real numbers:

In a computer, it is convention to preserve the max^m number of significant digits in a real number and also increase the range of values of real numbers stored. This representation is called the normalized floating point mode of representation and storing real numbers. In this mode a real number is expressed as a combination of a mantissa and an exponent. The mantissa is made less than 1 and greater than or equal to .1 and the exponent is the power of 10 which multiplies the mantissa. For example the number 44.85×10^6 is represented in this notation as .4485 E8 (E8 is used to represent 10^8). The mantissa is .4485 and the exponent is 8.

Note: The leading digit in the mantissa is always made non-zero by appropriately shifting it and adjusting the value of the exponent. For example,

$$.004854 = .4854 \times 10^{-2}$$
$$= .4854\,E{-2}$$

Arithmetic operations with Normalized Floating Point Numbers:

1) Addition:

If two numbers represented in normalized floating point notation are to be added the exponents of the two numbers must be made equal and the mantissa shifted appropriately.

Example 1.

Add $\cdot 4546E5$ and $\cdot 5433E5$

Solⁿ $\cdot 4546E5 + \cdot 5433E5$

$= \cdot 4546 \times 10^5 + \cdot 5433 \times 10^5$

$= (\cdot 4546 + \cdot 5433) \times 10^5$

$= \cdot 9979 \times 10^5$

$= \cdot 9979E5$ //

Example 2. Add $\cdot 4546E5$ to $\cdot 5433E7$

Solⁿ: Here the two exponents are not equal. In such case, the smaller exponent should be made equal to the larger exponent by adjusting mantissa.

Thus,

$\cdot 4546E5 + \cdot 5433E7$

$= \cdot 4546 \times 10^5 + \cdot 5433 \times 10^7$

$= \cdot 0045 \times 10^7 + \cdot 5433 \times 10^7$

$= (\cdot 0045 + \cdot 5433) \times 10^7$

$= (\cdot 5478) \times 10^7$

$= \cdot 5478E7$

[The mantissa should be consist of 4 digits (significant digits)]

Example 3: Add $\cdot 4546E3$ to $\cdot 5433E7$

Solⁿ:- $\cdot 4546E3 + \cdot 5433E7$

$= \cdot 0000 \times 10^7 + \cdot 5433 \times 10^7$

$= \cdot 5433 \times 10^7$

$= \cdot 5433E7$ //

Example 4: Add $\cdot 6434E3$ to $\cdot 4845E3$

Solⁿ:- $\cdot 6434E3 + \cdot 4845E3$

$= (\cdot 6434 + \cdot 4845)E3$

$= 1\cdot 1279E3$

Here mantissa is $m = 1\cdot 1279$. Since $m > 1$ in the normalized floating point system we must make it $m < 1$ and mantissa should contains 4 digits. Hence given expression $= \cdot 1127E4$ //

Note: If the result of an arithmetic operation gives a number smaller than $\cdot 1000E-99$ then it is called an underflow condition. Similarly any result greater than $\cdot 9999E99$ leads to an overflow condition.

Ex 5: Add $\cdot 6434E99$ to $\cdot 4845E99$

Solⁿ:- $\cdot 6434E99 + \cdot 4845E99$

$= \cdot 6434 \times 10^{99} + \cdot 4845 \times 10^{99}$

$= (\cdot 6434 + \cdot 4845) \times 10^{99}$

$= 1\cdot 1279 \times 10^{99}$

$= \cdot 11279 \times 10^{100}$

$= \cdot 1127E100$

Here exponent $n > 99$, so it is an overflow condition and hence the arithmetic unit will intimate an error condition.

2. Subtraction:

The operation of subtraction is nothing but adding a negative number.

Ex 1) Subtract $\cdot 9432E-4$ from $\cdot 5452\ E-3$.

Sol^n:- Here exponents are -4 and -3. Since $-3 > -4$, so exponent -4 should be made -3 by adjusting mantissa.

$$\cdot 5452\ E-3 - \cdot 0943\ E-3$$

$$= (\cdot 5452 - \cdot 0943)\ E-3$$

$$= \cdot 4509\ E-3$$

//

2. Ex) Subtract $\cdot 5424\ E3$ from $\cdot 5452\ E3$

Sol^n:-

$$\cdot 5452\ E3 - \cdot 5424E3$$

$$= (\cdot 5452 - \cdot 5424)\ E3$$

$$= \cdot 0028\ E3$$

$$= \cdot 2800\ E1$$

//

$$\begin{array}{r} \cdot 5452 \\ \cdot 5424 \\ \hline 0028 \end{array}$$

Ex.3) Subtract $\cdot 5424\ E-99$ from $\cdot 5452\ E-99$

Sol^n:-

$$\cdot 5452\ E-99 - \cdot 5424E-99$$

$$= (\cdot 5452 - \cdot 5424)\ E-99$$

$$= \cdot 0028\ E-99$$

$$= \cdot 2800\ E-101$$

Since -101 is less than -99, so it is an underflow condition and hence the arithmetic unit will signal an error condition.

3. Multiplication :-

Two numbers are multiplied in the normalized floating point mode by multiplying the mantissa and adding the exponents. After the multiplication of the mantissas the result mantissa is normalized as in addition / subtraction operation and the exponent appropriately adjusted.

Ex) Multiply $+.5543E12$ by $+.4111E-15$

Sol :- $(.5543E12) \times (.4111E-15)$

$= (.5543 \times 10^{12}) \times (.4111 \times 10^{-15})$

$= (.5543 \times .4111) \times 10^{12} \times 10^{-15}$

$= .22787273 \times 10^{-3}$

$= .2278E-3$

Ex) $.1111E10 \times .1234E15$

$= .01370974 \times 10^{25}$

$= .1370E24$

Ex) $.1111E51 \times .4444E50$

$= .04937284 \times 10^{101}$

$= .4937E100$, overflows.

Ex) $.1234E-49 \times .1111E-54$

$= .1370E-104$, underflows.

Ex) Briefly explain how two numbers are multiplied in the normalized floating point mode. What do you mean by 'over flow' in this context.

4. Division:-

In division a number by another the mantissa of the numerator is divided by that of the denominator. The denominator exponent is subtracted from the numerator exponent. The quotient mantissa is normalized to make the most significant digit non-zero and the exponent appropriately adjusted.

Ex)
$$\cdot 1000E5 \div \cdot 9999E3$$
$$= (\cdot 1000 \div \cdot 9999)(10^5 \div 10^3)$$
$$= \cdot 1000 \times 10^2$$
$$= \cdot 1000E2$$

Ex)
$$\cdot 9998E1 \div \cdot 1000E-99$$
$$= 9.998 \times 10^{100}$$
$$= \cdot 9998 \times 10^{101}$$
$$= \cdot 9998E101$$
, overflows.

Ex)
$$\cdot 9998E-5 \div \cdot 1000E98$$
$$= 9.998 \times 10^{-103}$$
$$= \cdot 9998 \times 10^{-102}$$
, underflows.
$$= \cdot 9998E-102$$

* Rounding off Numbers;

To round off a number to n-significant figures, all digits to the right of the n-th place are to be discarded. If the discarded number is less than half a unit in the n-th place, then the n-th digit is to be left unchanged; on the other hand, if it is greater than half, then 1 is to be added to the n-th digit. If the discarded number is exactly half a unit in the n-th place, then the n-th digit is to be left unaltered if it is even and is to be increased by 1 if it is odd, i.e., the number is to be rounded off in such a way that the n-th digit remains always even.

Some examples of rounding off numbers correct to five significant figures:

$5.2908623 \approx 5.2909$

$1.290813 \approx 1.2908$

$31.79992 \approx 31.800$

$4.39065 \approx 4.3906$

~~3.48755 ≈~~

$0.00324555 \approx 0.0032456$

$4632925 \approx 46329 \times 10^2$

Significant Figures: A significant figure is any one of the digits 1, 2, 3, ... , 9 and 0 is a significant figure except when it is used to fix the decimal point or to fill the places of unknown or discarded digits.

Thus, in the number 0.00134, the significant figures are 1, 3, 4. The zeros are used here merely to fix the decimal point and therefore not significant. But in the number 0.1204 the significant figures are 1, 2, 0, 4.

Specifically, the rules for identifying significant figures when writing or interpreting numbers are as follows:

(1) All non-zero digits are considered significant. For example, 91 has two significant figures (9 and 1) while 123.45 has five significant figures (1, 2, 3, 4, 5).

(2) ~~Zeros~~ Zeros appearing anywhere betn two non-zero digits are significant. Example: 101·12 03 has seven significant figures: 1, 0, 1, 1, 2, 0 and 3.

(3) Leading zeros are not significant. For example, 0·00052 has two significant figures: 5 and 2.

(4) Trailing zeros in a number containing a decimal point are significant. For example, 12·23 00 has six significant figures: 1, 2, 2, 3, 0 and 0. The number 0·000122300 still has only six significant figures (the zeros before the 1 are not significant).

<u>Note:</u> It should be ~~nodated~~ noted that in decimal representation, a significant digit of an approximate number is any non-zero digit or any zero lying within significant digits or used as a place holder to indicate a retained place. In scientific notation ($M \times 10^n$), all the digits explicitly in M are significant.

100. = 3 significant

100 = 1 significant

$48{\cdot}32 = 0{\cdot}4832 \times 10^2$

$7{\cdot}20 = 0{\cdot}720 \times 10^1$

$0{\cdot}00084 = 0{\cdot}84 \times 10^{-3}$

Approximate numbers:

We know that the numbers such as $3, \frac{1}{5}, \frac{1}{8}, 100$ etc. are exact numbers because these numbers involve no approximation or uncertaintity. On the otherhand, the numbers like π, e, $\sqrt{3}$ etc. are exact, but cannot be expressed exactly by a finite number of digits. They can be written in digital form as 3.1416, 2.7183, 1.7320, etc. which are only approximations to the ~~tue~~ ~~the~~ true values and are called approximate numbers.

Thus an approximate number is defined as a number approximated to the exact number and there is a slight difference betn the ~~exactt~~ exact and approximate numbers.

Errors in Arithmetic operations:

Absolute error:

The absolute error of computed result is equal to the difference betn the true value and the computed value.

Thus if x_T = true value

x_A = approximate value

$$\text{Absolute error} = |x_T - x_A|$$

For example, if we approximate $\frac{1}{3}$ by 0.3333, the absolute error E_A is given by

$$E_A = \left|\frac{1}{3} - 0.3333\right| \cong 0.00003$$

Relative error:

The relative error E_R is defined as the absolute error divided by the true value. Thus

$$E_R = \frac{E_A}{x_T} = \frac{|x_T - x_A|}{|x_T|} \quad , \text{Provided } x_T \neq 0$$

Hence the relative error in approximating $\frac{1}{3}$ by 0.3333 is given by $\frac{0.00003}{\frac{1}{3}} = 0.00009$.

Relative percentage error:

The accuracy of a computed result is better represented by the relative percentage error which is defined by

$$E_P = E_R \times 100 = \frac{|x_T - x_A|}{|x_T|} \times 100 \quad , \text{Provided } x_T \neq 0$$

Thus the relative percentage error in approximating $\frac{1}{3}$ by 0.3333 is given by

$$E_P = 0.00009 \times 100 = 0.009\,\%$$

Truncation error:

In numerical analysis and scientific computing, truncation error is the error made by truncating an infinite sum and approximating it by a finite sum. For example, if we approximate the sine function by the first two non-zero term of its Taylor series, as in $\sin(x) \approx x - \frac{1}{6}x^3$ for small x, the resulting error is truncation error.

Round-off error:

A round-off error is introduced or generated when a real number is converted to machine number with finite size.

Consider the following number with four-digit word length (which can be expressed in exponent form called floating point representation given within parenthesis)

(a) 223.4 (0.2234×10^3)

(b) 0.05124 (0.5124×10^{-1})

(c) 11.12 (0.1112×10^2)

The result of (a) $\times$ (b) is 11.447016. However, in four digit representation this product is rounded to four significant figures and is given by as 0.1145×10^2. The round-off error in this case is about 3×10^{-3}.

Again the sum (a) + (c) will be performed by shifting the mantissa of the number with smaller exponent to the right till its exponent agrees with that of the large number.

Thus $0.2234 \times 10^3 + .01112 \times 10^3 = 0.23452 \times 10^3$. The sum would then be rounded to a four-digit number giving the result as 0.2345×10^3 with an error $= 2 \times 10^{-2}$.

Ex) Write down approximate representation of $\frac{2}{3}$ correct upto four significant digits; find the absolute, relative and percentage error.

Sol^n Here, $\frac{2}{3} = 0.66666\text{-----}$

$\therefore \frac{2}{3} = 0.6667$, correct upto four significant digits.

$\therefore$ absolute error is

$$= \left| \frac{2}{3} - 0.6667 \right|$$

$$= 3.3 \times 10^{-5}$$

$$\text{Relative error} = \frac{\left| \frac{2}{3} - 0.6667 \right|}{\left| \frac{2}{3} \right|}$$

$$= 5 \times 10^{-5}$$

(Round-off)

and percentage error $= 5 \times 10^{-5} \times 10^{2}$

$$= 5 \times 10^{-3}$$

(Round-off)

Ex) If $\pi = 3.14$ is used to in place of 3.14156, find the absolute and relative errors.

Sol^n Here the true value $= 3.14156$ and approximate value $= 3.14$

$\therefore$ absolute error $= |3.14156 - 3.14|$

$$= .00156$$

and relative error $= \frac{|3.14156 - 3.14|}{|3.14156|}$

$$= .000496$$

(Round-off)