Floating point representation of numbers

There are two types of arithmetic operations available in a computer:

(i) Integer Arithmetic

(ii) Real or Floating point arithmetic

The integer arithmetic deals with integer operands and used mainly in counting. The real arithmetic uses numbers with fractional parts as operands and used in most computations.

Normalized floating point representation:

A real number is expressed as a combination of **mantissa** and **exponent**. The mantissa is made less than 1 and greater than .1 *i.e.* $(.1 \leq \text{mantissa} \leq 1)$, while the exponent is power of 10 which multiplies the mantissa.

Ex: The number 12.56×10^5 is represented in this notation $.1256E7$, where $E7$ is used to represent 10^7. Thus, the mantissa is $.1256$ and the exponent is 7.

$1.556 \times 10^5 = 0.1556E6$

$123 \times 10^6 = 0.123E9$

$.00453 = .453E-2$

- Moreover, shifting of mantissa to the left till its most significant digit is non-zero is called Normalization.

$.00453 = .453E-2$

Arithmetic operations with Normalized floating point numbers:

Addition and subtraction:

If two numbers represented in normalized floating point notation are to be added, the exponent of two numbers must be made equal and mantissa shifted appropriately. The operation of subtraction is nothing but adding a negative number. Thus, the principles are same.

Ex: Add the floating point numbers $.1254E6$ and $.4631E6$.

Solution: Here, the exponents are equal so mantissa are added.

$\therefore$ Sum= $.1254E6$

$+ .4631E6$

$.5885E6$

Ex: Add the floating point numbers $.4534E5$ and $.4231E6$.

Solution: Here, the exponents are not equal. The operand with larger exponent kept as it is. So, $.4534E5 = .0453E6$

$\therefore$ Sum= $.4231E6$

$+ .0453E6$

$.4684E6$

Ex: Add the floating point numbers $.6434E5$ and $.5231E5$.

Solution: Here, the exponents are equal.

$\therefore$ Sum= $.6434E5$

$+ .5231E5$

$1.1665E5$

Here, $1.1665E5 = .1166E6$. Thus, the resulting sum $= .1166E6$

Multiplication:

Two numbers are multiplied in the normalized floating point mode by multiplying the mantissa and adding the exponents. After multiplication, the resulting mantissa will be normalized.

Ex: Multiply the floating point numbers $.3532E5$ and $.4231E6$.

Solution: Here, multiplication= $.3532E5 \times .4231E6 = .1494E11$

Ex: Multiply the floating point numbers $.1123E5$ and $.2312E4$.

Solution: Here, multiplication= $.1123E5 \times .2312E4 = .02596E9 = .2596E8$

Thus, the resulting multiplication= $= .2596E8$

Ex: Multiply the floating point numbers $.3532E5$ and $.4231E-2$.

Solution: Here, multiplication= $.3532E5 \times .4231E-2 = .1494E3$

Division:

In division, the mantissa of numerator is divided by mantissa of denominator and the denominator exponent is subtracted from numerator exponent. Then, the quotient mantissa is normalized.

Ex: Perform the operation to the floating point numbers $.1000E5 \div .9999E3$.

Solution: $.1000E5 \div .9999E3 = .1000E2$

Ex: Perform the operation to the floating point numbers $.9998E1 \div .1000E-99$.

Solution: $.9998E1 \div .1000E-99 = 9.998E100 = .9998E101$

Thus, the resulting floating point number is $.9998E101$