U3A Maths

26-Mar-2024 Term 1 / Week 9

Floating Point Number

Homework Solution

Given $(0.35)_{10} \cong (0.01011)_2$ by ignoring recurring numbers

- Calculate the rounding-off error

$2^{-1}\ 2^{-2}\ 2^{-3}\ 2^{-4}\ 2^{-5}$

0.01011

$= 0\times 2^{-1} + 1\times 2^{-2} + 0\times 2^{-3} + 1\times 2^{-4} + 1\times 2^{-5}$

$= 1\times\frac{1}{4} + 1\times\frac{1}{16} + 1\times\frac{1}{32}$

$= 0.25 + 0.0625 + 0.03125$

$= 0.34375.$

$\therefore$ Error $= 0.35 - 0.34375$

$= 0.00625$

$\%\text{Error} = \frac{0.00625}{0.35}\times 100$

$= 1.79\ \%$

Trying next approximation,

$(0.35)_{10} = (0.0101100011)_2$

we get

$(0.0101100011)_2 \cong (0.34961)$

$\%\text{ Error} = \frac{0.35 - 0.34961}{0.35}\times 100$

$= 0.11\%$

-3-

Floating Point Number

- Reviewing the Example from last week, we have

$(25.625)_{10} = (11001.101)_2$

- This format is called "Fixed Point" – the fractional value follows the decimal or the binary point.

- For computer storage, it is convenient to 'set' the 'point' at one end. Normally, the 'point' is shifted to the left most end.

-4-

$(11001.101)_2 = 0.11001101\times 2^{+5}$

$= [0.11001101\times (10)^{+101}]$

- We have used 2^{α} for clarity. Note that $(2)_{10} = (10)_2$

- In the computer, the above value is stored as below:

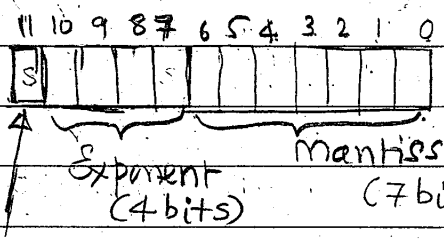
11001101 is called the Mantissa

+101 is called the Exponent

- Note that the "Exponent" is alway an Integer or a whole number.

- Negative Integers are stored as 2's complement!

- Assuming we have a 12-bit machine, the above number is stored as below:



'S' indicates the Sign bit of the mantissa!

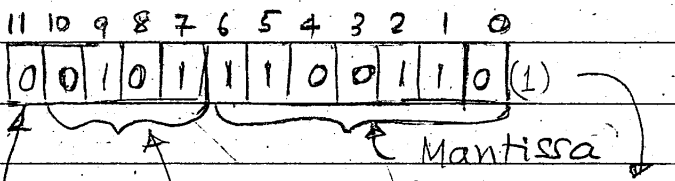
We have: Mantissa ⇒ 1100110 1

Sign of Mantissa ⇒ 0 (+ive)

- Exponent is stored as a 2's complement for negative values.

- We have a positive exponent (+5), hence no need for 2's complement.

- Hence the number is stored as below:



- IEEE 754 standard for floating point numbers uses a trick! (Next week!)

Example

Add 25.625 + 4.125

In floating point form

$25.75 = 0.25625 \times 10^2$

$4.125 = 0.4125 \times 10^1$

To equate exponents, let us adjust this exponent it should be 10^2!

$0.4125 \times 10^1 = 0.04125 \times 10^2$

∴ the sum is

$$\begin{array}{r} 0.25625 \times 10^2 \\ + \quad 0.04125 \times 10^2 \\ \hline 0.29750 \times 10^2 \end{array}$$

- The answer is 0.29750×10^2
- In "fixed point" form the answer is 29.75!
- In modern computers, the floating point arithmetic is done by dedicated hardware!

Homework

- Add the numbers in the above Example by changing both exponents to 1. {which method is preferable?}