CSD1100

Floating-Point Numbers

Vadim Surov

Floating Point Numbers

- In the decimal system, a decimal point (radix point) separates the whole numbers from the fractional part
- Examples:

  37.25 (whole=37, fraction=25)

  123.567

  10.12345678

Floating Point Numbers

- Where 37.25 has:

10^1	10^0	10^{-1}	10^{-2}
Tens	Units	Tenths	Hundredths
3	7	2	5

- Or

$$37.25 =$$

$$3 \text{ x } 10^1 + 7 \text{ x } 10^0 + 2 \text{ x } 10^{-1} + 5 \text{ x } 10^{-2} =$$

$$3 \text{ x } 10 + 7 \text{ x } 1 + 2 \text{ x } 1/10 + 5 \text{ x } 1/100$$

Floating Point Numbers

- In the binary representation of a floating point number the column values will be as follows:

$$\ldots\ 2^{6}\quad 2^{5}\quad 2^{4}\quad 2^{3}\quad 2^{2}\quad 2^{1}\quad 2^{0}\ .\ 2^{-1}\quad 2^{-2}\quad 2^{-3}\quad 2^{-4}\ \ldots$$

or

$$\ldots\ 64\quad 32\quad 16\quad 8\quad 4\quad 2\quad 1\ .\ 1/2\quad 1/4\quad 1/8\quad 1/16\ \ldots$$

or

$$\ldots\ 64\quad 32\quad 16\quad 8\quad 4\quad 2\quad 1\ .\ .5\quad .25\quad .125\quad .0625\ \ldots$$

Finding Binary Equivalent Of Fraction Part

- Converting .25 using **Multiplication method.**

Step 1 : multiply fraction by 2 until fraction becomes 0

.25
x 2
0.5
x 2
1.0

Step 2 Collect the whole parts and place them after the radix point

64	32	16	8	4	2	1	.	.5	.25	.125	.0625
							.	0	1		

Finding Binary Equivalent Of Fraction Part

- Converting .25 using **Subtraction method.**

Step 1: write positional powers of two and column values for the fractional part

.	2^{-1}	2^{-2}	2^{-3}	2^{-4}	2^{-5}
.	½	¼	1/8	1/16	1/32
.	.5	.25	.125	.0625	0.03125

Finding Binary Equivalent Of Fraction Part

- Converting .25 using **Subtraction method.**

Step 2: start subtracting the column values from left to right, place

0 if the value cannot be subtracted or

1 if it can be subtracted

Repeat step 2 until the fraction becomes 0.

$$\begin{array}{r} .25 \\ \underline{-\ .25} \\ .0 \end{array}$$

2	1	.	.5	.25	.125	.0625
		.	0	1		

Binary Equivalent Example

Given 37.25, convert 37 and .25 using subtraction method.

64	32	16	8	4	2	1	.	.5	.25	.125	.0625
2^6	2^5	2^4	2^3	2^2	2^1	2^0	.	2^{-1}	2^{-2}	2^{-3}	2^{-4}
	1	0	0	1	0	1	.	0	1		

$$\begin{array}{r} 37 \\ \underline{-\ 32} \\ 5 \\ \underline{-\ 4} \\ 1 \\ \underline{-1} \\ 0 \end{array} \qquad \begin{array}{r} .25 \\ \underline{-\ .25} \\ .0 \end{array}$$

$$37.25_{10} = 100101.01_2$$

So what is the Problem?

Given the following binary representation:

$37.25_{10} = 100101.01_2$

$7.625_{10} = 111.101_2$

$0.3125_{10} = 0.0101_2$

How we can represent the whole and fraction part of the binary rep. in 4 bytes?