Floating Point Number Representation

Floating Point Representation

Floating point representation of a number has two parts:

Mantissa: Signed fixed point number, either an integer or a fractional number

Exponent: Designates the position of the decimal or binary point

Example: The decimal number +6132.789 is represented in floating-point with a fraction and a exponent as follows:

Fraction	Exponent
+0.6132789	+04

This representation is equivalent to the scientific notation $+0.6132789\times 10^{+4}$

Floating Point Representation

In a floating point representation, the number is always represented in the following form:

$$\boldsymbol{m} \times \boldsymbol{r}^{\boldsymbol{e}}$$

where m is mantissa, e is exponent and r is the radix.

Mantissa(m) and exponent(e) are physically represented in the register.

Radix(r) and Radix-point position of the mantissa are always assumed.

Floating Point Representation

Example:

The binary number +1001.11 can be represented in a floating point representation with 8-bit fraction and 6-bit exponent as follows:

Fraction	Exponent
01001110	000100

A 0 in the leftmost position of the fraction denote positive and the exponent has a equivalent binary number +4.

The floating point number is equivalent to $+(.1001110)_2 \times 2^{+4}$

Normalization of Floating Point Numbers

- A floating point number is said to be Normalized if the most significant position of the mantissa is a Non-zero.

Example 1: The number 350 is normalized or not. (Yes)

Example 2: Normalized the number $(.0035)_{10} \longrightarrow .3500 \times 10^{-2}$

Example 3: 8-bit binary number $(.00010101)_2$ is normalized or not (No)

Example 4: Normalized the number $(.00010101)_2 \longrightarrow .10101000 \times 2^{-3}$

Questions

- Q1. Represent the number (+46.5), as a floating-point binary number with 24 bits. The normalized fraction mantissa has 16 bits and the exponent has 8 bits.

- Q2. Normalized value of

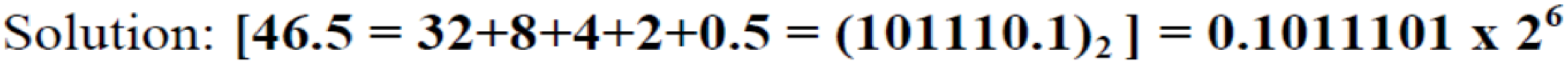
a) .000101010

b) 101010.101

c) 01010.1

Questions

- Q1. Represent the number (+46.5), as a floating-point binary number with 24 bits. The normalized fraction mantissa has 16 bits and the exponent has 8 bits.



Solution: $[46.5 = 32+8+4+2+0.5 = (101110.1)_2] = 0.1011101 \times 2^6$