Real Number to Floating-point Number Conversion Examples

Monday, January 30, 2023 5:47 PM

\# Given a computing system with the following specifications: $m=3$, $e_{min}=-1$, $e_{max}=3$, how does the system represent $x_1 = 5.875$ and $x_2 = 6.35$?

Note: $e_{min} \le e \le e_{max}$; $e \in \mathbb{Z}$

- **Normalized Form**

System

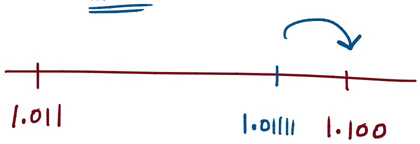
$$(\underbrace{1.d_1d_2d_3}_{m=3})_2 \times 2^e$$

$$x_1 = 5.875 = (101.111)_2 \times 2^0$$

Bring to normalized form

$$(1.01111)_2 \times 2^2 \checkmark$$

Notice that we can not represent all the 5 bits after the radix since $m=3$.



$$(1.01111)_2 \times 2^2 \longrightarrow (1.\overbrace{100}^{m=3})_2 \times 2^2 = fl(x_1)$$

$\in F_N$

F_N: floating-point numbers representable by the system. in the Normalized Form.

- **Denormalized Form**

System

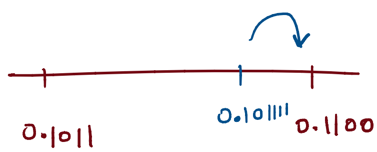
$$(0.1\underbrace{d_1d_2d_3}_{m=3})_2 \times 2^e$$

$$x_1 = 5.875 = (101.111)_2 \times 2^0$$

Bring to denormalized form

$$x_1 = (0.101111)_2 \times 2^3$$

Again, we can't represent all the bits after the radix.



$$(0.101111)_2 \times 2^3 \longrightarrow (0.1100)_2 \times 2^3 = fl(x_1)$$

$\in F_D$

F_D ~~F_N~~: floating-point numbers representable by the system. in the deNormalized Form.

- **Convention #1**

- Figure it out.
- Decimal to Binary Conversion by hand if required.

[Tutorial by an Indian gentleman.]

- Do the same for x_2.

- After representing x_1 and x_2 in the system, compute x_1^2, and find its representation $fl(x_1^2)$ in Denormalized form.

$x_1 = 5.875$ and we found that $fl(x_1) = (0.1100)_2 \times 2^3 = 0.75 \times 2^3 = 6 \checkmark$

Now, $fl(x^2) = fl(x_1) \cdot fl(x_1) \rightarrow 6 \times 6 = \underline{\underline{36}}$

$36 = (100100.0)_2 \times 2^0 = (0.100100)_2 \times 2^6 \notin F_D$

Here, we see that the exponent $e = 6$.

$e > e_{max}$ $\therefore$ $fl(x_1^2)$ is not representable even though we can perfectly represent the fractional part $(0.\underbrace{1001}_{m=3})_2$.

Note that after any computation such as $fl(x^2)$, $fl(x_1 + x_2)$, $fl(x_1 x_2)$, we have to represent the result according to the system representation.