Worksheet

01/22/2020

Name: ________________________________

Below is a toy model of IEEE double-precision floating-point format.
The sequence of 8 bits

$$\underbrace{c_0}_{\text{sign part}} \quad \underbrace{b_1 \quad b_2 \quad b_3 \quad b_4}_{\text{exponent part}} \quad \underbrace{a_1 \quad a_2 \quad a_3}_{\text{mantissa part}}$$

represents a number $x = \sigma \cdot \bar{x} \cdot 2^e$ where σ, $\bar{x}$, e are determined as follows:

$$\begin{aligned} \sigma &= \begin{cases} 1 & \text{if } c_0 = 0, \\ -1 & \text{if } c_0 = 1, \end{cases} \\ E &= (b_1 b_2 b_3 b_4)_2 \end{aligned}$$

- If $1 \leq E \leq 14$ then

$$\begin{aligned} e &= E - 7, \\ \bar{x} &= (1.a_1 a_2 a_3)_2 \end{aligned}$$

- If $E = 0$ then $e = -6$ and $\bar{x} = (0.a_1 a_2 a_3)_2$.
- If $E = 15$ then $x = \pm\infty$ (depending on the sign σ).

1) Perform the addition of floating-point numbers by following the procedure:

   (a) Rewrite the smaller number such that its exponent matches with the exponent of the larger number.
   (b) Add the significands.
   (c) Normalize the result by moving the floating point.
   (d) Round the result.

   (A) $(1.101)_2 \times 2^2 + (1.111)_2 \times 2^2$

   See Lecture 7

   (B) $(1.011)_2 \times 2^2 - (0.111)_2 \times 2^2$

(C) $-(1.010)_2 \times 2^1 + (1.011)_2 \times 2^{-1}$

2) Perform the below floating-point multiplication by following the procedure:

(a) Add two exponents.
(b) Multiply the significands.
(c) Normalize the result by shifting the floating point.
(d) Round the significand.

$(1.101)_2 \times 2^{-3} \times (1.111)_2 \times 2^2$

See Lecture 7

3) For each given x, find the next number (smallest number greater than x) that can be represented with exactness by the above floating-point format. Find the difference between two numbers (written as power of 2).

(a) $x = (0.000)_2 \times 2^{-6}$

(b) $x = (0.001)_2 \times 2^{-6}$

(c) $x = (1.101)_2 \times 2^{-2}$

(d) $x = (1.010)_2 \times 2^6$