Worksheet

10/9/2019

Name: __________________________________

Below is a toy model of IEEE double-precision floating-point format.
A number is represented by a sequence of 8 bits:

$$\underbrace{c_0}_{\text{sign}} \; \underbrace{c_1 \quad c_2 \quad c_3 \quad c_4}_{E} \; \underbrace{a_1 \quad a_2 \quad a_3}_{\bar{x}}$$

This sequence represents the number $x = \sigma \cdot \bar{x} \cdot 2^e$ where

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

$$\begin{aligned} \sigma &= \begin{cases} 1 & \text{if} \quad c_0 = 0, \\ -1 & \text{if} \quad c_0 = 1, \end{cases} \\ e &= E - 7, \\ \bar{x} &= (1.a_1a_2a_3)_2 \end{aligned}$$

- If $E = 0$ then $e = -6$ and $\bar{x} = (0.a_1a_2a_3)_2$.
- If $E = 15$ then the bit sequence represents $\pm\infty$ (depending on the sign σ).

1) Write 0 and 1 in the floating-point format described above (in form of $\sigma \cdot \bar{x} \cdot 2^e$).

$0 = (0.000)_2 \times 2^{-6}$

$1 = (1.000)_2 \times 2^0$

2) 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) $x = (0.000)_2 \times 2^{-6}$, $y = (0.001)_2 \times 2^{-6}$

$y - x = [(0.001)_2 - (0.000)_2] \times 2^{-6} = 2^{-3} \times 2^{-6} = 2^{-9}$

(b) $x = (0.001)_2 \times 2^{-6}$, $y = (0.010)_2 \times 2^{-6}$

$y - x = [(0.010)_2 - (0.001)_2] \times 2^{-6} = (0.001)_2 \times 2^{-6} = 2^{-9}$

(c) $x = (1.101)_2 \times 2^{-2}$, $y = (1.110)_2 \times 2^{-2}$

$y - x = [(1.110)_2 - (1.101)_2] \times 2^{-2} = (0.001)_2 \times 2^{-2} = 2^{-3} \times 2^{-2} = 2^{-5}$.

(d) $x = (1.010)_2 \times 2^6$, $y = (1.011)_2 \times 2^6$

$y - x = [(1.011)_2 - (1.010)_2] \times 2^6 = (0.001)_2 \times 2^6 = 2^{-3} \times 2^6 = 2^3$.