Worksheet
10/7/2019

Name: ______________________________

Below is a toy model of IEEE double-precision floating-point format to demonstrate how machine arithmetic is done.
A number is represented by a sequence of 8 bits:

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

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).

(a) Convert the number 28.375 from decimal system to binary system (exact form).

$28.375 = 28 + 0.375$

$28 = (11100)_2$

$0.375 = (0.0110)_2$

$28.375 = (11100.011)_2$

28	0
14	0
7	1
3	1
1	1
0	

$0.375 \times 2 = 0.75 \longrightarrow 0$

$0.75 \times 2 = 1.5 \longrightarrow 1$

$0.5 \times 2 = 1 \longrightarrow 1$

$0 \times 2 = 0 \longrightarrow 0$

(b) What is the (approximate) representation of 28.375 in this format?

$$(11100.011)_2 = (1.1100011)_2 \times 2^4 \approx (1.110)_2 \times 2^4$$

rounding

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

1. Add two exponents.
2. Multiply the significands.
3. Normalize the result by shifting the floating point.
4. Round the significand.

$$\underbrace{(1.101)_2 \times 2^{-3}}_{x} \times \underbrace{(1.111)_2 \times 2^{2}}_{y}$$

Note that $M = 7$ and $m = -6$

1) Add exponents:

$$2^{-3} \times 2^{2} = 2^{-1}$$

-1 is between m and M.

2) Multiply the significands:

$$
\begin{array}{rr}
 & 1.101 \\
\times & 1.111 \\
\hline
 & 1101 \\
+ & 1101 \\
 & 1101 \\
 & 1101 \\
\hline
 & 11.000011
\end{array}
$$

$1+1 = (10)_2$. Write 0, carry 1.

$1+1+1+1 = 4 = (100)_2$. Write 0, carry 10

$1+1+10 = (100)_2$. Write 0, carry 10

$1+1+10 = (100)_2$. Write 0, carry 10

$1+10 = 11$. Write 11

3) Shift the floating point:

$$xy = (11.000011)_2 \times 2^{-1}$$

$$= (1.1000011)_2 \times 2^{0}$$

4) Round off:

$$xy \approx (1.100)_2 \times 2^{0}$$