Lecture 6 (10/7/2019)

* Rounding in decimal system:

$x = 1.125964$

One digit after the point: $x \approx 1.1$

Two " " " " : $x \approx 1.13$

Three " " " " : $x \approx 1.126$

Four " " " " : $x \approx 1.1260$

* Rounding in binary system:

$a = (0.101011)_2$

Suppose we want to take only 3 digits after the point. Then

$a \approx (0.101)_2$

because $0.101\,\cancel{011}$

How about 4 digits?

$a \approx (0.1011)_2$

The fourth digit gets added one unit because the digit after it is 1.

$$\begin{array}{r} 0.1010\,\cancel{11} \\ +\ 0.0001 \\ \hline 0.1011 \end{array}$$

How about 5 digits:

$a \approx (0.10110)_2$

because $0.10101\,\cancel{1}$

$$\begin{array}{r} +\ 0.00001 \\ \hline 0.10110 \end{array}$$

* Multiplication of floating-point numbers:

$x = \sigma \cdot \bar{x} \cdot 2^e$, $y = \tau \cdot \bar{y} \cdot 2^f$

$$xy = (\bar{x}\bar{y})\, 2^{e+f}$$

Rule: 1) Add the exponents. If there is a specified range for the exponent, say between m and M, then do the following:

- If $e+f > M$, $xy = \infty$ or $-\infty$ (overflow)
- If $e+f < m$, $xy = 0$ (underflow)

2) Multiply the significands (mantissas)

3) Normalize by shifting the floating point.

4) Round the significand.

Recall that in the floating-point format in the worksheet last time, $M = 7$ and $m = -6$. For the IEEE-754 standard, $M = 1023$ and $m = -1022$.

Ex:

Perform the multiplication of the following numbers in the floating-point format specified in the last worksheet.

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

$$y = (1.101)_2 \times 2^1$$

$$\begin{array}{rl} \bar{x} = & 1.010 \\ \times\ \bar{y} = & \times\ 1.101 \\ \hline & 1010 \\ & 0000 \\ + & 1010 \\ \hline & 1010 \\ \hline \bar{x}\bar{y} = & 11.000010 \end{array}$$

$$xy = (11.00001)_2\, 2^3$$
$$= (1.100001)_2\, 2^4 \quad \text{(normalized)}$$
$$\approx (1.100)_2\, 2^4 \quad \text{(roundoff)}$$

Next, we consider some consequences of error by floating-point arithmetic:
- loss of significant digits
- overflow, underflow
- random behavior in "micro"-scale.

1) Loss of significant digits:

This is caused by adding or multiplying two numbers that are too different in size. For example,

$$(1.011)_2 \times 2^6 + (1.010)_2 \times 2^{-3} = (1.011)_2 \times 2^6$$
$$\overbrace{(0.00\cdots0101)_2 \times 2^6}$$

The smaller number is completely ignored in this addition, leading to the error $x+y=x$.