* 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 \( 0.101011 \)

How about 4 digits?

\[ a \approx (0.1011)_2 \]

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

\[ 0.101011 \]

\[ + \quad 0.0001 \]

\[ \frac{0.1011}{0.1011} \]

How about 5 digits:

\[ a \approx (0.10110)_2 \]

because \( 0.101011 \)

\[ + \quad 0.00001 \]

\[ \frac{0.10110}{0.10110} \]

* Multiplication of floating-point numbers:

\[ x = 5 \cdot 1.2 \cdot 2^e, \quad y = 1.4 \cdot 2^f \]
\[ x \cdot y = (\varepsilon_x) (\varepsilon_y) \cdot e^{x+y} \]

**Rule:**
1. Add the exponents. If there is a specified range for the exponent, say between \( m \) and \( M \), then do the following:
   - If \( e^{x+y} > M \), \( x \cdot y = \infty \) or \(-\infty\) (overflow)
   - If \( e^{x+y} \leq m \), \( x \cdot y = 0 \) (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:

\[ x = (1.010)_2 \times 2^2 \]
\[ y = (1.101)_2 \times 2^1 \]

\[ \frac{x}{\overline{y}} = \frac{1.010}{1.101} = \frac{1010}{1010} \]

\[ \begin{array}{cccc}
  \hline
  1 & 0 & 1 & 0 \\
  + & 0 & 0 & 0 \times 2^0 \\
  \hline
  1 & 0 & 1 & 0 \\
  \hline
  1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0
\end{array} \]

\[ \bar{x} \overline{y} = 1.10000010 \]
\[ xy = (1.00001)_2 \times 2^3 \]
\[ \quad = (1.100001)_2 \times 2^4 \quad \text{(normalized)} \]
\[ \quad \approx (1.100)_2 \times 2^4 \quad \text{(roundoff)} \]

Next, we consider some consequences of error by floating-point arithmetic:

\[ \{ \begin{align*}
\text{loss of significant digits} \\
\quad \text{overflow, underflow} \\
\quad \text{random behavior in “micro-scale”}
\end{align*} \]

1) Loss of significant digits:

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

\[
(1.011)_2 \times 2^6 + (1.010)_2 \times 2^{-3} = (1.011)_2 \times 2^6
\]

\[
\underline{(0.00...0101)_2 \times 2^6}
\]

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