Problem 2.1

1. If the Marc-32 did not round off numbers correctly but simply dropped excess bits, what would the unit roundoff be?

9. Let $x = (1.11...11100...)_2 \times 2^16$, in which the fractionall part has 26 1's followed by 0's. For the Marc-32, determine x_- x_+, $fl(x)$, $x - x_-$, $x_+ - x$, $x_+ - x_-$, and $|x - fl(x)|/|x|$.

10. Let $x = 2^3 + 2^{-19} + 2^{-22}$. Find the machine numbers on the Marc-32 that are just to the right and just to the left of x. Determine $fl(x)$, the absolute error $|x - fl(x)|$, and the relative error $|x - fl(x)|/|x|$. Verify that the relative error in this case does not exceed 2^{-24}.

14. Which of these is not necessarily true on the Marc-32? (Here x, y, and z are machine numbers and $|\delta| \leq 2^{-24}$.)

a. $fl(xy) = xy(1 + \delta)$

b. $fl(x + y) = (x + y)(1 + \delta)$

c. $fl(xy) = (xy)/(1 + \delta)$

d. $|fl(xy) - xy| \leq |xy|2^{-24}$

e. $fl(x + y + z) = (x + y + z)(1 + \delta)$

29. What's the unit roundoff error for a decimal machine that allocates 12 decimal places to the mantissa? Such a machine stores numbers in the form $x = \pm r \times 10^n$ with $1/10 \leq r < 1$.

Problem 2.2

2. How many bits of precision are lost in a computer when we carry out the subtraction $x - sinx$ for $x = 1/2$?

6. Find a way of computing $\sqrt{x^4 + 4} - 2$ without undue loss of significance.

15. Consider the function $f(x) = x^{-1}(1 - cosx)$.

a. what is the correct definition of $f(0)$; that is, the value that makes f continuous?

b. Near what points is there a loss of significance if the given formula is used?

c. How can we circumvent the difficulty in part b? Find a method that does not use the Taylor series.

d. If the new formula that you gave in part c involves subtractive cancellation at some other point, describe how to avoid that difficulty.

17. If at most 2 bits of precision are to be lost in the computation $y = sqrtx^2 + 1 - 1$, what restriction must be placed on x?