Section 0.2:

Problem 1c. Convert 79 to Binary.

Solution:

$$\begin{aligned} 79/2 &= 39r1 \\ 39/2 &= 19r1 \\ 19/2 &= 9r1 \\ 9/2 &= 4r1 \\ 4/2 &= 2r0 \\ 2/2 &= 1r0 \\ 1/2 &= 0r1 \end{aligned} \tag{1}$$

Answer = 1001111; □

Problem 3d. Convert 12.8 to Binary

Solution:

$$\begin{aligned} 12/2 &= 6r0 \\ 6/2 &= 3r0 \\ 3/2 &= 1r1 \\ 1/2 &= 0r1 \end{aligned} \qquad \begin{aligned} 0.8 * 2 &= 1.6 \\ 0.6 * 2 &= 1.2 \\ 0.2 * 2 &= 0.4 \\ 0.4 * 2 &= 0.8 \end{aligned}$$

Answer = $1100.\overline{1100}$ □

Problem 4b. convert 2/3 to binary

Solution:

$$\begin{aligned} \overline{.66} * 2 &= 1.\overline{33} \\ \overline{.33} * 2 &= 0.\overline{66} \\ \overline{.66} * 2 &= 1.\overline{33} \\ \overline{.33} * 2 &= 0.\overline{66} \end{aligned}$$

Answer = $.\overline{10}$ □

Problem 7b. convert 1011.101 to Decimal

Solution:

$$2^3 + 2^1 + 2^0 + 2^{-1} + 2^{-3} = 11\frac{5}{8} \tag{2}$$

□

Problem 7c. convert $10111.\overline{01}$ to decimal

Solution:

$$\begin{aligned} 2^4 + 2^2 + 2^1 + 2^0 &= 23 \\ 2^2 x - x = 01 \therefore x &= \frac{1}{3} \end{aligned} \tag{3}$$

Answer $= 23\frac{1}{3}$ □

Section 0.3:

Problem 1b. convert 1/3 to Binary rounding to the nearest rule.

Solution:

$$\begin{aligned} 1/3 = 0.\overline{33} * 2 &= 0.\overline{66} \\ \overline{.66} * 2 &= 1.\overline{33} \\ \overline{.33} * 2 &= 0.\overline{66} \\ \overline{.66} * 2 &= 1.\overline{33} \\ &\dots \end{aligned}$$

Answer = .01010101010101010101010101010101010101010101010101 Does not round because 53 bit would be 0. □

Problem 1d. Convert the following base 10 numbers to binary and express each as a floating point number fl(x) by using the Rounding to Nearest Rule:

Solution:

$$\begin{aligned} 0.9 * 2 &= 1.8 \\ 0.8 * 2 &= 1.6 \\ 0.6 * 2 &= 1.2 \\ 0.2 * 2 &= 0.4 \\ 0.8 * 2 &= 1.6 \end{aligned}$$

Answer = 11100110011001100110011001100110011001100110011001101 Does round because 53rd bit would be 1. □

Problem 2b. Convert the following base 10 numbers to binary and express each as a floating point number $fl(x)$ by using the Rounding to Nearest Rule:

Solution:

$$\begin{aligned} 9/2 &= 4r1 \\ 4/2 &= 2r0 \\ 2/2 &= 1r0 \\ 1/2 &= 0r1 \end{aligned} \qquad \begin{aligned} .6 * 2 &= 1.2 \\ .2 * 2 &= 0.4 \\ .4 * 2 &= 0.8 \\ .8 * 2 &= 1.6 \end{aligned}$$

$1001.\overline{1001} = 1 * 2^3$ The last 53 bit would be 1; therefore, round up.
Answer = $1.00110011001100110011001100110011001100110011001100110011\underline{10}$

□

Problem 3. For which positive integers k can the number $5 + 2^{-k}$ be represented exactly (with no rounding error) in double precision floating point arithmetic?

Solution:

$$5/2 = 2r1$$
$$2/2 = 1r0$$
$$1/2 = 0r1$$

Answer = $1.01 * 2^2$; 52 - 2 = 50; The numbers $k = \{x : 0 < x \leq 50\}$ will not produce an overflow. □

Problem 5a. Do the following sums by hand in IEEE double precision computer arithmetic, using the Rounding to Nearest Rule. (Check your answers, using MATLAB.) $(a)(1 + (2 - 51 + 2 - 53)) - 1$

Solution:

$$\begin{aligned}
&1.00000000000000000000000000000000000000000000000000010\\
+&1.0000000000000000000000000000000000000000000000000001\underline{1}\\
&1.00000000000000000000000000000000000000000000000000010\\
-&1.00000000000000000000000000000000000000000000000000000\\
=&0.00000000000000000000000000000000000000000000000000010
\end{aligned}$$

On line two, the col 53 gets truncated; Therefore, the answer is $1 * 2^{-51}$

□

Section 0.4:

Problem 1c. Identify for which values of x there is subtraction of nearly equal numbers, and find an alternate form that avoids the problem.

Solution: For values that are small; $x \approx 0$; The formula needs to be changed. We can revaluate the formula by multiplying by the conjugate.

$$\begin{aligned}
&= \frac{1}{1+x} - \frac{1}{1-x}\\
&= \frac{1-x-(1+x)}{(1+x)(1-x)}\\
&= \frac{-2x}{x^2-1}
\end{aligned} \tag{4}$$

□

Problem 2. Find the roots of the equation $x^2 + 3x - 8^{-14} = 0$ with three-digit accuracy.

Solution: Using the quadratic formula, we find the roots of the problem to be:

$$\frac{-3 \pm \sqrt{9 - 4(8^{-14})}}{2} \tag{5}$$

When the sign is negative, we get

$$x_1 = \frac{-3 - \sqrt{9 - 4(8^{-14})}}{2} \tag{6}$$

if the sign is positive, we must revaluate the equation because the answer will be very close to 0.

$$x_2 = \frac{-3 + \sqrt{9 - 4(8^{-14})}}{2} \tag{7}$$

We can muliply the formula by $(3+\sqrt{9 - 48^{-}14})$ and simplify. We are able to get the answer to 3 significant digits.

$$x_2 = \frac{-2(8^{-14})}{3 + \sqrt{9 - 4(8^{-14})}} \tag{8}$$

Therefore, the answers are $x_1 = 2.99, x_2 = -7.58;$ □