HW Problem 5: Representing numbers using unsigned fixed-point representation

Monday, September 21, 2020 8:15 PM

(A) For the number $M = +98.6$ determine the round-off approximation $\overline{\beta_r}$ for M, the rounded value $\overline{M}$, and the round-off error ΔM using the $[16, -3]$ unsigned fixed-point representation scheme.

$N = M * 2^f = 98.6 * 2^3 = 788.8$

$M_0 = \frac{788}{2^3} = 98.5$

$M_1 = \frac{789}{2^3} = 98.625$

$\Delta M_0 = |98.6 - 98.5| = 0.1$

$\Delta M_1 = |98.6 - 98.625| = 0.025$

$\Delta M_1 < \Delta M_0 \rightarrow \overline{M} = M_1 = 98.625$

$\overline{\beta_r} = 0000\ 0011\ 0001\ 0101$

(B) For the number $M = +98.6$ determine the round-off approximation $\overline{\beta_r}$ for M, the rounded value $\overline{M}$, and the round-off error ΔM using the $[16, -5]$ unsigned fixed-point representation scheme.

$N = M * 2^f = 98.6 * 2^5 = 3155.2$

$M_0 = \frac{3155}{2^5} = 98.59375$

$M_1 = \frac{3156}{2^5} = 98.625$

$\Delta M_0 = |98.6 - 98.59375| = 0.00625$

$\Delta M_1 = |98.6 - 98.625| = 0.025$

$\Delta M_0 < \Delta M_1 \rightarrow \overline{M} = M_0 = 98.59375$

$\overline{\beta_r} = 0000\ 1100\ 0101\ 0011$

(C) For the number $M = +98.6$ determine the round-off approximation $\overline{\beta_r}$ for M, the rounded value $\overline{M}$, and the round-off error ΔM using the $[16, -8]$ unsigned fixed-point representation scheme.

$$N = M * 2^f = 98.6 * 2^8 = 25241.6$$

$$M_0 = \frac{25241}{2^8} = 98.59765625$$
$$M_1 = \frac{25242}{2^8} = 98.6015625$$

$$\Delta M_0 = |98.6 - 98.59765625| = 0.00234375$$
$$\Delta M_1 = |98.6 - 98.6015625| = 0.0015625$$

$$\Delta M_1 < \Delta M_0 \rightarrow \overline{M} = M_1 = 98.6015625$$

$$\overline{\beta_r} = 0110\ 0010\ 1001\ 1010$$

(D) For the number $M = +98.6$ determine the round-off approximation $\overline{\beta_r}$ for M, the rounded value $\overline{M}$, and the round-off error ΔM using the $[32, -24]$ unsigned fixed-point representation scheme.

$$N = M * 2^f = 98.6 * 2^{24} = 1654233497.6$$

$$M_0 = \frac{1654233497}{2^{24}} = 98.5999999964237$$
$$M_1 = \frac{1654233498}{2^{24}} = 98.600000023842$$

$$\Delta M_0 = |98.6 - 98.5999999964237| = 3.5763E^{-8}$$
$$\Delta M_1 = |98.6 - 98.600000023842| = 2.3842E^{-8}$$

$$\Delta M_1 < \Delta M_0 \rightarrow \overline{M} = M_1 = 98.600000023842$$

$$\overline{\beta_r} = 0110\ 0010\ 1001\ 1001\ 1001\ 1001\ 1001\ 1010$$

(E) Find the smallest n and f such that $M = +98.6$ has an $[n, -f]$ unsigned fixed-point representation.

For the integer bits to represent 98, $2^6 < 98 < 2^7$, so 7 bits would be required for the integer part. However, for the fractional part of M, there is no finite number of bits that will represent 0.6.

0.6 = $\frac{3}{5}$, So there are two cases to be able to represent 0.6

1) $\frac{3}{5} = \frac{k}{2^x}$, where k is an integer and x is the number of bits required. With some manipulation it can be seen that $5k = 3 * 2^x$. There is no such integer K to satisfy this expression, which can hopefully be reasoned through without a rigorous proof.
2) $\frac{1}{5} = \frac{k}{2^x}$, where again k is an integer and x is bits required. In this case a multiple of this expression could be used to obtain the representation of $\frac{3}{5}$. However, the same issue occurs where for $5k = 2^x$, there is no integer k to satisfy.

This means that there is no exact finite fixed-point representation for 98.6.