HW Problem 4: Representing numbers using signed fixed-point representation

Wednesday, September 16, 2020 1:23 PM

(A) For the number $M = -468.421875$ determine the round-off approximation $\overline{\beta_r}$ for M, the rounded value $\overline{M}$, and the round-off error ΔM using the $[16, -3]$ (signed) fixed-point representation scheme.

N = M * 2^f = -468.421875 * 2^3 = -3747.375

$$M_0 = \frac{-3748}{2^3} = -468.375$$
$$M_1 = \frac{-3747}{2^3} = -468.5$$

$\Delta M_0 = |-468.375 - -468.421875| = 0.078125$
$\Delta M_1 = |-468.5 - -468.421875| = 0.046875$
$\Delta M_1 < \Delta M_0$, therefore $\overline{M} = M_1 = -468.375$ with error $\Delta M = 0.046875$
$\overline{\beta_r} = 1111\ 0001\ 0101\ 1101$

(B) For the number $M = -468.421875$ determine the round-off approximation $\overline{\beta_r}$ for M, the rounded value $\overline{M}$, and the round-off error ΔM using the $[16, -5]$ (signed) fixed-point representation scheme.

N = M * 2^f = -468.421875 * 2^5 = -14989.5

$$M_0 = \frac{-149890}{2^5} = -468.4375$$
$$M_1 = \frac{-14989}{2^5} = -468.40625$$

$\Delta M_0 = |-468.4375 - -468.421875| = 0.015625$
$\Delta M_1 = |-468.40625 - -468.421875| = 0.015625$
$\Delta M_0 = \Delta M_1$, therefore $\overline{M} = M_0 = -468.4375$ with error $\Delta M = 0.015625$ In this case, the error it the same for both M_0 and M_1 so both would be equally approximate representations of M.
$\overline{\beta_r} = 1100\ 0101\ 0111\ 0010$

(C) For the number $M = -468.421875$ determine the round-off approximation $\overline{\beta_r}$ for M, the rounded value $\overline{M}$, and the round-off error ΔM using the $[16, -8]$ (signed) fixed-point representation scheme.

For [16, -8] fixed point representation, 16-f bits are integer bits. In this example, there are 8 integer bits. The most negative number in a signed two's complement number representation is -2^{n-1} which in this case would be $-2^7 = -128$. Since M is smaller than -128, the closest approximation of M in [16, -8] fixed point representation is -128.

$\overline{M} = -128$
$\Delta M = |-468.40625 - -128| = 340.421875$
$\overline{\beta_r} = 1111\ 1111\ 1000\ 0000$

(D) Find the smallest n and f such that $M = -468.421875$ has an $[n, -f]$ (signed) fixed-point representation.

For the integer part of M, the smallest representable number in a signed two's complement representation is -2^{n-1}. In this case, $(-2^{9-1} = -256) < -468 < (-2^{10-1} = -512)$. This means that the smallest number of integer bits that can represent -468 is 10.

For the fractional part of M, $-0.421875 = -\frac{27}{64}$. Since $-\frac{27}{64}$ is irreducible and $\frac{1}{64} = \frac{1}{2^6}$, then 6 bits is the smallest amount of bits required to have enough precision to represent M.

This means that the smallest amount of n bits required to represent -468.421875 is 16 bits in [10, -6] fixed point representation.