Conversión de decimal a coma flotante

(Autor: Rodrigo Russo)

Ejercicio: Convertir 5,0205 a coma flotante IEEE con 32 bits (1 bit de signo, 11 bits de mantisa, 4 bits de exponente con exceso $2^{4-1} - 1 = 7$).

1. Convierto la parte entera (5) de decimal a binario. El procedimiento consiste en realizar la división por 2 utilizando el número y cada uno de los cocientes y tomar el resto de abajo hacia arriba:

$$\begin{array}{cccc}
5 & \multicolumn{1}{|c}{2} & & \\ \cline{2-2}
\mathbf{1} & 2 & \multicolumn{1}{|c}{2} & \\ \cline{3-3}
 & \mathbf{0} & 1 & \multicolumn{1}{|c}{2} \\ \cline{4-4}
 & & \mathbf{1} & 0
\end{array}$$

   $5_{10} = 101_2$.

2. Convierto la parte decimal a binario. Para esto, se multiplica sucesivamente por dos hasta caer en 1 o truncando la operación cuando sea conveniente y se toma, desde arriba hacia abajo, la parte entera de la multiplicación. Al llegar a un valor mayor que uno, se descarta la parte entera:

$$\begin{aligned}
0,0205 \times 2 &= \mathbf{0},0410 \\
0,0410 \times 2 &= \mathbf{0},0820 \\
0,0820 \times 2 &= \mathbf{0},1640 \\
0,1640 \times 2 &= \mathbf{0},3280 \\
0,3280 \times 2 &= \mathbf{0},6360 \\
0,6360 \times 2 &= \mathbf{1},2720 \\
0,2720 \times 2 &= \mathbf{0},5440 \\
0,5440 \times 2 &= \mathbf{1},0880 \\
0,0880 \times 2 &= \mathbf{0},1760 \\
0,1760 \times 2 &= \mathbf{0},3520 \\
0,3520 \times 2 &= \mathbf{0},7040 \\
0,7040 \times 2 &= \mathbf{1},4080 \\
0,4080 \times 2 &= \mathbf{0},8160 \\
0,8160 \times 2 &= \mathbf{1},6320 \\
0,6320 \times 2 &= \mathbf{1},2640 \\
0,2640 \times 2 &= \mathbf{0},5280 \\
0,5280 \times 2 &= \mathbf{1},0560 \\
0,0560 \times 2 &= \mathbf{0},1120
\end{aligned}$$

   $0,0205_{10} = 0,000001010001011010\ldots_2$

3. Junto parte entera y parte decimal:

$$5,0205_{10} \approx 101,00000101000_2$$

4. Utilizando la expresión de la convención de IEEE $x = (-1)^s \cdot 1, m \cdot 2^{\epsilon - z}$ donde s corresponde al signo (0 positivo, 1 negativo), m la mantisa, ϵ el exponente y z el exceso; se lleva el número obtenido a esa fórmula.

$$101,0000010100 \cdot 2^{-2} = 1,010000010100$$
$$101,0000010100 = 1,010000010100 \cdot 2^{2}$$

$$2 = \epsilon - 7 \Rightarrow \epsilon = 2 + 7 = 9$$

$$101,0000010100 = (-1)^0 1,010000010100 \cdot 2^{9-7}$$

$$\boxed{\begin{aligned} s &= 0 \\ m &= 010000010100 \\ \epsilon &= 9_{10} = 1001_2 \end{aligned}}$$

Número convertido: $\underbrace{0}_{s}\underbrace{01000001010}_{m}\underbrace{1001}_{\epsilon}$

Procedimiento inverso: binario a decimal.

1. $x = (-1)^s \cdot 1, m \cdot 2^{\epsilon - 7}$. Por el número a convertir se conoce que $s = 0$ y $\epsilon = 1001_2 = 9_{10}$.

$$\begin{aligned} x &= (-1)^0 \cdot 1,01000001010 \cdot 2^{9-7} \\ &= 1,01000001010 \cdot 2^2 \\ &= 101,000001010_2 \end{aligned}$$

2. Convertir a decimal:

$$\begin{aligned} 101,000001010_2 &= 2^2 + 2^0 + 2^{-6} + 2^{-8} \\ &= 4 + 1 + \frac{1}{64} + \frac{1}{256} \\ &= 5,01953125_{10} \end{aligned}$$

3. Comparo el número original (5,0205) con el guardado en el formato de coma flotante (5,01953125):

$$\epsilon_A = |5{,}0205 - 5{,}01953125| = 0,00096875$$

El número original y el guardado resultan ser levemente distintos, **¿por qué?**