Sistemas de Representación de números fraccionarios: Punto Fijo

Organización de Computadoras 2019

Universidad Nacional de Quilmes

1. Introducción

El sistema numérico de punto fijo es una generalización de los sistemas enteros que permite representar números fraccionarios. En un sistema de punto fijo, se destinan una cierta cantidad de bits a la parte entera y el resto a la parte fraccionaria, considerando que existe un punto (o coma) que los separa, tal como en el sistema decimal. Sin embargo, en los sistemas binarios de punto fijo, el punto no está representado explícitamente, sino que se asume una posición determinada.

Por ejemplo, se puede construir un sistema binario sin signo con punto fijo de 8 bits, considerando 5 bits para la parte entera y 3 bits para la parte fraccionaria.

En cuanto a la notación además del valor **n** que describe la cantidad total de bits se agrega un valor adicional **m** que hace referencia a la cantidad de bits fraccionarios enteros, quedando cada sistema de la siguiente manera:

- BSS(n,m): sistema de punto fijo en BSS con n bits en total, de los cuales m son fraccionarios.
- SM(n,m): sistema de punto fijo en SM con n bits en total, de los cuales 1 es de signo, y m de parte fraccionaria. Esto implica que $n - m - 1$ corresponden a la parte entera.

Los bits fraccionarios, al igual que los enteros, tienen un peso, pero están asociados a una potencia de 2 negativa, así, el primer bit fraccionario (de izquierda a derecha) tiene un peso de 2^{-1}, el segundo de 2^{-2}, y así sucesivamente hasta el último bit fraccionario. Entonces, para interpretar una cadena en punto fijo, se debe interpretar por un lado la parte entera en el sistema correspondiente y por otro lado la parte fraccionaria.

2. Interpretación

Para **Interpretar** una cadena en punto fijo se tiene dos mecanismos. El mecanismo **Literal** (o *por partes*) y el mecanismo **Escalado**.

mecanismo por partes

El mecanismo literal requiere que se trabaje por un lado la parte entera en sistema BSS y por otro lado la parte fraccionaria, aplicando las potencias negativas mencionadas previamente.

Por ejemplo, queremos interpretar la cadena `10011101` en Bss(8,3). Esto quiere decir que se toman 5 bits para la parte entera (de los 8 bits en total):

$$1 \cdot 2^4 + 0 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0$$

y 3 bits para la parte fraccionaria:

$$1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3}$$

Quedando la composición de ambas partes de la siguiente manera:

$$I_{BSS(8,3)}(10011101) = 1 \cdot 2^4 + 0 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 + 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3}$$

$$= 16 + 2 + 1 + 0,5 + 0,125 = 19,625$$

Por otra parte el mecanismo **Escalado** aprovecha la interpretación BSS adaptando el valor que se obtiene a la posición de la **coma desplazada**, es decir, m lugares. Este desplazamiento de coma tiene en cuenta la escala entre el sistema BSS(n) y el *mecanismo escalado*. Por lo que el escalado termina resultando 2^{-m}.

cadena $\longrightarrow^{I_bss}$ **valor entero** $\longrightarrow^{escala}$ **valor fraccionario**

Aplicando este mecanismo al ejemplo anterior, la cadena `10011101` se interpreta en BSS como:

$$I_{BSS(8,3)}(10011101) = 1 \cdot 2^7 + 0 \cdot 2^6 + 0 \cdot 2^5 + 1 \cdot 2^4 + 1 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0$$

$$128 + 16 + 8 + 4 + 1 = 157$$

Finalmente se escala el valor entero:

$$157 \cdot 2^{-3} = \frac{157}{2^3} = \frac{157}{8} = 19,625$$

(notar que 2^{-3} es la escala del sistema, con respecto a BSS)
Pasando en limpio:

`10011101` $\longrightarrow^{I_{bss}}$ **157** $\longrightarrow^{\times 2^{-3}}$ **valor 19,625**

3. Representación y error

mecanismo por partes

Tal como ocurre con la interpretación, la representación también tiene estos dos posibles mecanismos.

El mecanismo por partes requiere partir el problema y representar por un lado la parte entera (usando la representación de BSS) y por otro lado la parte fraccionaria realizando sucesivas multiplicaciones.

Por ejemplo para representar el valor **3,8** en el sistema BSS(7,4) necesitamos:

1. Representar el **valor 3** sólo usando los bits enteros (3 bits): `011`

2. Construir la cadena fraccionaria que aproxime el valor **0,8** multiplicando por 2 tantas veces como bits fraccionarios se tiene, y en cada multiplicación separar la parte entera del resultado, que es lo que corresponde a cada nuevo bit.
   Volviendo al ejemplo:

   a) 0,8 * 2 = **1**,6 $\Rightarrow b_{-1} = 1$

   b) 0,6 * 2 = **1**,2 $\Rightarrow b_{-2} = 1$

   c) 0,2 * 2 = **0**,4 $\Rightarrow b_{-3} = 0$

   d) **0**,4 * 2 = **0**,8 $\Rightarrow b_{-4} = 0$

   En este caso multiplicamos 4 veces x 2 porque disponemos de 4 bits para la parte fraccionaria.

   Este mecanismo establece un criterio de corte basado en la cantidad de bits disponibles, pero como en el caso del sistema decimal, esto puede obtener una cadena que no aproxima de la mejor manera. Por ejemplo, en el sistema decimal si contamos con sólo 2 dígitos fraccionarios para aproximar el valor 0,009, existe una gran diferencia entre aproximarlo con 0,00 o aproximarlo con 0,01. En el ultimo caso se llevó a cabo un **redondeo** mientras que en el primer caso se truncó la cadena ocasionando más pérdida.

   redondeo

   Entonces, para evitar que la cadena de bits quede truncada se hace un paso más (m+1 pasos en total) a los fines del redondeo, y si el bit obtenido en este último paso es un `1`, se suma el valor 1 a la cadena de bits fraccionaria resultante.
   Siguiendo el ejemplo, el último paso sería:

   0,8 * 2 = **1**,6 $\Rightarrow b_{-5} = 1$

   Como en este último paso obtuvimos un 1, debemos sumar el valor 1 a la cadena fraccionaria original (sin este último bit), quedando la cadena fraccionaria resultante de la siguiente manera:

   $1100 + 1 = 1101$

3. Finalmente ambas subcadenas se componen en la cadena resultante: `011 1101`

Una vez finalizada la representación es oportuno controlar el trabajo realizado aplicando la interpretación (con cualquiera de los mecanismos) sobre la cadena obtenida. En el ejemplo:

$$I_{bss(7,4)}(0111101) = 2^1+2^0+2^{-1}+2^{-2}+2^{-4} = 2+1+0,5+0,25+0,0625 = 3,8125$$

Sin embargo, se quería representar el valor **3,8**. Esto ocurre porque no todos los números del rango son representables en el sistema, por lo tanto puede haber un **error de representación** que es el valor absoluto de la diferencia entre el número que se deseaba representar y el número que efectivamente se logró representar. En este ejemplo, la diferencia es:

error de representacion

$$|3,8 - 3,8125| = 0,0125$$

.

Mecanismo escalado

El **mecanismo escalado** requiere llevar el valor a representar a un entero para luego aplicar la representación de BSS (R_{bss}). Dicho entero se obtiene al escalar y redondear el valor original, teniendo en cuenta, como en la interpretación que el valor de escala es 2^{-m}.

$$\textbf{valor fraccionario} \longrightarrow^{escala} \textbf{valor entero} \longrightarrow^{R_{bss}} \textbf{cadena}$$

Por ejemplo, representar **el valor 3,8** en BSS(7,4)

1. Escalar: $3,8 * 2^4 = 60,8$
2. Redondeo: $60,8 \approx 61$
3. $R_{bss(7)}(61) = 0111101$

$$\textbf{valor 3,8} \longrightarrow^{\times 2^4} \textbf{60,8} \longrightarrow^{redondeo} \textbf{61} \longrightarrow^{R_{bss}} \texttt{0111101}$$

4. Rango y Resolución

Al igual que en los sistemas enteros (BSS, SM y CA2), el rango en Punto Fijo está determinado por el intervalo de números representables. Por ejemplo, el rango del sistema BSS(2,1) está dado por:

Mínimo : $I_{bss(2,1)}(00) = 0$

Máximo : $I_{bss(2,1)}(11) = 2^0 + 2^{-1} = 1 + 0,5 = 1,5$

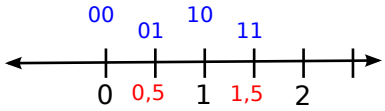
Y el del sistema sm (4,2) por:

Mínimo :$I_{sm(4,2)}(1111) = -(2^0 + 2^{-1} + 2^{-2}) = -(1 + 0,5 + 0,5) = -1,75$

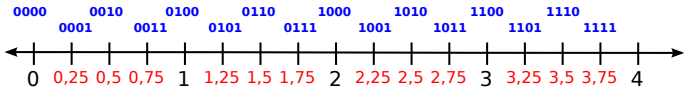
Máximo :$I_{sm(4,2)}(0111) = 2^0 + 2^{-1} + 2^{-2} = 1 + 0,5 + 0,5 = 1,75$

Sin embargo, el rango se refiere a todos los números representables en el intervalo. En los sistemas enteros esto es trivial (todos los enteros del intervalo), pero en punto fijo, para determinar exactamente qué números están representados dentro del intervalo es necesario el concepto de **resolución**: *la resolución es la distancia entre dos números representables consecutivos.*

Por ejemplo en el sistema bss(2,1) la resolución es 0,5 y por lo tanto los números representables son:



Y en el sistema bss (4,2)la resolución es 0,25, y los números representables son:



Referencias

[1] Williams Stallings, *Computer Organization and Architecture*, octava edicion, Editorial Prentice Hall, 2010. **Capitulo 8: Artimética del Computador, subcapítulo 8.2: Representación en coma fija.**