Sistema de punto flotante

Floating Point System

Autor 1: Camilo Eduardo Muñoz Albornoz
Risaralda, Universidad Tecnológica de Pereira, Pereira, Colombia
Correo-e: camilo.munoz2@utp.edu.co

I. INTRODUCCIÓN

Se dara a conocer el tema sobre sistemas de punto flotante y ejemplos de este

II. CONTENIDO

¿Qué es el sistema de punto flotante?

La representación de coma flotante (en inglés floating point, que significa «punto flotante») es una forma de notación científica usada en los computadores con la cual se pueden representar números reales extremadamente grandes y pequeños de una manera muy eficiente y compacta, y con la que se pueden realizar operaciones aritméticas. El estándar actual para la representación en coma flotante es el IEEE 754.

¿Por qué son necesarios los números de punto flotante?

Como la memoria de los ordenadores es limitada, no puedes almacenar números con precisión infinita, no importa si usas fracciones binarias o decimales: en algún momento tienes que cortar. Pero ¿cuánta precisión se necesita? ¿Y dónde se necesita? ¿Cuántos dígitos enteros y cuántos fraccionarios?

Para un ingeniero construyendo una autopista, no importa si tiene 10 metros o 10.0001 metros de ancho ─ posiblemente ni siquiera sus mediciones eran así de precisas.

Para alguien diseñando un microchip, 0.0001 metros (la décima parte de un milímetro) es una diferencia enorme ─ pero nunca tendrá que manejar distancias mayores de 0.1 metros.

Un físico necesita usar la velocidad de la luz (más o menos 300000000) y la constante de gravitación universal (más o menos 0.0000000000667) juntas en el mismo cálculo.

Para satisfacer al ingeniero y al diseñador de circuitos integrados, el formato tiene que ser preciso para números de órdenes de magnitud muy diferentes. Sin embargo, solo se necesita precisión relativa. Para satisfacer al físico, debe ser posible hacer cálculos que involucren números de órdenes muy dispares.

Básicamente, tener un número fijo de dígitos enteros y fraccionarios no es útil ─ y la solución es un formato con un punto flotante.

Notación científica

Como la representación en coma flotante es casi idéntica a la notación científica tradicional, con algunos añadidos y algunas diferencias, se describirá la notación científica para entender cómo funciona,

Representación

La notación científica se usa para representar números reales. Siendo r el número real a representar, la representación en notación científica está compuesta de tres partes:

$$r = c \cdot b^e$$

Fecha de Recepción: (Letra Times New Roman de 8 puntos)
Fecha de Aceptación: Dejar en blanco

Donde

- c es el coeficiente, formado por un número real con un solo dígito entero seguido de una coma (o punto) y de varios dígitos fraccionarios.
- b es la base, que en nuestro sistema decimal es 10, y en el sistema binario de los computadores es 2.
- e es el exponente entero, el cual eleva la base a una potencia.

Coeficiente

Un signo en el coeficiente indica si el número real es positivo o negativo.
El coeficiente tiene una cantidad determinada de dígitos significativos, los cuales indican la precisión del número representado, cuantos más dígitos tenga el coeficiente, más precisa es la representación. Por ejemplo, π lo podemos representar en notación científica, con 3 cifras significativas, $3{,}14 \times 10^0$, o con 12 cifras significativas, $3{,}14159265359 \times 10^0$ teniendo en la segunda representación mucha más precisión que la primera.

Base y exponente

El coeficiente es multiplicado por la base elevada a un exponente entero. En nuestro sistema decimal la base es 10. Al multiplicar el coeficiente por la base elevada a una potencia entera, lo que estamos haciendo es desplazando la coma del coeficiente tantas posiciones (tantos dígitos) como indique el exponente. La coma se desplaza hacia la derecha si el exponente es positivo, o hacia la izquierda si es negativo.

Ejemplo de cómo cambia un número al variar el exponente de la base:

- $2{,}71828 \times 10^{-2}$ representa al número real 0,0271828
- $2{,}71828 \times 10^{-1}$ representa al número real 0,271828
- $2{,}71828 \times 10^{0}$ representa al número real 2,71828 (el exponente cero indica que la coma no se desplaza)
- $2{,}71828 \times 10^{1}$ representa al número real 27,1828
- $2{,}71828 \times 10^{2}$ representa al número real 271,828
-

Ejemplo

Un ejemplo de número en notación científica es el siguiente:
$-1{,}234\ 567\ 89 \times 10^3$

El coeficiente es -1,23456789, tiene 9 dígitos significativos, y está multiplicado por la base diez elevada a la 3. El signo del coeficiente indica si el número real representado por la notación científica es positivo o negativo.

El valor de la potencia nos indica cuántas posiciones (cuántos dígitos) debe ser desplazada la coma del coeficiente para obtener el número real final. El signo de la potencia nos indica si ese desplazamiento de la coma debe hacerse hacia la derecha o hacia la izquierda. Una potencia positiva indica que el desplazamiento de la coma es hacia la derecha, mientras que un signo negativo indica que el desplazamiento debe ser hacia la izquierda. Si el exponente es cero, la coma no se desplaza ninguna posición. La razón de la denominación de "coma flotante", es porque la coma se desplaza o "flota" tantos dígitos como indica el exponente de la base, al cambiar el exponente, la coma "flota" a otra posición.

En el número representado en la notación científica anterior, $-1{,}23456789 \times 10^3$, el exponente es 3 positivo, lo que indica que la coma del coeficiente -1,23456789 debe ser desplazada 3 posiciones hacia la derecha, dando como resultado el número real equivalente: -1234,567 89

¿Cómo funcionan los números de punto flotante?

La idea es descomponer el número en dos partes:

- Una mantisa (también llamada coeficiente o significando) que contiene los dígitos del número. Mantisas negativas representan números negativos.

- Un exponente que indica dónde se coloca el punto decimal (o binario) en relación al inicio de la mantisa. Exponentes negativos representan números menores que uno.

Este formato cumple todos los requisitos:

- Puede representar números de órdenes de magnitud enormemente dispares (limitado por la longitud del exponente).
- Proporciona la misma precisión relativa para todos los órdenes (limitado por la longitud de la mantisa).
- Permite cálculos entre magnitudes: multiplicar un número muy grande y uno muy pequeño conserva la precisión de ambos en el resultado.
- Los números de coma flotante decimales normalmente se expresan en notación científica con un punto explícito siempre entre el primer y el segundo dígitos. El exponente o bien se escribe explícitamente incluyendo la base, o se usa una e para separarlo de la mantisa.

Ejemplo

Mantisa	Exponente	Notación científica	Valor en punto fijo
1.5	4	$1.5 \cdot 10^4$	15000
-2.001	2	$-2.001 \cdot 10^2$	-200.1
5	-3	$5 \cdot 10^{-3}$	0.005
6.667	-11	6.667e-11	0.0000000000667

El estándar

Casi todo el hardware y lenguajes de programación utilizan números de punto flotante en los mismos formatos binarios, que están definidos en el estándar IEEE 754. Los formatos más comunes son de 32 o 64 bits de longitud total:

Formato	Bits totales	Bits significativos	Bits del exponente	Número más pequeño	Número más grande
Precisión sencilla	32	23 + 1 signo	8	$\sim 1.2 \cdot 10^{-38}$	$\sim 3.4 \cdot 10^{38}$
Precisión doble	64	52 + 1 signo	11	$\sim 5.0 \cdot 10^{-324}$	$\sim 1.8 \cdot 10^{308}$

Hay algunas peculiaridades:

- La secuencia de bits es primero el bit del signo, seguido del exponente y finalmente los bits significativos.
- El exponente no tiene signo; en su lugar se le resta un desplazamiento (127 para sencilla y 1023 para doble precisión). Esto, junto con la secuencia de bits, permite que los números de punto flotante se puedan comparar y ordenar correctamente incluso cuando se interpretan como enteros.
- Se asume que el bit más significativo de la mantisa es 1 y se omite, excepto para casos especiales.
- Hay valores diferentes para cero positivo y cero negativo. Estos difieren en el bit del signo, mientras que todos los demás son 0. Deben ser considerados iguales aunque sus secuencias de bits sean diferentes.
- Hay valores especiales no numéricos (NaN, «not a number» en inglés) en los que el exponente es todo unos y la mantisa no es todo ceros. Estos valores representan el resultado de algunas operaciones indefinidas (como multiplicar 0 por infinito, operaciones que involucren NaN, o casos específicos). Incluso valores NaN con idéntica secuencia de bits no deben ser considerados iguales.

1. Las notas de pie de página deberán estar en la página donde se citan. Letra Times New Roman de 8 puntos

Sistema binario con punto flotante

Un valor real se puede extender con una cantidad arbitraria de dígitos. La coma flotante permite representar solo una cantidad limitada de dígitos de un número real, solo se trabajará con los dígitos más significativos, (los de mayor peso) del número real, de tal manera que un número real generalmente no se podrá representar con total precisión sino como una aproximación que dependerá de la cantidad de dígitos significativos que tenga la representación en coma flotante con que se está trabajando. La limitación se halla cuando existen dígitos de peso menor al de los dígitos de la parte significativa. En dicho caso estos suelen ser redondeados, y si son muy pequeños son truncados. Sin embargo, y según el uso, la relevancia de esos datos puede ser despreciable, razón por la cual el método es interesante pese a ser una potencial fuente de error.

En la representación binaria de coma flotante, el bit de mayor peso define el valor del signo, 0 para positivo, 1 para negativo. Le siguen una serie de bits que definen el exponente. El resto de bits son la parte significativa.

Debido a que la parte significativa está generalmente normalizada, en estos casos, el bit más significativo de la parte significativa siempre es 1, así que no se representa cuando se almacena sino que es asumido implícitamente. Para poder realizar los cálculos ese bit implícito se hace explícito antes de operar con el número en coma flotante. Hay otros casos donde el bit más significativo no es un 1, como con la representación del número cero, o cuando el número es muy pequeño en magnitud y rebasa la capacidad del exponente, en cuyo caso los dígitos significativos se representan de una manera denormalizada para así no perder la precisión de un solo golpe sino progresivamente. En estos casos, el bit más significativo es cero y el número va perdiendo precisión poco a poco (mientras que al realizar cálculos este se haga más pequeño en magnitud) hasta que al final se convierte en cero.

Recordemos que para formar números con coma o parte fraccionaria se sigue un mismo mecanismo, sólo que con exponentes negativos, ya que:
$10^{-1} = 1/10 = 0,1$
$10^{-2} = 1/100 = 0,01$
$10^{-3} = 1/1.000 = 0,001$
$10^{-4} = 1/10.000 = 0,0001$
$10^{-5} = 1/100.000 = 0,00001$

Cada una de estas potencias de 10 representan una columna de la parte fraccionaria o decimal de un número (hacia la derecha de la coma) y son multiplicadas por un dígito ubicado en dicha columna; después se suman todos los productos de las multiplicaciones para obtener la parte fraccionaria del número.

Por ejemplo 125,144 se forma de la siguiente manera:

$(\mathbf{1} \times 10^{2}) + (\mathbf{2} \times 10^{1}) + (\mathbf{5} \times 10^{0}) + (\mathbf{1} \times 10^{-1}) + (\mathbf{4} \times 10^{-2}) + (\mathbf{4} \times 10^{-3})$
$= (\mathbf{1} \times 100) + (\mathbf{2} \times 10) + (\mathbf{5} \times 1) + (\mathbf{1} \times 0,1) + (\mathbf{4} \times 0,01) + (\mathbf{4} \times 0,001)$
$= \mathbf{100} + \mathbf{20} + \mathbf{5} + \mathbf{0,1} + \mathbf{0,04} + \mathbf{0,004} = 125,144$

En binario es exactamente igual, sólo que en lugar de potencias de 10 se utilizan potencias de 2 con exponente negativo para representar la parte fraccionaria de un número, o sea la parte que se encuentra por sobre la derecha de la coma del número.

$2^{-1} = 1/2 = 0,5$
$2^{-2} = 1/4 = 0,25$
$2^{-3} = 1/8 = 0,125$
$2^{-4} = 1/16 = 0,0625$
$2^{-5} = 1/32 = 0,03125$

Por ejemplo pasemos 10,1001 a sistema decimal:
$(\mathbf{1} \times 2^{1}) + (\mathbf{0} \times 2^{0}) + (\mathbf{1} \times 2^{-1}) + (\mathbf{0} \times 2^{-2}) + (\mathbf{0} \times 2^{-3}) + (\mathbf{1} \times 2^{-4})$
$= (\mathbf{1} \times 2) + (\mathbf{0} \times 1) + (\mathbf{1} \times 0,5) + (\mathbf{0} \times 0,25) + (\mathbf{0} \times 0,125) + (\mathbf{1} \times 0,0625)$
$= \mathbf{2} + \mathbf{0} + \mathbf{0,5} + \mathbf{0} + \mathbf{0} + \mathbf{0,0625} = 2,5625$

Por lo tanto 10,1001 binario es equivalente a 2,5625 en decimal.

Veamos otro ejemplo:

Pasemos 100,1 a sistema decimal:

($\mathbf{1} \times 2^2$) + ($\mathbf{0} \times 2^1$) + ($\mathbf{0} \times 2^0$) + ($\mathbf{1} \times 2^{-1}$)
= ($\mathbf{1}$ x 4) + ($\mathbf{0}$ x 2) + ($\mathbf{0}$ x 1) + ($\mathbf{0}$ x 0,5)
= **4** + **0** + **0** + **0,5** = 4,5

Pasemos 0,01 a sistema decimal:

($\mathbf{0} \times 2^0$) + ($\mathbf{0} \times 2^{-1}$) + ($\mathbf{1} \times 2^{-2}$)
= ($\mathbf{0}$ x 1) + ($\mathbf{0}$ x 0,5) + ($\mathbf{1}$ x 0,25)
= **0** + **0** + **0,25** = 0,25

Pasemos 100110,101110 a decimal:

($\mathbf{1} \times 2^5$) + ($\mathbf{0} \times 2^4$) + ($\mathbf{0} \times 2^3$) + ($\mathbf{1} \times 2^2$) + ($\mathbf{1} \times 2^1$) + ($\mathbf{0} \times 2^0$) + ($\mathbf{1} \times 2^{-1}$) + ($\mathbf{0} \times 2^{-2}$) + ($\mathbf{1} \times 2^{-3}$) + ($\mathbf{1} \times 2^{-4}$) + ($\mathbf{1} \times 2^{-5}$) + ($\mathbf{0} \times 2^{-6}$)
= ($\mathbf{1}$ x 32) + ($\mathbf{0}$ x 16) + ($\mathbf{0}$ x 8) + ($\mathbf{1}$ x 4) + ($\mathbf{1}$ x 2) + ($\mathbf{0}$ x 1) + ($\mathbf{1}$ x 0,5) + ($\mathbf{0}$ x 0,25) + ($\mathbf{1}$ x 0,125) + ($\mathbf{1}$ x 0,0625) + ($\mathbf{1}$ x 0,03125) + ($\mathbf{0}$ x 0,015625)
= **32** + **0** + **0** + **4** + **2** + **0** + **0,5** + **0** + **0,125** + **0,0625** + **0,03125** + **0,015625** = 38,734375

Por lo tanto 100110,101110 binario es equivalente a 38,734375 en decimal.

Ahora número decimal con coma flotante a binario.

- Dado un número decimal con coma; primero se debe convertir a binario la parte entera,
- Después se convierte la parte fraccionaria del número, multiplicandola por 2, si su resultado es igual o mayor a 1 se agrega un dígito 1 en la parte fraccionaria del número binario, si es menor a 1 se agrega un 0.
- Después se vuelve a multiplicar por 2 a la parte fraccionaria del número que obtuvimos en la multiplicación anterior; si su resultado es mayor o igual a 1 se agrega un 1 en la parte fraccionaria del número binario, si es menor a 1 se agrega un 0.
- Se repiten los pasos anteriores hasta quedarnos con 1,0 tras la última multiplicación. Si no se puede llegar a 1,0 se genera un número binario con dígitos periódicos o parte fraccionaria infinita.

Si queremos convertir 5,3125:

Primero convertimos la parte entera 5 a binario:

Obtenemos 5 = 101 (binario).

Resultado parcial **101,.....**

Ahora continuamos con la parte fraccionaria 0,3125:

- Multiplicamos 0,3125 x 2 = 0,625 >> Resultado parcial **101,0**
- Multiplicamos 0,625 x 2 = 1,25 >> Resultado parcial **101,01**
- Quitamos o restamos 1 de 1,25 y nos quedamos con la parte fraccionaria 0,25
- Multiplicamos 0,25 x 2 = 0,5 >> Resultado parcial **101,010**
- Multiplicamos 0,5 x 2 = 1,0 >> Resultado final **101,0101**

Cada vez que obtenemos un producto mayor a 1, agregamos 1 en la parte fraccionaria del número binario y quitamos o restamos 1 del producto quedándonos solamente con la parte fraccionaria. Dicha parte fraccionaria debe ser multiplicada por 2 y se repite lo mismo de antes; si tenemos un producto menor a 1 se agrega 0 en la parte fraccionaria del número binario, de lo contrario

1. Las notas de pie de página deberán estar en la página donde se citan. Letra Times New Roman de 8 puntos

agregamos 1 y restamos una unidad del producto obtenido y volvemos a multiplar por 2 su parte fraccionaria. Este proceso se repite hasta llegar a un producto igual a 1,0; con lo que la conversión a binario concluye.

Ejemplo

Vamos a convertir 2,125 a binario:

Primero hay que convertir la parte entera y seguimos con la parte fraccionaria:

2 >> Resultado parcial 10, (parte entera)
0,125 x 2 = 0,25 >> Resultado parcial 10,0
0,25 x 2 = 0,5 >> Resultado parcial 10,00
0,5 x 2 = 1,0 >> Resultado final 10,001

Ejemplo
convertir 0,4 a binario:

0 >> Resultado parcial **0,** (parte entera)
0,40 x 2 = 0,80 Resultado parcial **0,0**
0,80 x 2 = 1,60 Resultado parcial **0,01**
0,60 x 2 = 1,20 Resultado parcial **0,011**
0,20 x 2 = 0,40 Resultado parcial **0,0110**
0,40 x 2 = 0,80 Resultado parcial **0,01100**
0,80 x 2 = 1,60 Resultado parcial **0,011001**
0,60 x 2 = 1,20 Resultado parcial **0,0110011**
0,20 x 2 = 0,40 Resultado parcial **0,01100110**
0,40 x 2 = 0,80 Resultado parcial **0,011001100**
0,80 x 2 = 1,60 Resultado parcial **0,0110011001**
0,60 x 2 = 1,20 Resultado parcial **0,01100110011**
0,20 x 2 = 0,40 Resultado parcial **0,011001100110**
0,40 x 2 = 0,80 Resultado parcial **0,0110011001100**

Como se puede ver, convertir 0,40 de decimal a binario, nos da un número periódico; en donde *0011* se repite indefinidamente. En este ejemplo llegamos hasta **0,0110011001100** que es igual a 0,39990234375. Cuanto más continuemos más nos acercaremos a 0,40; sin embargo como se trata de un número binario periódico, no importa la cantidad de dígitos fraccionarios, siempre se acercará a 0,40 pero no llegará a ser exactamente igual; por lo que se redondea luego de un cierto número de dígitos.

Esto se puede probar de la siguiente manera:

$0{,}0110011001100 = (0 \times \mathbf{2^{0}}) + (0 \times \mathbf{2^{-1}}) + (1 \times \mathbf{2^{-2}}) + (1 \times \mathbf{2^{-3}}) + (0 \times \mathbf{2^{-4}}) + (0 \times \mathbf{2^{-5}}) + (1 \times \mathbf{2^{-6}}) + (1 \times \mathbf{2^{-7}}) + (0 \times \mathbf{2^{-8}}) + (0 \times \mathbf{2^{-9}}) + (1 \times \mathbf{2^{-10}}) + (1 \times \mathbf{2^{-11}}) + (0 \times \mathbf{2^{-12}}) + (0 \times \mathbf{2^{-13}})$

$= 0 + 0 + 0{,}25 + 0{,}125 + 0 + 0 + 0{,}015625 + 0{,}0078125 + 0 + 0 + 0{,}0009765625 + 0{,}00048828125 + 0 + 0$

$= 0{,}39990234375$

Que se puede redondear a 0,4

REFERENCIAS

[1]. Wikipedia. Coma flotante. [Online]. Available https://es.wikipedia.org/wiki/Coma_flotante

[2]. La guia del punto flotante. NUMEROS DE PUNTO FLOTANTE. [Online]. Available: http://puntoflotante.org/html

[3]. G. Threepwood. (2002, Jun.). Cómo funciona el sistema de numeración binario 2. [Online]. Available: https://www.youbioit.com/es/article/shared-information/8228/como-funciona-el-sistema-de-numeracion-binario-2

1. Las notas de pie de página deberán estar en la página donde se citan. Letra Times New Roman de 8 puntos