4. Representación de números en coma flotante

Introducción

Para representar números enteros muy grandes o números fraccionarios muy pequeños es necesarios utilizar muchos bits

La notación científica es una forma muy útil de expresar estos números sin aumentar demasiado la cantidad de bits usados

Los objetivos de este tema son:

- Introducir la representación de números en coma flotante, que se basa en la notación científica
- Describir las operaciones básicas de la aritmética utilizando números en coma flotane

Estructura del tema

Introducción

Números en coma flotante

Representación de números binarios en coma flotante

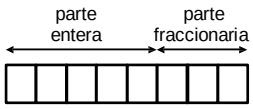
Aritmética binaria en coma flotante

Resumen y bibliografía

Números en coma fija

Los números en coma fija son aquellos números en los que la coma fraccionaria tiene una posición fija

Por ejemplo, si tenemos números de 8 bits podemos dedicar 5 bits a la parte entera y 3 a la fraccionaria



Esto supone una severa limitación

- No podemos usar más de 5 bits para la parte entera aunque la fraccionaria sea cero
- No podemos usar más de 3 bits para la parte fraccionaria aunque la entera sea cero

Números en coma flotante

El sistema de numeración en coma flotante está basado en la notación científica que se utiliza para representar números muy grandes o muy pequeños

Un número en coma flotante tiene tres partes:

- Signo: indica si el número es positivo o negativo
- Mantisa: representa la magnitud del número
- Exponente: indica el desplazamiento de la coma fraccionaria

$$+/- \text{ Mantisa} \times \text{Base}^{\text{exponente}}$$

Números en coma flotante

Los números decimales en coma flotante se normalizan, desplazando la coma fraccionaria de manera que la parte entera del número siempre valga cero

$$241506800 = 0'2415068 \times 10^{9}$$

No es necesario representar la base del número, ya que está implícita en el formato

Dado que la parte entera de un número normalizado siempre es cero, tampoco es necesario representarla

signo:	+
mantisa:	2415068
exponente:	9

Estructura del tema

Introducción

Números en coma flotante

Representación de números binarios en coma flotante

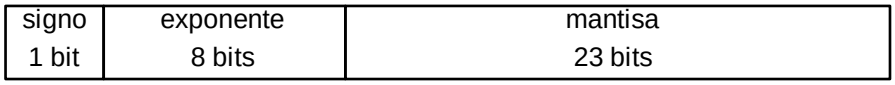
Aritmética binaria en coma flotante

Resumen y bibliografía

Números binarios en coma flotante

El formato utilizado para la representación de números binarios en coma flotante está definido por el estándar 514-1985 ANSI/IEEE

Nos centraremos en el formato de simple precisión, que utiliza 32 bits para la representación



Existen otros formatos similares pero con distinto número de bits, como el de doble precisión (64 bits)

Mantisa en simple precisión

La magnitud de un número binario se almacena en la mantisa de forma normalizada

$$0'0000101 \qquad 0'101 \times 2^{-4}$$

La forma normalizada siempre tiene un 1 en la posición más a la izquierda, por lo que no hay que almacenarlo, está implícito en la representación

Incluso podemos aprovechar otro bit si normalizamos el número de manera que el 1 más a la izquierda esté en la parte entera

$$0'0000101 \qquad 0'101 \times 2^{-4} \quad 1'01 \times 2^{-5}$$

Exponente en simple precisión

Para simplificar el formato, sería interesante evitar la necesidad de un bit de signo adicional para representar el exponente

Por este motivo, el exponente se representa usando un formato denominado en exceso, que consiste en sumar un desplazamiento al valor real del número

- En los formatos de exponente en coma flotante, este desplazamiento se calcula restando 1 al entero más grande que se pueda representar y luego dividiendo por 2
- Si utilizamos 8 bits para representar el exponente, hay 2^8 combinaciones posibles, desde 0 hasta 255
- El desplazamiento será (255 – 1) / 2 = 127

Exponente en simple precisión

El formato de simple precisión tiene un exponente de 8 bits en formato exceso 127

Este formato se utiliza para representar exponentes con valores entre –126 y 127

exponente =	-126	-126 + 127 = 1	00000001
exponente =	-125	-125 + 127 = 2	00000010
		...	
exponente =	0	0 + 127 = 127	01111111
exponente =	1	1 + 127 = 128	10000000
		...	
exponente =	126	126 + 127 = 253	11111101
exponente =	127	127 + 127 = 254	11111110

Casos particulares en simple precisión

El valor –127 (00000000) del exponente se reserva para representar dos casos especiales

- Si la mantisa también es cero se está representando el cero
- Si la mantisa es distinta de cero se trata de un número no normalizado, es decir, supondremos que el bit implícito de la parte entera no es 1 sino 0, y el exponente es –126

El valor 128 (11111111) del exponente se reserva para representar dos casos especiales

- Si la mantisa es cero se está representando el infinito
- Si la mantisa es distinta de cero se está representando el resultado de una operación inválida, es decir, el valor no es un número (NaN – Not a Number)

Ejemplos en simple precisión

Utilizaremos como ejemplo el número +1011010010001

Se trata de un número positivo, por lo que el bit de signo será un 0

El siguiente paso consiste en normalizar el número

1011010010001 $1'011010010001 \times 2^{12}$

Esto nos permite calcular el valor de la mantisa

mantisa 011010010001

Ejemplos en simple precisión

Ahora falta por calcular el valor del exponente

exponente 12

exponente 12 + 127 = 139

exponente 10001011

Con esta información podemos representar el número

1 bit	8 bits	23 bits
0	10001011	01101001000100000000000

Ejemplos en simple precisión

Ahora haremos un ejemplo en sentido inverso usando el número

1	10010001	10001110001000000000000

Dado que el bit de signo es 1, el número es negativo

La mantisa nos proporciona la magnitud del número

1'10001110001

Ejemplos en simple precisión

El valor del exponente se calcula de la siguiente forma:

exponente 10010001

exponente 145

exponente 145 - 127 = 18

Con esta información obtenemos el número

$$-1'10001110001 \times 2^{18}$$

-11000111000100000000

Ventajas de la coma flotante

El formato en coma flotante se usa para representar valores muy grandes (exponente positivo) o valores muy pequeños (exponente negativo)

El formato en coma flotante de simple precisión permite representar números de 128 bits con solo 32

El formato es flexible: se pude dedicar cualquier cantidad de bits a la parte entera o a la parte fraccionaria según convenga

Estructura del tema

Introducción

Números en coma flotante

Representación de números binarios en coma flotante

Aritmética binaria en coma flotante

Resumen y bibliografía

Suma y resta en coma flotante

La suma y la resta se llevan a cabo con las reglas habituales, pero los dos exponentes deben ser iguales

Antes de operar hay que igualar los exponentes

- Se desplaza a la derecha la mantisa del número con menor exponente
- Cada posición desplazada a la derecha implica incrementar en uno el valor del exponente
- El desplazamiento se repite hasta que los dos números tengan el mismo valor en el exponente

Suma en coma flotante

En la suma existe la posibilidad de desbordamiento a infinito (overflow) del valor de la mantisa

El desbordamiento se debe corregir desplazando la mantisa una posición a la derecha y sumando uno al exponente

Es importante controlar la posibilidad de que el exponente se desborde al realizar esta operación, ya que obtendríamos un número no representable

Resta en coma flotante

En la resta existe la posibilidad de obtener un número con uno o varios ceros en los bits más significativos, lo que obligaría a normalizar el resultado

La normalización se realiza desplazando la mantisa a la izquierda y restando uno al exponente por cada posición desplazada

Es importante controlar la posibilidad de que el exponente se desborde a cero (underflow)

- Una posible solución sería usar el formato no normalizado
- Si este formato tampoco es suficiente, entonces el número no es representable

Multiplicación en coma flotante

La multiplicación es más sencilla porque no es necesario que los exponentes sean iguales

- Las mantisas se multiplican como enteros en coma fija
- Los exponentes se suman
- Si cualquiera de los operandos es cero, el resultado también

Existe la posibilidad de que sea necesario normalizar el resultado de la multiplicación

También es posible que se produzca un desbordamiento a cero o a infinito del exponente, por lo que hay que controlar que el resultado sea representable

División en coma flotante

La división consiste en dividir las mantisas y restar al exponente del dividendo el exponente del divisor

- Si el dividendo es cero el resultado es cero
- Si el divisor es cero se considera desbordamiento
- Si el dividendo y el divisor son cero, el resultado se identifica como un número desconocido

Existe la posibilidad de que sea necesario normalizar el resultado de la división

También es posible que se produzca un desbordamiento a cero o a infinito del exponente, por lo que hay que controlar que el resultado sea representable

Estructura del tema

Introducción

Números en coma flotante

Representación de números binarios en coma flotante

Aritmética binaria en coma flotante

Resumen y bibliografía

Resumen

Los números en coma flotante permiten representar números muy grandes o muy pequeños con una cantidad relativamente reducida de bits

Existe un formato para representar números binarios en coma flotante, lo que determina cómo se realizan las operaciones aritméticas con ellos

Los números en coma flotante son muy frecuentes en las aplicaciones de cálculo numérico, por lo que los ordenadores suelen utilizar unidades separadas específicas para operaciones en coma flotante

Bibliografía

Sistemas Digitales - Principios y Aplicaciones - Tocci - 10th