Guía de ejercicios # 7 - Punto Flotante

Organización de Computadoras 2017

UNQ

1 Motivación

Punto flotante llega a nosotros por una razón: necesitamos tener resolución variable. La resolución es la distancia entre dos números consecutivos. Por ejemplo, si midiesemos la resolución en los numeros representables con 2 digitos, de los cuales 1 es fraccionario (por ejemplo 0,5; 1,4; 3,2) la resolución es de 0,1: ya que si tomamos dos números consecutivos (9,8 y 9,9) la diferencia es 0,1.
El problema de punto fijo viene cuando queremos representar numeros muy grandes o numeros muy chicos con el mismo sistema: Suponiendo que mi resolución es de 0,5

1. Si representamos el número 100.000,3 obtendremos el número 100.000,5 (un error relativo muy bajo)
2. Si representamos el número 0,26 obtendremos el número 0,5 (un error relativo muy grande)

Entonces lo que hacemos es tener resolución variable: Para números chicos, tenemos una mejor resolución y para números más grandes la resolución es mas baja. La forma de lograrlo es añadiendo el exponente a nuestras representaciones:

$$m \times 2^e$$

Donde m es la mantisa y e es el exponente. Al cambiar el exponente estamos cambiando la resolución.

2 Interpretación

Interpretemos la cadena 1101101 en el siguiente sistema de punto flotante:

mantisa BSS(4)	exponente BSS(3)

1. Primero interpretamos la mantisa, tomando los primeros 4 bits, como indica el formato: 1101. Como el sistema de la mantisa es BSS, el resultado es 13.
2. Luego interpretamos el exponente, tomando los últimos 3 bits, como indica el formato: 101. Como el sistema del exponente es BSS, el resultado es 5.
3. Por último reemplazamos los resultados obtenidos en la formula

$$m \times 2^e$$

quedando como resultado final:

$$13 \times 2^5$$

2.1 Ejercicios

1 Interpretar las siguientes cadenas de bits en el sistema dado:
Donde:

mantisa $BSS(5)$	exponente $BSS(3)$

a) `1110 1110`
b) `1111 1111`
c) `1110 0000`
d) `0010 0000`
e) `0000 0100`

2 Interpretar las siguientes cadenas de bits en el sistema dado:
Donde:

mantisa $SM(5,4)$	exponente $CA2(3)$

a) `1110 1110`
b) `1111 1111`
c) `0110 0100`
d) `1110 0100`
e) `0010 0000`

3 Interpretar las siguientes cadenas de bits en el sistema dado:
Donde:

mantisa $SM(9,7)$	exponente $SM(7)$

a) `1110 1110 0101 1111`
b) `1111 1111 1111 1110`
c) `0110 0100 0001 1011`
d) `0110 0011 1100 0110`
e) `0010 0001 1000 1100`

3 Normalización

1. ¿Para qué sirve la Normalización de cadenas? ¿Cuál es su consecuencia?
   a) Para no tener 2 representaciones del 0
   b) Para perder la representación del 0
   c) Para no tener multiples representaciones de la mayoría de los números
   d) Para tener una mejor resolución máxima y mínima

2. Interpretar las siguientes cadenas de bits en el sistema dado: Donde: | mantisa $SM(10+1,10)$ | exponente $CA2(5)$ | Notar que los 10 bits de la magnitud de la mantisa son fraccionarios, 9 de ellos explícitos y uno implícito

   a) `010 0010 1110 1110`

   b) `111 1111 1111 1111`

   c) `111 1111 1110 0000`

   d) `000 0000 0010 0000`

   e) `000 0000 0000 0000`

   f) `100 0000 0000 0000`

   g) `000 0000 0111 0011`

   h) `000 0000 0001 1111`

   i) `000 0000 0011 1111`

3. Interpretar las cadenas del ejercicio anterior en un sistema de punto flotante con

   Mantisa: Normalizada y con bit implícito $SM(9+1,9)$

   Exponente: $SM(6)$

   Teniendo en cuenta el siguiente formato: | magnMant(8) | signoMant(1) | signoExp(1) | magnExp(5) | Notar que los 9 bits de la magnitud de la mantisa son fraccionarios, 8 de ellos explícitos y uno implícito

4 Rango

Para calcular el rango, en general, buscamos la cadena que representa al número más chico y la que representa al número más grande. La diferencia es que ahora tenemos que tener en cuenta el exponente, el cual puede tener un sistema distinto al de la mantisa. La cadena más grande es bastante intuitiva: utilizamos la mantisa positiva de mayor magnitud y el exponente positivo de mayor magnitud para calcularla. La cadena más chica utiliza la mantisa negativa de mayor magnitud y el exponente positivo de mayor magnitud.

Calculemos el rango para el siguiente sistema | $SM(4+1,4)$ | $CA2(4)$ |

La mantisa negativa de mayor magnitud es: 1 1111. Como todos sus bits son fraccionarios representa:

$$M_n = -(2^{-1} + 2^{-2} + 2^{-3} + 2^{-4})$$

La mantisa positiva de mayor magnitud es: 0 1111. Como todos sus bits son fraccionarios representa:

$$M_p = 2^{-1} + 2^{-2} + 2^{-3} + 2^{-4}$$

El exponente positivo de mayor magnitud es: 0111, y representa:

$$E = 2^0 + 2^1 + 2^2 = 7$$

Por lo tanto el rango es el siguiente:

$$Rango = [M_n \times 2^E; M_p \times 2^E]$$

4.1 Ejercicios

1. Calcular el rango de un sistema de punto flotante con

   Mantisa: $BSS(5)$

   Exponente: $BSS(3)$

2. Calcular el rango de un sistema de punto flotante con

   Mantisa: $SM(5,4)$

   Exponente: $CA2(3)$

3. Calcular el rango de un sistema de punto flotante con

   Mantisa: $SM(9,7)$

   Exponente: $SM(7)$

4. Calcular el rango de un sistema de punto flotante con

   Mantisa: Normalizada y con bit implícito $SM(5+1,4)$

   Exponente: $Ex(8,128)$

   . Notar que en la mantisa tenemos 5 bits + 1 implicito de los cuales sólo 4 son fraccionarios.

5. Calcular el rango de un sistema de punto flotante con

   Mantisa: Normalizada y con bit implícito $SM(4+1,4)$

   Exponente: $CA2(3)$

6. Calcular el rango de un sistema de punto flotante con

   Mantisa: Normalizada y con bit implícito $SM(9+1,9)$

   Exponente: $Ex(5,16)$

7. Calcular el rango de un sistema de punto flotante con

   Mantisa: Normalizada y con bit implícito $SM(7+1,6)$

   Exponente: $Ex(7,64)$

5 Resolucion variable

No es ninguna novedad: en punto flotante la resolución es variable. Si mantenemos la mantisa constante en cantidad de bits, la resolución depende del exponente. Al depender del exponente, tenemos muchas resoluciones (tantas como el sistema del exponente permita representar). Las que nos van a interesar son la máxima y la mínima, ya que estas nos hablan de la precisión del sistema.

Resolución Mínima: es la mínima diferencia que puedo lograr entre dos números consecutivos, tomando un sistema como base. Para calcularla, tengo que elegir dos cadenas consecutivas, ambas con el exponente negativo de mayor magnitud.

Resolución Máxima: es la máxima diferencia que puedo lograr entre dos números consecutivos, tomando un sistema como base. para calcularla, tengo que elegir dos cadenas consecutivas, ambas con el exponente positivo de mayor magnitud.

Calculemos la resolución máxima y mínima del siguiente sistema de punto flotante con

Mantisa: $SM(4+1,4)$

Exponente: $CA2(4)$

a) Para la resolución mínima necesito el mínimo exponente, es decir, 1000 en nuestro sistema. El cual representa, como ya vimos, al número -8.

Las cadenas que voy a utilizar son 0000 1000 y 0001 1000, las cuales son consecutivas y como las mantisas están normalizadas y tienen bit implícito, la primera vale

$$2^{-1} \times 2^{-8}$$

y la segunda

$$(2^{-1} + 2^{-4}) \times 2^{-8}$$

Para saber la resolución tengo que restar los valores y para eso distribuyo en el segundo valor, quedando:

$$2^{-1} \times 2^{-8} + 2^{-4} \times 2^{-8}$$

Y si resto los valores, me queda:

$$2^{-1} \times 2^{-8} + 2^{-4} \times 2^{-8} - 2^{-1} \times 2^{-8}$$

Por lo tanto, mi resolución mínima es:

$$2^{-4} \times 2^{-8}$$

b) Para la resolución máxima necesito el máximo exponente, es decir, 0111 en nuestro sistema. El cual representa, como ya vimos, al número 7.

Las cadenas que voy a utilizar son 0000 0111 y 0001 0111, las cuales son consecutivas y como las mantisas están normalizadas y tienen bit implícito, la primera vale

$$2^{-1} \times 2^{7}$$

y la segunda

$$(2^{-1} + 2^{-4}) \times 2^{7}$$

Para saber la resolución tengo que restar los valores y para eso distribuyo en el segundo valor, quedando:

$$2^{-1} \times 2^{7} + 2^{-4} \times 2^{7}$$

Y si resto los valores, me queda:

$$2^{-1} \times 2^{7} + 2^{-4} \times 2^{7} - 2^{-1} \times 2^{7}$$

Por lo tanto, mi resolución máxima es:

$$2^{-4} \times 2^{7}$$

5.1 Ejercicios

1. Calcular la resolución máxima y mínima de un sistema de punto flotante con

   Mantisa: $BSS(5)$
   Exponente: $BSS(3)$

2. Calcular la resolución máxima y mínima de un sistema de punto flotante con

   Mantisa: $SM(5,4)$
   Exponente: $CA2(3)$

3. Calcular la resolución máxima y mínima de un sistema de punto flotante con

   Mantisa: $SM(9,7)$
   Exponente: $SM(7)$

4. Calcular la resolución máxima y mínima de un sistema de punto flotante con

   Mantisa: Normalizada y con bit implícito $SM(5+1,4)$
   Exponente: $Ex(8,128)$

   .

5. Calcular la resolución máxima y mínima de un sistema de punto flotante con

   Mantisa: Normalizada y con bit implícito $SM(4+1,4)$
   Exponente: $CA2(3)$

6. Calcular la resolución máxima y mínima de un sistema de punto flotante con

   Mantisa: Normalizada y con bit implícito $SM(9+1,9)$
   Exponente: $Ex(5,16)$

7. Calcular la resolución máxima y mínima de un sistema de punto flotante con

   Mantisa: Normalizada y con bit implícito $SM(7+1, 6)$

   Exponente: $Ex(7, 64)$

6 IEEE 754

El estándar IEEE 754 es un estandar adoptado para la representación de números con punto flotante. Cuenta con 5 grupos de números:

1) **Numeros Normalizados:** esta familia contiene a todos los numeros cuyo exponente es distinto de 00000000 y distinto de 11111111

2) **Numeros Denormalizados:** esta familia contiene a todos las cadenas cuyo exponente es igual a 00000000 y la mantisa es distinta de 0

3) **Infinito:** una cadena representa infinito cuando su exponente es 11111111 y su mantisa es 0.

4) **Cero:** una cadena representa el cero cuando su exponente es 00000000 y su mantisa es 0.

5) **NaN:** Not a number, estos cadenas no representan un número, como su nombre lo dice. Una cadena es NaN cuando el exponente es 11111111 y la mantisa distinta de 0

En el estándar IEEE 754, la mantisa está en signo magnitud y el exponente en exceso. Son diferentes para números Normalizados y Denormalizados:

1) **Numeros Normalizados:** La Mantisa es Normalizada SM(24+1, 23) y el exponente Ex(8, 127)

2) **Numeros Denormalizados:** La Mantisa $SM(24, 23)$ y el exponente $Ex(8, 126)$

Por último, el estandar define dos precisiones:

Simple Precisión Mantisa $SM(24+1, 23)$ normalizada con bit implícito y exponente en exceso de $Ex(8, 127)$.

signoMant(1b)	exp(8b)	magnMant(23b)

Doble Precisión Mantisa $SM(53+1, 52)$ normalizada con bit implícito y exponente en exceso de $Ex(11, 1023)$.

signoMant(1b)	exp(11b)	magnMant(52b)

6.1 Ejercicios

1 Calcular el rango y la resolución máxima y mínima de los números normalizados de ambos formatos del estándar IEEE 754:

2 ¿Qué valores están representados por las siguientes cadenas en formato IEEE de simple precisión?

a) `0 11000100 00000000000000000000000`

b) `1 11111110 10100000000000000000000`

c) `0 00000000 00000000000000000000001`

d) `1 00000000 00100000000000000000000`

e) `1 00000000 00000000000000000000000`

f) `1 00100000 01000000000000000000000`

3 Escribir la siguiente subrutina:

```
;----------------------------extraerExponente
; REQUIERE En R5 y R6 un valor en IEEE simple
;      precision (en ese orden)
; MODIFICA ??
; RETORNA En los 8 bits de la derecha de R4, los
;      8 bits del exponente
;-------------------------------------------------
```

Por ejemplo, si la cadena IEEE almacenada en R5/R6 es `1 01010101 11110000111100001111000` entonces en R4 se debe obtener `00000000 01010101`

4 Ejecute el siguiente programa e indique el valor final de los registros R2 y R3

```
;----------------------------sumarSiEsNormalizado
; REQUIERE En R5 y R6 un valor en IEEE simple
;     precision. Asume que R7 es un contador.
; MODIFICA R7
; RETORNA suma 1 a R7 si es un número normalizado
;-------------------------------------------------

;---------------------sumarSiEsDenormalizado
; REQUIERE En R5 y R6 un valor en IEEE simple
;     precision. Asume que R7 es un contador.
; MODIFICA R7
; RETORNA suma 1 a R7 si es un número
;     denormalizado
;-------------------------------------------------

MOV R5, 0x0000
MOV R6, 0x0001
MOV R7, 0x0000
call sumarSiEsNormalizado
MOV R2, R7
MOV R7, 0x0000
call sumarSiEsDenormalizado
MOV R3, R7
```

5 Interpretar las siguientes cadenas (abreviadas en hexadecimal) mediante el estándar IEEE 754:

a) `C28FFF00`

b) `42E48000`

c) `00800000`

d) `40000000`

e) `45500430`

f) `3FE00000`

g) `C0066666`

h) `CFFFFF34`

6 Indicar si las siguientes afirmaciones son verdaderas o falsas. Justificar.

a) Los números desnormalizados en IEEE sirven para indicar que ocurrió una condición de error

b) En punto flotante la resolución es infinita.

c) Punto fijo tiene error de representación, mientras que punto flotante no.

d) Punto flotante tiene sólo dos resoluciones: maxima y mínima

e) La cadena mas grande en un sistema de punto flotante | mantisa SM(5,4) | exponente $CA2(4)$ | es 011110111

7 ¿Para qué sirve que la mantisa no esté normalizada cuando el exponente es 0 y la mantisa no es nula?

8 ¿Qué ventajas tiene la representación IEEE 754 en simple precisión sobre un sistema de mantisa fraccionaria normalizada con bit implícito | mantisa SM(24+1,24) | exponente $SM(8)$ | ?

9 ¿Cuántas resoluciones diferentes puedo tener en el siguiente sistema? | mantisa SM(5,4) | exponente $CA2(4)$ |

10 ¿Cuántas resoluciones diferentes puedo tener en el siguiente sistema? | mantisa SM(5,4) | exponente $SM(4)$ |

Referencias

(1) *Williams Stallings, Computer Organization and Architecture. Editorial Prentice Hall.* ***Capítulo 9, sección 4***