Digital II - Anexo Capítulo I

Sistemas de representación restringidos a n bits

Los sistemas digitales disponen de registros y buses de tamaños específicos que limitan la cantidad de bits disponibles para la representación de los datos. Es habitual mencionar que el sistema trabaja con datos de 8, 16, 32 bits, o en punto flotante de simple/doble precisión. Por lo tanto, las representaciones que se utilizan tienen limitaciones, y los cálculos están siempre sujetos a aproximaciones y por ende a errores. Para caracterizar los sistemas de representación y compararlos se definen tres parámetros importantes:

- *Capacidad de representación*: Es la cantidad de tiras distintas que se pueden representar. Por ejemplo, si tengo un sistema restringido a 5 bits, sería 2^5 tiras, es decir, 32.
- *Resolución*: Es la mínima diferencia entre un número representable y el siguiente. Por ejemplo, en binario con dos dígitos fraccionarios es 0.01.
- *Rango*: El rango de un sistema está dado por el número mínimo y el número máximo representables. Por ejemplo, en binario con cinco dígitos es [0, 31]

Vamos a ejemplificar la obtención de estos parámetros para dos sistemas de representación limitados a 8 bits, uno en punto flotante y otro en punto fijo.

Ejemplo 1: Sistema en punto flotante de 8 bits, con 4 bits de mantisa fraccionaria normalizada positiva sin signo y 4 de exponente positivo sin signo. Calcular capacidad de representación, rango y resolución.

- Capacidad de representación = $2^8 = 256$
- Máximo valor representable $= (.1111 \times 2^{1111})_2 = (0.5 + 0.25 + 0.125 + 0.0625) \times 2^{15} = 0.9375 \times 32768 = (30720)_{10}$
- Mínimo valor representable = $(0.1 \times 2^{0000})_2 = (0.5)_{10}$
- Rango = [0.5, ..., 0.9375×2^{15}] = [0.5,.., 30720]
- Resolución en el extremo superior: $R_s = (0.1111 - 0.1110) \times 2^{1111} = (0.0001 \times 2^{1111})_2 = (0.0625 \times 2^{15})_2 = (2048)_{10}$
- Resolución en el extremo inferior: $R_i = (0.1001 - 0.1000) \times 2^{0000} = (0.0001 \times 2^{0000})_2 = = (0.0625)_{10}$

Ejemplo 2: Consideremos enteros de 8 bits positivos sin signo. Calcular capacidad de representación, rango y resolución:

- Capacidad de representación = $2^8 = 256$
- Rango = [0, .., 255]
- Resolución en el extremo superior: $R_s = 255 - 254 = 1$
- Resolución en el extremo inferior: $R_i = 1 - 0 = 1$

Comparando los resultados obtenidos en los ejemplos anteriores podemos afirmar:

- El rango en la representación en punto flotante es mayor.
- La cantidad de combinaciones binarias distintas es la misma en ambos sistemas: $2^8 = 256$, ya que está definida por la cantidad de bits disponibles.
- El sistema en punto flotante presenta una resolución variable a lo largo del intervalo representable. El sistema en punto fijo, en cambio, tiene resolución constante.

Digital II - Anexo Capítulo I

Conclusiónes:

En el sistema de punto flotante el rango es mayor. Podemos representar números más grandes ó más pequeños que en un sistema de punto fijo ***(para igual cantidad de bits)***, pero pagamos el precio de que los números no están igualmente espaciados en el rango, como ocurre en punto fijo.
Si comparamos sistemas en punto flotante restringidos a n bits, y repartimos estos n bits en diferentes proporciones entre mantisa y exponente, podremos obtener otras conclusiones interesantes:

- Al incrementar los bits del exponente (y por ende reducir los de la mantisa), aumenta el rango representable pero empeora la resolución, ya que los números quedan más espaciados en el rango. Esto es lógico pues el número total de representaciones es siempre la misma (2n) y el rango aumenta. El desmejoramiento de la resolución trae aparejado un incremento en los errores.
- Cuando se distribuyen los bits entre mantisa y exponente en un sistema en punto flotante se debe llegar a una solución de compromiso entre rango y resolución.

Se recomienda a los alumnos que verifiquen las conclusiones con ejemplos.

Ejercicios:

Ejercicio 1) Indicar cuál es la capacidad de representación, la resolución y el rango de un sistema binario con 4 bits para la parte entera y 3 para la parte fraccionaria.

Ejercicio 2) Especificar la capacidad de representación, el rango y la resolución de un sistema de representación binaria entera con n bits, en C1, y en C2.

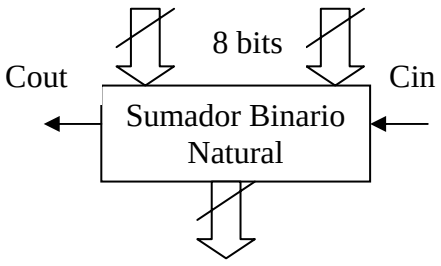
Ejercicio 3) Resuelva la siguiente operación trabajando en binario con la convención de complemento a 1, representando los números con 9 bits: $S = (2000)_4 - (128)_{10}$

a) ¿Es correcto el resultado? ¿Por qué?
b) Explique si lo comentado en el punto anterior implica algún inconveniente, si es así comente que solución propondría.

Ejercicio 4) Resuelva la siguiente operación utilizando el sistema binario con la convención de complemento a 1 con la mínima cantidad de bits para que no se presente overflow: $S = (- 672)_8 - (D9)_{16.}$ Repita el ejercicio para C2.

Ejercicio 5) Utilizando el circuito indicado, opere en binario usando C2 y C1 e indique en cada caso si el resultado es ***correcto***. Si no lo es, justifique con claridad el/los motivos que lo causan y explique, detallando en qué consiste, alguna modificación al circuito que permita obtener el resultado correcto.

- $S = - (24)_{16} - (43)_5$
- $S = - (167)_8 - (31)_4$



Ejercicio 6) ¿Qué valores están representados por las siguientes cadenas en formato de simple precisión de la norma IEEE 754?

- 0 11000100 00000000000000000000000

- 1 11111110 10100000000000000000000
- 0 00000000 00000000000000000000001

Ejercicio 7) Hallar la representación de los siguientes números en simple precisión:

- 1
- 13
- 257
- - 40000
- - 0,0625