Representación de números reales

- Los números reales se pueden representar de dos formas:
  - Punto fijo
  - Punto flotante
- Punto fijo:
  - La coma decimal se considera fija en un punto.
    - Ejemplo: Datos de 32 bits, utilizar 20 bits para la parte entera y 12 bits para los decimales
  - Fácil realizar las operaciones de sumas, restas, multiplicaciones y divisiones vistas hasta ahora
  - Notación: Qm,n
    - m: número de bits de la parte entera (opcional)
    - n: número de bits para la parte decimal
    - Se utiliza un bit adicional para el signo (en total hacen falta m+n+1 bits).
    - Ejemplos: Q16,16, Q.32, etc.

Representación de números reales

- Punto flotante:
  - La coma decimal es "flotante"
  - Se descompone el número en dos partes: mantisa y exponente:

    $$N = M \times b^E$$

    Ejemplos:
    - $2547{,}35_{10} = 2{,}54735 * 10^3$
    - $0{,}0035_{10} = 3{,}5 * 10^{-3}$
    - $111{,}0110_2 = 1{,}11011_2 * 2^2$
    - $0{,}001101_2 = 1{,}101_2 * 2^{-3}$
  - Se utiliza un número fijo de bits para la mantisa, otro para el exponente y otro adicional para el signo
  - Normalización: Fija la posición de la coma decimal en la descomposición (para tener una representación única)

Representación de números reales

- Standard IEEE 754: Precisión simple (32 bits)
  - Se utiliza 1 bit para el signo
  - Se utilizan 8 bits para el exponente (E)
  - Se utilizan 23 bits para la mantisa (M)
  - Números normalizados: $N= (-1)^s * 2^{E-127} * 1.M$
    - E puede tomar valores entre 1 y 254 (el exponente está desplazado por -127)
  - El cero se representa como todo ceros en los campos E y M (es una excepción)
  - Algunas otras excepciones: E = 255 denota infinito (utilizado en casos de overflow, por ejemplo)
  - También se pueden representar números no normalizados (E=0). La descomposición es diferente, la mantisa es 0.M
- Standard IEEE 754: Precisión doble (64 bits)
  - Representación similar
  - 1 bit para signo, 11 bits para exponente, 52 bits para mantisa

Representación de números reales

- Ejemplo: Representar -7.625_{10} en precisión sencilla

$$-7.625_{10} = -111.101_2$$

Descomponiendo en la forma $N= (-1)^s * 2^{E-127} * 1.M$:

$$-111.101_2 = (-1)^1 * 1.11101 * 2^2$$

$$S = 1$$

$$M = 11101$$

$$2 = E - 127 \rightarrow E = 129_{10} = 10000001_2$$

Por tanto el número representado con precisión sencilla:

1 10000001 11101000000000000000