Estándar de IEEE para Números en Coma Flotante

El estándar IEEE 754 para la representación de números en coma flotante contempla dos tipos básicos de precisiones: simple precisión (32 bits) y doble precisión (64 bits). En la Figura 1 se muestran estos dos formatos.

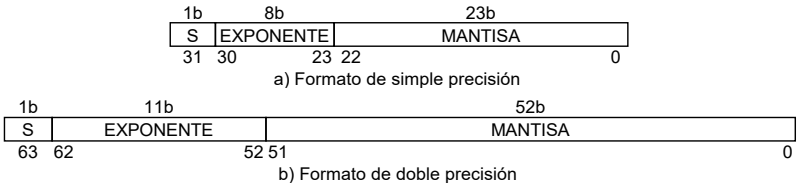

Figura 1. Formatos de simple y doble precisión.

El orden de los campos (signo, exponente y mantisa), así como la representación utilizada para cada uno de ellos están pensados para acelerar las operaciones de comparación. En ambos formatos se utiliza un bit para especificar el signo de la mantisa. La mantisa se representa en valor absoluto; es fraccionaria y normalizada con un uno implícito a la izquierda de la coma decimal. Esto quiere decir que las mantisas reales serán de la forma 1,XXXXXX… y que el primer uno nunca estará representado dentro del campo destinado a la mantisa. Al estar la mantisa normalizada los valores posibles se encontraran entre 1,0000… y 1,1111…

La base del exponente es siempre 2, y el exponente está representado en exceso a 2^{q-1}-1 donde q representa el número de bits que se destinan al campo de exponente (8 u 11). Lo que supone un exceso a 127 en el formato de simple precisión y un exceso a 1023 en el de doble precisión. Esta forma de representar el exponente, junto con la forma de ordenar los campos es lo que hace que se pueda comparar de una forma rápida dos números en este formato[1].

Así pues, si S, M y Exp son los contenidos de los campos de signo, mantisa y exponente de un número en este formato, el valor de ese número será:

- $(-1)^S \times 1{,}M \times 2^{Exp-127}$ en el caso de simple precisión y
- $(-1)^S \times 1{,}M \times 2^{Exp-1023}$ en el caso de doble precisión.

En general, si $m_{-1}m_{-2}m_{-3}m_{-4}\ldots$ son los bits del campo de mantisa leídos de izquierda a derecha y E es el valor del campo de exponente (Exp - 127 ó Exp - 1023, dependiendo de la representación) el valor de un número en el formato IEEE 754 será:

$$(-1)^S \times (1 + m_{-1} \times 2^{-1} + m_{-2} \times 2^{-2} + m_{-3} \times 2^{-3} + \ldots) \times 2^E$$

Nótese que en el caso de simple precisión el campo de exponente oscila entre los valores[2] $0 < exp < 255$ y esto a su vez implica que el exponente real se encuentre entre los

[1] Las comparaciones pueden ser llevadas a cabo utilizando técnicas de aritmética entera sin signo. Los signos se procesan por separado, los números negativos (S = 1) son menores que los positivos (S = 0). El exponente y la mantisa se pueden procesar concatenados como si se tratase de números enteros positivos, ya que el formato de representación del exponente (exceso a 2^{q-1}-1) así lo permite.

[2] Los valores 2^q-1 y 0 para el campo de exponente están reservados, ver Tabla 1.

valores $-126 \leq E \leq 127$. En el caso de doble precisión $0 < exp < 2047$ luego el valor real del exponente estará entre $-1022 \leq E \leq 1023$.

Este tipo de codificación es la que se emplea por defecto, sin embargo existen unos casos especiales para representar números muy pequeños, resultados de operaciones sin sentido, valores infinitos, etc. Estos casos especiales se reflejan en la Tabla 1.

E	*M*	*Valores*
2^q-1	$\neq 0$	NaN (No un Número)
2^q-1	0	$+\infty$ ó $-\infty$ según el signo S
0	0	Cero
0	$\neq 0$	Números desnormalizados

Tabla 1. Casos especiales en la codificación.

Cuando el campo de exponente está todo a unos y el campo de mantisa contiene algún bit distinto de cero se utiliza para representar valores para los que no existe representación posible como números fraccionarios, como el resultado de operaciones tales como 0/0 ó $\sqrt{-1}$.

Si el campo de exponente está todo a unos y el campo de mantisa contiene el valor cero se utiliza para representar el $+\infty$ o el $-\infty$ según el signo S de la mantisa.

Cuando tanto el campo de mantisa como el de exponente se encuentran a cero indican el valor cero que tiene dos representaciones, +0 y -0, dependiendo del signo de la mantisa.

La combinación con el campo de exponente a cero y el campo de mantisa distinto de cero se utiliza para representar aquellos números muy pequeños mediante un formato desnormalizado con valores de la forma $0,M \times 2^{-126}$ para el formato de simple precisión ó $0,M \times 2^{-1022}$ para el formato de doble precisión.

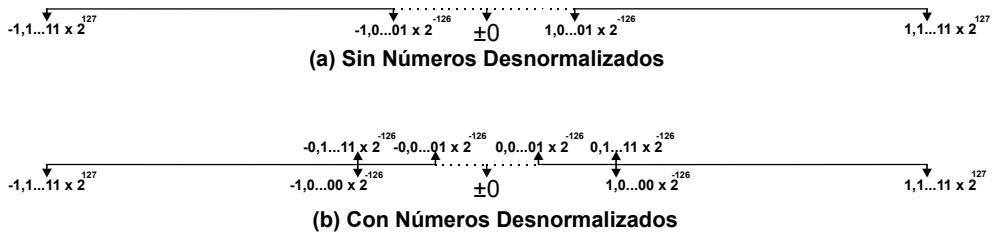

Figura 2. Representación de números en simple precisión.

Con este esquema de representación, tal y como se pone de manifiesto mediante la Figura 2 para el formato de simple precisión, se logra ampliar el rango de representación a costa de una menor precisión para los números más próximos a cero. En la Tabla 2 se puede observar el rango de representación que ofrece el estándar para los números en coma flotante en simple precisión.

S	*Exp*	*Mantisa*	*IEEE 754*
1	11…11	11…11	NaN
…	…	…	
1	11…11	00…01	
1	11…11	00…00	$-\infty$
1	11…10	11…11	($\approx -2 \times 2^{127}$)
…	…	…	Normalizados
1	00…01	00…00	($= -1 \times 2^{-126}$)
1	00…00	11…11	($\approx -1 \times 2^{-126}$)
…	…	…	Desnormalizados
1	00…00	00…01	($= -2^{-23} \times 2^{-126}$)
1	00…00	00…00	-0
0	00…00	00…00	+0
0	00…00	00…01	($= 2^{-23} \times 2^{-126}$)
…	…	…	Desnormalizados
0	00…00	11…11	($\approx 1 \times 2^{-126}$)
0	00…01	00…00	($= 1 \times 2^{-126}$)
…	…	…	Normalizados
0	11…10	11…11	($\approx 2 \times 2^{127}$)
0	11…11	00…00	$+\infty$
0	11…11	00…01	NaN
…	…	…	
0	11…11	11…11	

Tabla 2. Rango de representación para números en simple precisión.

Redondeo.

Muchas veces tras una operación con números en coma flotante se genera una mantisa M de longitud más larga (*p* bits) que la prevista en el formato (*m* bits). En esta mantisa resultado los *m* primeros bits de la mantisa M se llaman bits *retenidos.*

En general, nos podemos tras una operación con números en coma flotante nos podemos encontrar con los siguientes casos respecto de la mantisa:

- M es representable de forma *exacta* en el formato: los *p*–*m* bits no retenidos son 0 y se pueden eliminar: 010000 $\rightarrow$ 0100.
- M se encuentra entre dos valores representables M_- y M_+ ($M_- < M < M_+$) y hay que *redondear*: escoger uno de ellos como representación *inexacta* de M

La norma IEEE-745 admite cuatro *modos de redondeo*:

- Hacia $+\infty$
- Hacia $-\infty$
- Hacia 0
- Hacia el más próximo de los dos (este es el modo por omisión). En este método hay que resolver de alguna forma una situación de empate. En el caso que nos ocupa se redondea hacia la mantisa par, es decir la

que acabe en 0. La figura 3 incluye un ejemplo de este modo de redondeo.

M	se elige	M resultante
010000	(exacta)	**0100**
010001	M_-(más próxima)	**0100**
010010	M_- (par)	**0100**
010011	M_+(más próxima)	**0101**
010100	(exacta)	**0101**
010101	M_-(más próxima)	**0101**
010110	M_+ **(par)**	**0110**
010111	M_+(más próxima)	**0110**
011000	(exacta)	**0110**

Figura 3. Ejemplo de redondeo de la mantisa sesgado al par.