Sistema de representación de coma flotante IEEE 754

31 Marzo 2011

El estándar IEEE 754 define representaciones para números de coma flotante, con diferentes tipos de precisión: simple y doble[1], utilizando anchos de palabra de 32 y 64 bits, respectivamente. Estas representaciones son las que utilizan los procesadores de la familia x86, entre otros.

Estos sistemas, a diferencia de los anteriores, permiten representar también valores especiales, los cuales serán tratados posteriormente.

En la representación IEEE 754 de 32 bits, el bit de más alto peso es utilizado para almacenar el signo de la mantisa, los siguientes 8 bits guardan la representación del exponente, y los primeros 23 bits almacenan la mantisa. El exponente se representa en CD con frontera no equilibrada ($\frac{b^d}{2} - 1$), cuyo valor es 127.

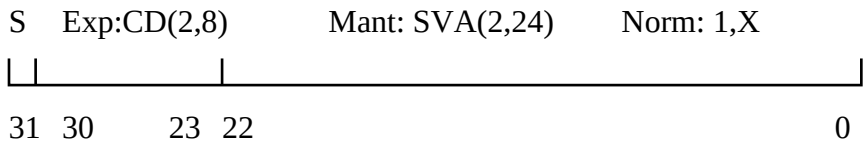

Figura 1: Representación de coma flotante IEEE 754 de 32 bits.

De manera similar, en la representación IEEE 754 de 64 bits, el bit de más alto peso es utilizado para almacenar el signo de la mantisa, los siguientes 11 bits representan el exponente, y los primeros 52 bits almacenan la mantisa. El exponente se representa en CD con frontera no equilibrada ($\frac{b^d}{2} - 1$), cuyo valor es 1023.

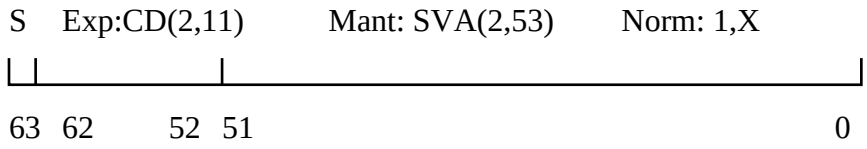

Figura 2: Representación de coma flotante IEEE 754 de 64 bits.

En ambos casos se tiene la mantisa normalizada como "1,X", que quiere decir "1, seguido del valor de la mantisa", donde el 1 que se encuentra a la izquierda de la coma está implícito u oculto, es decir, no es almacenado en la representación, permitiendo así ganar precisión.

La tabla 1 describe los parámetros usados en las representaciones de precisión simple y doble.

[1] En realidad, el estándar IEEE 754 también provee representaciones en otras precisiones que no serán tratadas en este apunte.

	Precisión Simple	Precisión Doble
Cantidad de bits de la mantisa *	24	53
Cantidad de bits del exponente	8	11
e_{min}	-126	-1022
e_{max}	127	1023
Cantidad total de bits	32	64

** Incluyendo el 1 oculto.*

Cuadro 1: Parámetros utilizados en las diferentes precisiones.

La ecuación 1 indica cómo se obtiene el valor representado r a partir de la mantisa y el exponente.

$$r = 1, m \times 10_b^e \tag{1}$$

Las ecuaciones 2 y 3 muestran un ejemplo de conversión de un número a las representaciones IEEE 754 de 32 y 64 bits, respectivamente.

$-58,875 = -111010,111_b = -1,11010111_b \times 10_b^5$

En IEEE 754 de precisión simple

$$\begin{aligned} m = &\quad \underline{1}\,1\,\underbrace{11010111_b}_{\text{explícito}} \qquad \text{SVA(2,24)} \\ e = &\quad 5 = 10000100_b \qquad \text{CD(2,8)}, \qquad f = 127 = 01111111_b \\ r_1 = &\quad 1\,10000100_b\,11010111_b \\ r_1 = &\quad C26B8000_h \end{aligned} \tag{2}$$

En IEEE 754 de precisión doble

$$\begin{aligned} m = &\quad \underline{1}\,1\,\underbrace{11010111_b}_{\text{explícito}} \qquad \text{SVA(2,53)} \\ e = &\quad 5 = 10000000100_b \qquad \text{CD(2,11)}, \qquad f = 1023 = \quad 01111111111_b \\ r_2 = &\quad 1\,10000000100_b\,11010111_b \\ r_2 = &\quad C04D700000000000_h \end{aligned} \tag{3}$$

Representación de valores especiales

Una cuestión de interés es analizar qué sucede cuando una operación arroja como resultado un número indeterminado o un complejo. En estos casos el resultado constituye un valor especial para el sistema y se almacena como NaN (Not a Number), tal como ocurre al hacer, por ejemplo, ∞/∞ ó $\sqrt{-4}$.

A veces sucede que el resultado de una operación es muy pequeño y menor que el mínimo valor representable[2], en este caso se almacenará como $+0$ o -0, dependiendo del signo del resultado. También se observa que al existir un 1 implícito en la mantisa no se puede representar el ± 0 como un número normal, por lo que éste es considerado un valor especial.

[2]Este valor depende de la precisión que se esté utilizando

Por otro lado, ante una operación que arroje un resultado excesivamente grande (en valor absoluto), este se almacenará como $\pm\infty$[3].

De las situaciones recién mencionadas surge la necesidad de una representación para los valores especiales.

Es importante detenerse en la representación del exponente, que como se ha visto, utiliza el sistema CD con frontera no equilibrada $f = \frac{b^d}{2} - 1$, lo que permite almacenar exponentes comprendidos en el rango [-127,128] en el sistema de precisión simple ([-1023,1024] en doble precisión). Pues, puede verse en la tabla 1 que el rango comprendido entre e_{min}y e_{max}no cubre todo el rango disponible. Esto se debe a que en este sistema de representación, se reservan $e_{min} - 1$ y $e_{max} + 1$ para representar los valores especiales. Nótese que esta elección no es arbitraria: la representación de $e_{min} - 1$ en ambas precisiones, está compuesta exclusivamente por ceros, y la representación de $e_{max} + 1$ en ambas precisiones, está compuesta exclusivamente por unos; por lo tanto, ambos son valores fácilmente reconocibles.

Adicionalmente pueden representarse números subnormales, es decir números no normalizados, de la forma $r = \pm 0, m \times 2^{e_{min}}$, que se extienden en el rango comprendido entre el mayor número normal negativo y el menor número normal positivo. Notar que los números subnormales no tienen bit implícito.

La tabla 2 indica cómo se representan en IEEE 754 los valores especiales: ± 0, $\pm\infty$[4] y NaN.

Exponente	Mantisa	Valor
$e_{min} - 1$	0	± 0
$e_{min} - 1$	$\neq 0$	Números subnormales: $r = \pm 0, m \times 2^{e_{min}}$
$e_{max} + 1$	0	$\pm\infty$
$e_{max} + 1$	$\neq 0$	NaN

Cuadro 2: Representación de valores especiales en IEEE 754.

Como ya se vio, cuando el exponente se encuentra en el rango [e_{min}, e_{max}], lo representado es un número normal.

[3] El lector recordará que ∞ no es un número, sino un valor límite.

[4] El signo corresponde al signo de la mantisa.