Estándar IEEE 754

Organización de computadoras 2014

Universidad Nacional de Quilmes

El estándar IEEE 754 define representaciones para números de coma flotante con diferentes tipos de presición: simple y doble, utilizando anchos de palabra de 32 y 34 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 de 32 bits, el bit mas significativo es utilizado para almacenar el signo de la mantisa, los siguientes 8 bits guardan la representación del exponente, y los restantes 23 bits almacenan la mantisa. El exponente se representa en exceso de 8 bits, con un desplazamiento de 127.

S	Exp:Exc(8,127)	Mant: SM(24,23) Norm c/bi

De manera similar, en la representaciín IEEE de doble precisión, el bit mas significativo es utilizado para almacenar el signo de la mantisa, los siguientes 11 bits representan el exponente y los restantes 52 bits representan la mantisa. El exponente se representa en exceso de 11 bits, con un desplazamiento de 1023.

S	Exp:Exc(11,1023)	Mant: SM(53,52) Norm c/bi

En ambos casos se tiene una mantisa normalizada con un bit entero y los restantes fraccionarios, es decir que tiene la forma "1,X", donde X es el valor de los bits fraccionarios. Además, como se tiene un bit implícito, el dígito 1 (entero) está oculto y por lo tanto no es almacenado en la representación, permitiendo así ganar precisión.

Sin embargo, los parámetros usados en las representaciones de simple y doble precisión son los que se describen en la siguiente tabla:

	P. simple	P. doble
Cant. total de bits	32	64
Cant. de bits de la mantisa (*)	24	53
Cant. de bits del exponente	8	11
Mínimo exponente (emin)	-126	-1022
Máximo exponente (emax)	127	1023

(* incluyendo el bit implícito)

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 $\frac{\infty}{\infty}$ ó $\sqrt{-4}$.

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

Por otro lado, ante una operación que arroje un resultado excesivamente grande (en valor absoluto), este se almacenará como $+\infty$ ó $-\infty$.

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

El exponente lo dice todo

Es importante detenerse en la representación del exponente, que como se ha visto, utiliza el sistema Exceso con frontera no equilibrada (127 o 1023), lo que permite almacenar exponentes comprendidos en el rango [-127,128] en el sistema de precisión simple o [-1023,1024] en el sistema de precisión doble. Pues, puede verse en la tabla de la sección anterior que el rango entre emin y emax no cubre todo el rango disponible, y esto se debe a que se reservan las representaciones de emin-1 y emax+1 en ambas precisiones para representar valores especiales. Nótese que esta elección no es arbitraria: la cadena que representa emin-1 está compuesta de ceros y la cadena que representa el valor emax+1 está compuesta por unos, ambos fácilmente reconocibles.

Adicionalmente pueden representarse valores subnormales o denormalizados, es decir números **no normalizados**, de la forma $\pm 0, X * 2^{emin}$, que se extienden en el rango comprendido entre el mayor número normal negativo y el menor número normal positivo. Notar que estos números no tienen bit implícito (ó es cero).

La siguiente tabla indica cómo se clasifican los valores especiales.

Exponente	Mantisa	Tipo de número
emin - 1	0	± 0
emin - 1	$\neq 0$	Denormalizados: $\pm 0, X * 2^{emin}$
emax + 1	0	$\pm \infty$
emax + 1	$\neq 0$	NaN
[emin,emax]	cualquiera	Normalizados: $\pm 1, X * 2^{e}$

Ejercicio

Interpretar las siguientes cadenas del formato de precisión simple:

1. 1100 0010 0110 1011 1000 0000 0000 0000
2. 0100 0010 0110 1011 1000 0000 0000 0000
3. 1000 0010 0110 1011 1000 0000 0000 0000