UNIVERSIDAD DE SAN BUENAVENTURA CALI

Conversión de un número Decimal

A un número en Formato IEE 754

El título completo del estándar es IEEE Standard for Binary Floating-Point Arithmetic (ANSI/IEEE Std 754-1985)

Precisión simple 32-bits:

Un número en coma flotante de precisión simple se almacena en una palabra de 32 bits (1+8+23).

```
 1    8                23              <-- tamaño en bits
+-+--------+-----------------------+
|S|  Exp   |  Mantisa              |
+-+--------+-----------------------+
31 30    23 22                    0    <-- índice del bit
                                          (0 a la derecha)
   Desplazado +127
```

Donde **S** es el bit de signo y **Exp** es el campo exponente.

(Para el signo: 0=Positivo; 1= Negativo).

El exponente es desplazado en el un número en precisión simple, un exponente en el rango −126 a +127 es desplazado mediante la suma de 127 para obtener un

valor en el rango 1 a 254 (0 y 255 tienen valores especiales descritos más adelante).

Cuando se interpreta el valor en coma flotante, el número es desplazado de nuevo para obtener el exponente real.

El conjunto de valores posibles pueden ser divididos en los siguientes:

- ceros
- números normalizados
- números desnormalizados
- infinitos
- NaN (no es un número, como por ejemplo, la raíz cuadrada de un número negativo)

Las clases se distinguen principalmente por el valor del campo Exp.

Considera Exp y Fracción como campos de números binarios sin signo (Exp se encuentra en el rango 0–255):

Clase	Exp	Fracción
Ceros	0	0
Números desnormalizados	0	distinto de 0
Números normalizados	1-254	cualquiera
Infinitos	255	0
NaN (Not a Number)	255	distinto de 0

Para números normalizados, los más comunes, Exp es el exponente desplazado y Fracción es la parte fraccional del significando; El número tiene valor V:

$$V = s \times 2^e \times m$$

Donde

S = +1 (números positivos) cuando S es 0

S = −1 (números negativos) cuando S es 1

E = Exp + 127 (en otras palabras, al exponente se le suma 127 y se almacena, a esto también se le llama "biased with 127" en inglés)

M = 1, Fracción en binario (esto es, el significando es el número binario 1 seguido por la coma decimal seguido por los bits de Fracción). Por lo tanto, 1 ≤ M < 2.

Ejemplo:

Codifiquemos el número decimal -118,625 usando el sistema IEEE coma flotante.

Necesitamos obtener el signo, el exponente y la fracción.

Dado que es un número negativo, el bit de signo es "1".

Primero, escribimos el número (sin signo, es decir 118,625) usando notación binaria. Consulta el sistema de numeración binario para ver cómo hacer esto.

El resultado es 1110110,101.

Ahora, movamos la coma decimal a la izquierda, dejando sólo un 1 a su izquierda.

$1110110{,}101 = 1{,}110110101 \cdot 2^6$ **Esto es un número en coma flotante normalizado**.

El significante es la parte a la derecha de la coma decimal, rellenada con ceros a la derecha hasta que obtengamos todos los 23 bits.

Es decir 11011010100000000000000.

El exponente es 6, pero necesitamos convertirlo a binario y desplazarlo (de forma que el exponente más negativo es 0, y todos los exponentes son solamente números binarios no negativos).

Para el formato IEEE coma flotante, el desplazamiento es 127, así es que:

6 + 127 = 133, En binario, esto se escribe como 10000101.

De una manera más gráfica:

```
 1     8               23             <-- Tamaño en bits
+-+--------+-----------------------+
|S|  Exp   |     Significante      |
|1|10000101|11011010100000000000000|
+-+--------+-----------------------+
31 30    23 22                    0   <-- índice del bit
                                         (0 a la derecha)
  Desplazado +127
```

UNIVERSIDAD DE SAN BUENAVENTURA CALI