Representación de Números Punto Flotante – Estándar IEEE-754

Los números de punto flotante binarios son almacenados en una magnitud con signo de la siguiente forma:

Donde el bit más significante es el bit de signo, el exponente son los siguientes bits, y la mantissa son los bits menos significantes.
El bit más significante de la mantissa es determinado por el valor del exponente:

- Si 0 < exponente < $2^e - 1$, el bit más significante de la mantissa es 1 y el número está normalizado.
- Si exponente = 0, el bit más significante de la mantissa es 0 y el número está desnormalizado.

Tres especiales casos son:

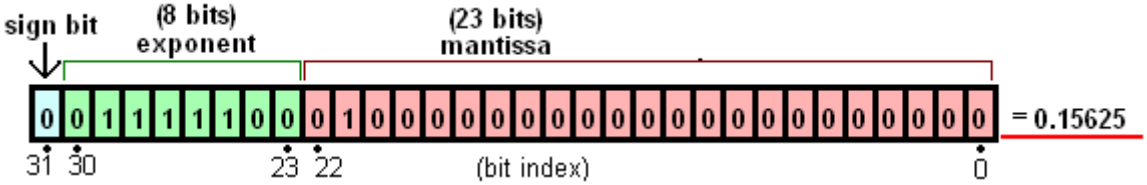
1. Si el exponente es 0 y la mantissa es 0, el número es ±0 (depende del bit del signo).
2. Si el exponente = $2^e - 1$ y la mantissa es 0, el número es ±infinito (depende del bit del signo).
3. Si el exponente = $2^e - 1$ y la mantissa es distinto de 0 (puros 1), el número representado no es un número.

Las clases se distinguen principalmente por el valor del campo Exponente, siendo modificada ésta por el campo Mantissa. Considera Exponente y Mantissa como campos de números binarios sin signo (el exponente se encuentra en el rango 0–255):

Tipos	Exponente	Mantissa
Ceros	0	0
Números Desnormalizados	0	distinto de 0
Números Normalizados	1 a ($2^e - 2$)	cualquiera
Infinitos	$2^e - 1$	0
NaNs	$2^e - 1$	distinto de 0

Precisión Simple 32-bits

Un número en punto flotante de precisión simple se almacena en una palabra de 32 bits.



Donde S es el bit de signo y Exponente es el campo del exponente. (Para el signo: 0=Positivo y 1= Negativo).
El exponente es desplazado en el sentido de la IEEE para una palabra. El valor almacenado es el *offset* (desplazado 127 en este caso) del valor actual. El desplazamiento ocurre porque los exponentes pueden ser valores con signo, para permitir la representación de valores pequeños y grandes, pero la representación en complemento a dos haría esta tarea más difícil. Para resolver esto, el exponente es desplazado antes de ser almacenado, ajustando su valor para ponerlo dentro de un rango

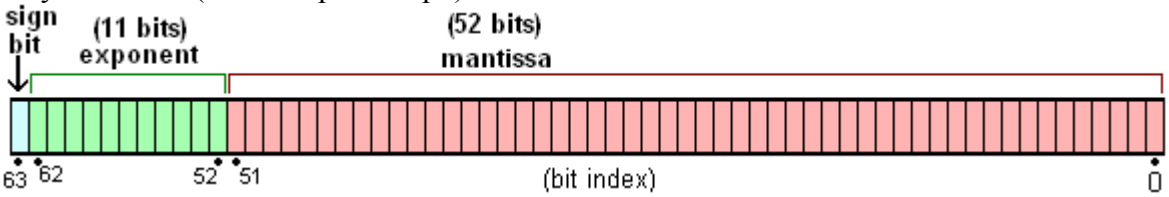
sin signo adaptable a una comparación. Así, para un número en precisión simple, un exponente en el rango −126 a +127 es desplazado mediante la suma de 127 para obtener un valor en el rango 1 a 254 (0 y 255 tienen valores especiales descritos más adelante). Cuando se interpreta el valor en punto flotante, el número es desplazado de nuevo para obtener el exponente real.
Para números normalizados, los más comunes, Exponente es el exponente desplazado y Mantissa es la parte fraccional del significando. El número tiene valor v: $\mathbf{v = s \times 2^e \times m.}$
Donde:
s = +1 (números positivos) cuando S es 0.
s = −1 (números negativos) cuando S es 1.
e = Exponente − 127 (en otras palabras, el exponente se almacena con 127 sumado a él, también llamado *"biased with 127"* en inglés).
m = 1.Mantissa en binario (esto es, el significando es el número binario 1 seguido por el punto decimal seguido por los bits de Fracción). Por lo tanto, $1 \leq m < 2$.
Nota:
Los **números desnormalizados** son iguales excepto que **e = −126** y **m es 0.Fracción**. (e NO es -127: El significando ha de ser desplazado a la derecha por un bit más, de forma que incluya el bit principal, que no siempre es 1 en este caso. Esto se balancea incrementando el exponente a -126 para el cálculo.).
Nota:

1. −126 es el menor exponente para un número desnormalizado.
2. Hay dos ceros: +0 (S es 0) y −0 (S es 1).
3. Hay dos infinitos: +∞ (S es 0) y −∞ (S es 1).
4. Los NaN s pueden tener un signo y un significando, pero estos no tienen otro significado que el que puedan aportar en pruebas de diagnostico; el primer bit del significando es a menudo utilizado para distinguir *NaN s señalizados* de *NaN s silenciosos*
5. los NaN s y los infinitos tienen todos los bits a 1 en el campo Exp.

Tipos	Exponente	Mantissa
Ceros	0	0
Números Desnormalizados	0	distinto de 0
Números Normalizados	1 – 254	cualquiera
Infinitos	255	0
NaNs	255	distinto de 0

Precisión doble 64-bits

La precisión doble es esencialmente lo mismo, exceptuando que los campos son de mayor tamaño (más bits por campo):



Los NaN s y los infinitos son representados con todos los bits de los Exponente siendo 1 (2047 en decimal).

Para los números normalizados, el exponente es desplazado +1023 (así e = Exponente - 1023) Para números desnormalizados es exponente es -1022 (el mínimo exponente para un número desnormalizado). Como antes, ambos infinitos y los ceros contienen signo.

Tipos	Exponente	Mantissa
Ceros	0	0
Números Desnormalizados	0	distinto de 0
Números Normalizados	1 – 2046	cualquiera
Infinitos	2047	0
NaNs	2047	distinto de 0