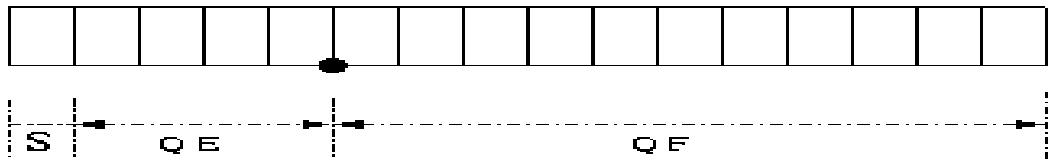
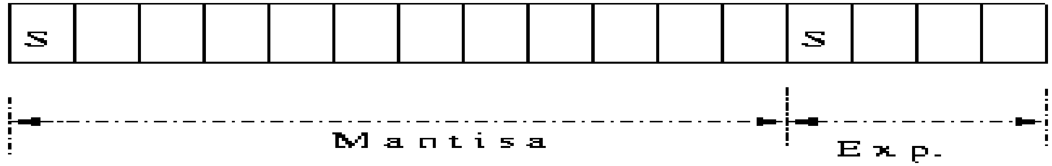
FORMATOS NUMÉRICOS

En cualquier sistema digital o una computadora, los números son representados como una combinación finita de números binarios con valores cero y uno, que ocasiona **errores** por los efectos de precisión finita.
Los formatos más frecuentes para representar los números en una computadora son: **Punto fijo y Punto flotante.**

FORMATO DE PUNTO FIJO



FORMATO DE PUNTO FLOTANTE



ERRORES NUMÉRICOS

Fuentes de errores en operaciones aritméticas en un sistema de PDS

- Efecto de conversión de una señal analógica a digital.

- La representación de los coeficientes.

- Truncamiento o redondeo de los resultados cuando se almacenan.

FORMATOS DE PUNTO FIJO

Signo magnitud (SM): el bit más significativo (MSb) es utilizado para representar el signo del número y el resto de bits representa la magnitud.

Complemento a uno (C1): los números positivos se representan de forma similar al formato SM y los negativos en complemento a uno.

Complemento a dos (C2): los números positivos se representan de forma similar al formato SM y los negativos en complemento a dos

FORMATOS DE PUNTO FIJO
L = 4 bits

Valor numérico	Signo magnitud (SM)	Complemento a uno (C1)	Complemento a dos (C2)
+7	0111	0111	0111
+6	0110	0110	0110
+5	0101	0101	0101
+4	0100	0100	0100
+3	0011	0011	0011
+2	0010	0010	0010
+1	0001	0001	0001
+0	0000	0000	0000
-0	1000	1111	—
-1	1001	1110	1111
-2	1010	1101	1110
-3	1011	1100	1101
-4	1100	1011	1100
-5	1101	1010	1011
-6	1100	1001	1010
-7	1111	1000	1001
-8	—	—	1000

FORMATO FRACCIONARIOS DE PUNTO FIJO Qi

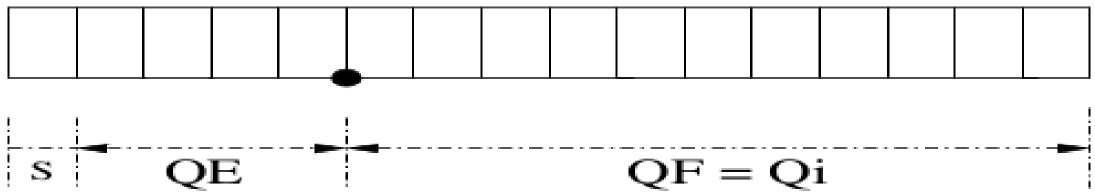
Una palabra digital L bits se particiona en:

L* = S + *QE* + *QF

QE : bits para la parte entera

QF = Qi : bits para la parte fraccionaria

S : un bit de signo, 0 positivo, 1, negativo



Un número *X* positivo se puede representar en formato de punto entero como:

$$X=\sum_{i=-QE}^{QF} b_i r^{-i}=\left(b_{-QE},\ldots,b_{-1},b_0\ldots,b_{QF}\right) \quad 0\le b_i\le\left(r-1\right)$$

INTERVALOS DINAMICOS
Y PRECISION NUMERICA L = 16 b

$$-2^{QE} < ID < 2^{QE} - 2^{-QF}$$

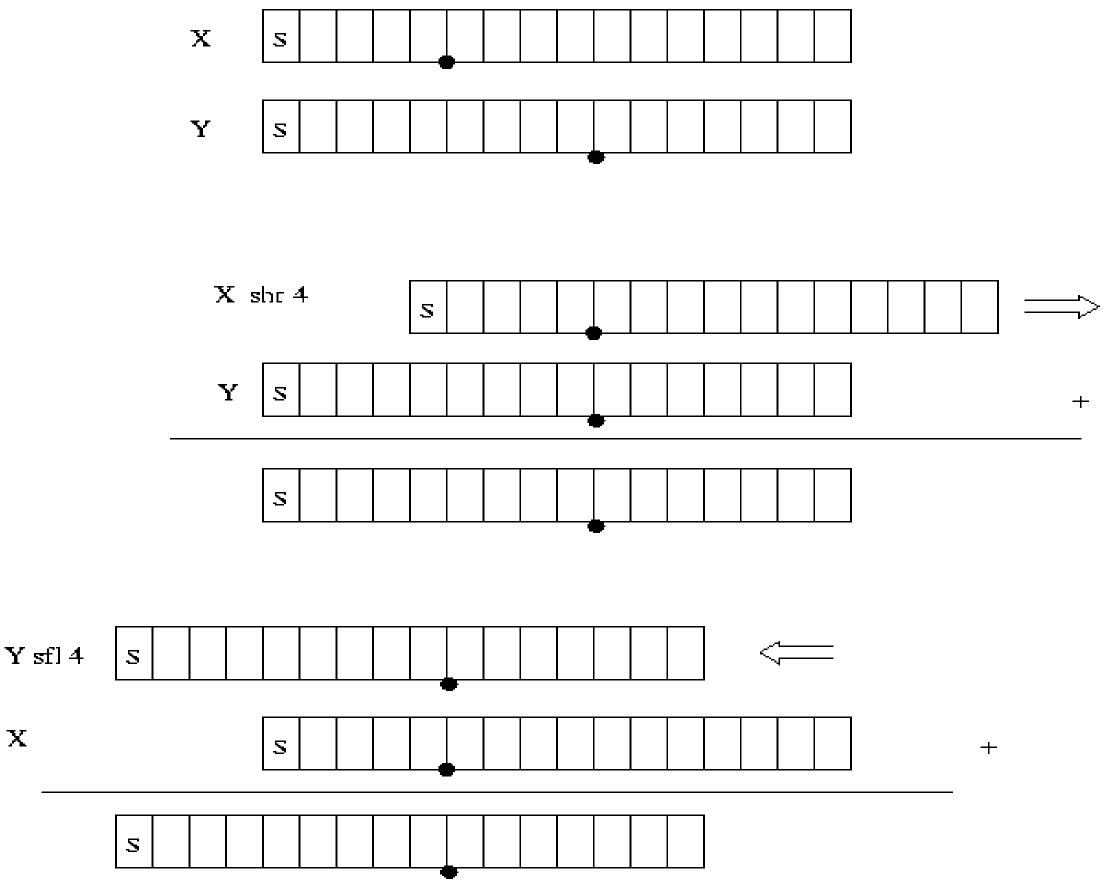
$$p = 2^{-QF}$$

Formato Qi	Mínimo	Máximo	Precisión
Q15	-1	0.9999694	0.0000305175
Q14	-2	1.9999389	0.0000610351
Q12	-8	7.9997558	0.0002441140
Q8	-128	127.96093	0.0039062500
Q4	-2,048	2047.9375	0.0625000000
Q1	-16,384	16,383.5	0.5000000000
Q0	-32,768	32,767	1.0000000000

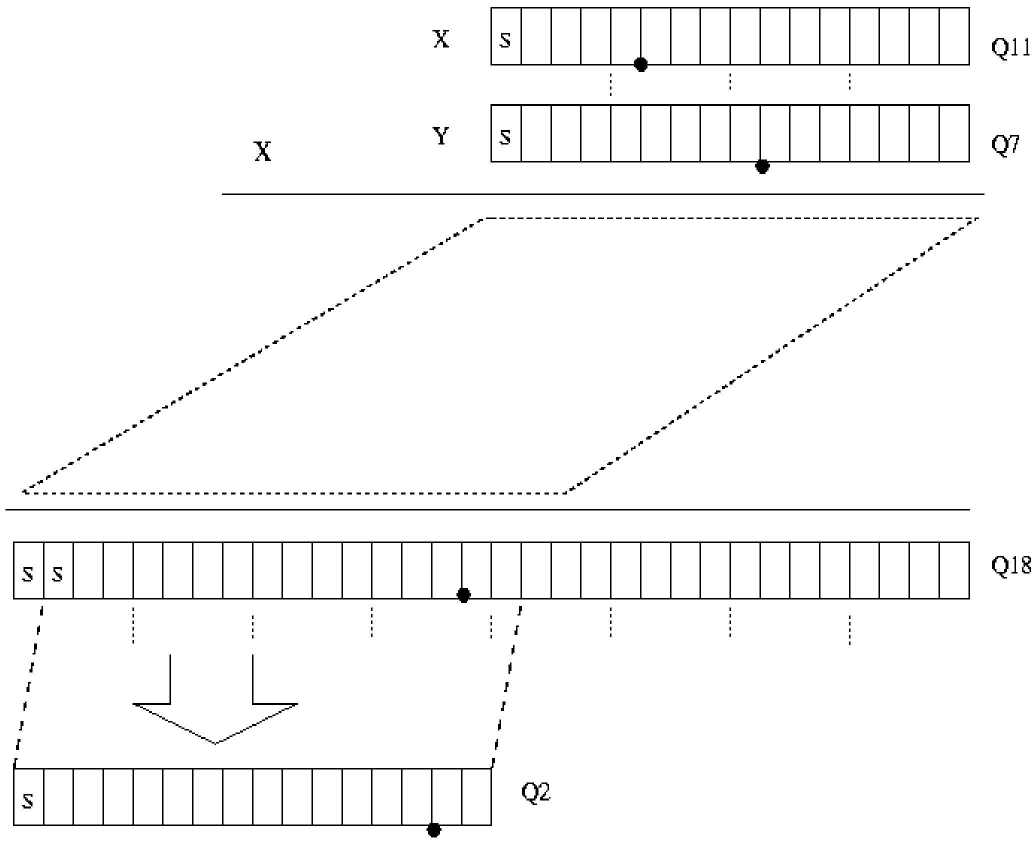
INTERVALOS DINAMICOS Y PRECISION NUMERICA L = 32 b

Formato Qi	Mínimo	Máximo	Precisión
Q31	-1	0.999999999	0.00000000046
Q30	-2	1.999999999	0.000000001
Q28	-8	7.999999996	0.000000004
Q24	-128	127.999999940	0.000000060
Q20	-2048	2047.999999046	0.000000954
Q16	-32768	32767.999984741	0.000015259
Q12	-524288	524287.999755859	0.000244141
Q8	-8388608	8388607.99609375	0.003906250
Q4	-134217728	134217727.937500	0.062500000
Q1	-1073741824	1073741824.50000	0.500000000
Q0	-2147483648	2147483647.00000	1.000000000

SUMA EN PUNTO FIJO



MULTIPLICACION EN PUNTO FIJO



Ejemplo de multiplicación en punto fijo, formato Qi

```
L = 4 bits
A = 2.5  en Q1          010.1              A = 2.5  en Q1           010.1
B = 1.75 en Q2        x 01.11              B =-1.75 en Q2 C2      x 10.01

                   ------------                                ------------
                         0101                                      0101
                        0101                                      0000
                       0101                                      0000
                  +   0000                                  +   1011    (C2 de A)

                  -------------                          ---------------
                      0100011             Ax(-B)=-AxB =-4.375   1011101
     AxB=4.375 =     0100.011             Ax(-B) en C2= 4.375   0100.011
```

Multiplicación por Hardware y Microcódigo

Los DSP ejecutan multiplicaciones y sumas por Hardware en un ciclo de reloj.

Ejemplo: multiplicación a 4-bits (sin signo)

```
   Hardware                 Microcódigo
       1011                      1011
     x 1110                    x 1110
   ---------                 ---------
   10011010                      0000    Ciclo 1
                                1011.    Ciclo 2
                               1011..    Ciclo 3
                              1011...    Ciclo 4
                             ---------
                             10011010    Ciclo 5
```

FORMATO DE PUNTO FLOTANTE

Un número X puede expresarse como
$X = M\,2^E$
donde:
M: mantisa
E: exponente

Ejemplo

```
 X  =  28.79546    En binario      0001 1100 . 1100 1011 1010 (1C.CBAh)
 Normalizando X para que  mantisa 0 < M < 0.99999999
Xn = M  = 000.1110 0110 0101 1101    e = 5 (0101)

Xn  = 0001.1100 1100 1011 1010        =  1.m,  1 < 1.m < 1.999999
         __/  \__________    _______
Bit implícito              \/
                          mantisa
```

OPERACIONES EN PUNTO FLOTANTE

Dados dos números X e Y en punto flotante

$$\text{si } X = (-1)^{sx} X_M 2^{Ex}$$
$$Y = (-1)^{sy} Y_M 2^{Ey} \text{ y si } E_y > E_x$$

Suma

$$W = [(-1)^{sy} Y_M + (-1)^{sx} X_M >> (E_y - E_x)] 2^{Ey}$$

Multiplicación

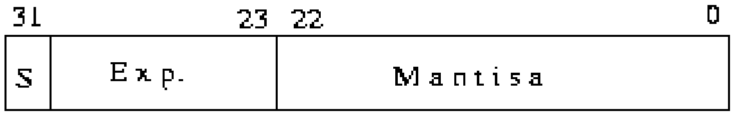
$$W = [(-1)^{sx} X_M \cdot (-1)^{sy} Y_M] 2^{Ex+Ey}$$

División

$$W = \frac{(-1)^{sx} X_M}{(-1)^{sy} Y_M} 2^{Ex-E_y}$$

FORMATO DE PUNTO FLOTANTE IEEE 754

$$X = (-1)^S M 2^{E-127}$$



Precisión simple (32 bits)

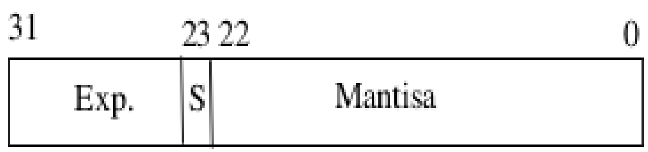
El formato IEEE tiene la siguiente interpretación y casos especiales [10]:

```
Si E=255 y M != 0 , entonces X no es un número       (NAN)
Si E=255 y M = 0 , entonces X = (-1)^s(infinito),   (X = infinito)
Si 0 < E < 255,  entonces X = (-1)^s(1.M)2^(E-127)
Si E=0 y M != 0, entonces X=(-1)^s (0.M)2^(E-126), (X aprox 0)
Si E=0 y M = 0,  entonces X=(-1)^s(0),              (X = 0)
```

FORMATO DE PUNTO FLOTANTE MICROSOFT

$$X = (-1)^S M 2^{E-129}$$

a) Precisión simple (32 bits)



b) Precisión doble (64 bits)