Representação em ponto flutuante

Na representação em ponto fixo, a vírgula não é representada, mas assumida em uma posição.
Nas máquinas atuais, utiliza-se a vírgula na posição mais a direita para representar inteiros.
Números fracionários utilizam a representação em ponto flutuante.
Exemplo: 3,141592565
2,71828
$1,0 \times 10^{-9}$
$3.155.160.000 = 3,15576 \times 10^{9}$
Maior número com 32 bits: 2.147.483.648
Utiliza-se a notação científica:
Um dígito à esquerda do ponto decimal
$1,0 \times 10^{-9,\ 3,15576} \times 10^{9}$
Notação científica normalizada: não possui zero antes da vírgula
$0,1 \times 10^{-8}$, $10,0 \times 10^{-10}$, não normalizadas
$1,0 \times 10^{-9}$, $1,0 \times 10^{-9}$, normalizadas
Números binários em notação científica normalizada
$(1,0)_2 \times 2^{-1}$
Representação de um número na base 2 em ponto flutuante
$(1,mmmmm)_2 \times 2^{yyy}$
Vantagens de se usar esta notação:

- Padrão para trocar dados
- Simplifica aritmética pois números padronizados
- Aumenta precisão dos números que podem ser representados porque não desperdiça dígitos.
  Exemplo: $0,000012 \times 10^{-2} = 1,2 \times 10^{-7}$

O que tem que ser representado ?

```
N = (+/-) 1,mmmmmmm x 2^yyyyyyyyy

    Sinal   Mantissa  Expoente


|--------------------------------------|

|  S | Expoente     |Mantissa           |

|--------------------------------------|

1 bit  7 bits        24 bits
```

Questões:

- Quantos bits utilizar para mantissa e expoente ?
- Qual a representação utilizada para mantissa e expoente? sinal e magnitude, complemento a 2 ?

Exemplo:
Suponha que S=0, positivo, S=1, negativo
Expoente representado em sinal e magnitude
Mantissa em binário
$N = (+/-)1,M \times 2^{E}$
$N=+407,375 = (110010111,011)_2 = 1,M \times 2^{E} = 1,10010111011 \times 2^{+8}$

S = 0

E = +8, utilizando-se sinal e magnitude com 7 bits, teremos a

representação: 0001000

M=10010111011 (11 bits) completa para 24 com 0s.

```
    Sinal   Mantissa  Expoente
|-----------------------------------------------------|
|  0 | 0001000      |10010111011000000000000000000000|
|-----------------------------------------------------|
1 bit  7 bits        24 bits
```

Dado o número em ponto flutuante, qual o decimal a ele associado ?

Exemplo: 04D00000

$N = 1,M \times 2^{E}$

0000 0100 1101 0000 0000 0000 0000 0000

S=0, E=+4, M=1101

$N=+1,1101 \times 2^{4}=11101=+29,0$

Formato IEEE 754 para ponto flutuante

```
|31|30|29|28|27|26|25|24|23|22|21|.....................|0|
|-------------------------------------------------------|
|S |Expoente               | Mantissa                   |
|-------------------------------------------------------|
 1    8                       23
 bit  bits                    bits
```

$N=(-1)^{S}x(1+M)x2^{E}=(-1)^{S} \times (1,M)x2^{E}$

Precisão simples: 8 bits para expoente e 23 para mantissa
Precisão dupla: 11 bits para expoente e 52 bits para mantissa

Decisões do IEEE 754

- Sinal mais à esquerda para facilitar a comparação com zero, porque em complemento a 2 bit mais significativo é zero para números positivos e 1 para números negativos, então pode-se utilizar instruções que comparem inteiros
- Expoente antes da mantissa porque números com maior expoente são maiores e a representação do inteiro é maior.
  Suponha que temos 3 bits para expoente, 6 para mantissa e 1 para sinal. Se usarmos complemento a 2 para expoente:
  $1,0 \times 2^{-1} = 0111000000$
  $1,0 \times 2^{+1} = 0001000000$
  Comparação de números inteiros não ia funcionar.
  Utiliza-se a notação excesso de N (ou notação deslocada ou *biased exponent*).

Representação em excesso

Padrao de bits	SM	C2
000	0	0
001	+1	+1
010	+2	+2
011	+3	+3
100	0	-4
101	-1	-3
110	-2	-2
111	-3	-1

Utilizando as representações SM e C2 não podemos comparar dois números como inteiros sem sinal.
A representação em excesso permite que possamos comparar o padrão de bits referentes a cada inteiro com sinal, considerando que o padrão de bits representa inteiros sem sinal.
$x_{bias}=x + bias$

Exemplo:

Padrao de bits	ISS	Excesso de 2
000	0	-2
001	1	-1
010	2	0
011	3	+1

100	4	+2
101	5	+3
110	6	+4
111	7	+5

Necessita-se definir o número de bits a ser utilizado e o valor de bias.
Escolhendo o bias: Para obter uma distribuição homogênea de valores acima e abaixo de 0, escolhe-se bias igual a 2^{n-1} ou $((2^{n-1})-1)$, onde n é o número de bits disponíveis
Dados 4 bits, o valor de bias será $2^3=8$, de modo que o resultado represente metade de números positivos e metade de negativos.
Exemplo: valor a ser representado= +3, soma com bias +8, resultado 11, cuja representação é 1011
Dada a representação 0110, qual número ele está representando?
0110 = 6 = x + bias = x + 8, x = 6-8=-2

Na notação em escesso, o expoente mais negativo tem a representação 0000...000, e o mais positivo tem a representação 111...11

Com 3 dígitos, excesso ou bias = $2^{3-1}=4$
-4+4=000
-2+4=010
-1+4=011
0+4=100
1+4=101
2+4=110
3+4=111
N=+1,01 x 2^{-4}=0000010000
N=+1,01 x 2^{+1}=0101010000

Representação em ponto flutuante para expoente, temos 8 bits, o excesso pode ser 2^7 ou 2^7-1. Escolheram o segundo que é 127 para precisão simples e 1023 para precisão dupla.