Representação de ponto fixo

- A representação segue a regra

$$S\ I\ F$$

onde:

S - sinal. 0 para positivo e 1 para negativo

I - parte inteira do valor. Representação binária tradicional. Obs.: deve-se estabelecer previamento a quatidade de bits para essa representação.

F - parte fracionária do valor. O expoente da base inicia em -1 e vai sendo decrementado. Idem observação acima.

- Por exemplo 011101,101:

0	1	1	1	0	1	1	0	1
+	1×2^4	1×2^3	1×2^2	0×2^1	1×2^0	1×2^{-1}	0×2^{-2}	1×2^{-3}
+	16	8	4	0	1	0,5	0	0,125

- Portanto o decimal equivalente é: 29,625
- Qual é a representação binária equivalente a 19,5?

Representação de ponto fixo

- A representação segue a regra

$$S\ I\ F$$

onde:

S - sinal. 0 para positivo e 1 para negativo

I - parte inteira do valor. Representação binária tradicional. Obs.: deve-se estabelecer previamento a quatidade de bits para essa representação.

F - parte fracionária do valor. O expoente da base inicia em -1 e vai sendo decrementado. Idem observação acima.

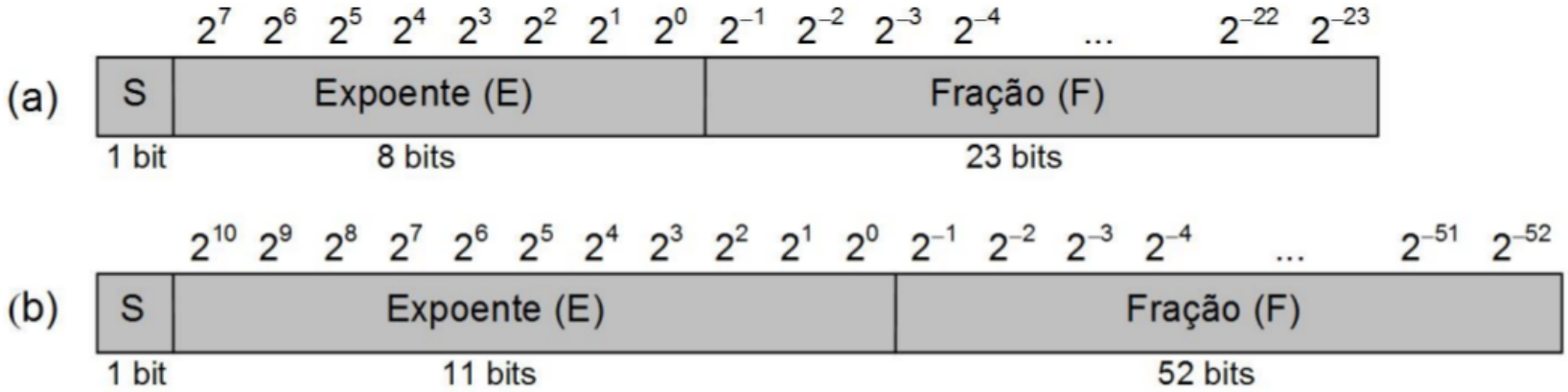
- Por exemplo 011101,101:

0	1	1	1	0	1	1	0	1
+	1×2^4	1×2^3	1×2^2	0×2^1	1×2^0	1×2^{-1}	0×2^{-2}	1×2^{-3}
+	16	8	4	0	1	0,5	0	0,125

- Portanto o decimal equivalente é: 29,625
- Qual é a representação binária equivalente a 19,5?: 10011,1.
- Como representar a vírgula na representação binária?

Representação por ponto flutuante

- Além do ponto fixo, o mais usual para representar números reais é empregar-se o padrão IEEE 754 – código para representar números com ponto flutuante.



a) Ponto flutuante de precisão simples: 1 bit de sinal (S) + 8 bits para expoente (E) + 23 bits para fração (F)
b) Ponto flutuante de precisão dupla: 1 bit de sinal (S) + 11 bits para expoente (E) + 52 bits para fração (F)

- Considera-se que é representado com a utilização de notação científica normalizada: exatamente 1 dígito diferente de zero antes da vírgula.

Precisão simples e dupla

- O número decimal equivalente é obtido por:
  - **precisão simples**: $y = (-1)^S \times (1 + F) \times 2^{E-127}$
  - **precisão dupla**: $y = (-1)^S \times (1 + F) \times 2^{E-1023}$

  E é o expoente deslocado e e é o expoente verdadeiro (pode ser negativo): $e = E - 127$ ($E = e + 127$) ou $e = E - 1023$ ($E = e + 1023$)
- Ex1: Encontre o valor decimal equivalente para o valor abaixo que está em ponto flutuante com precisão simples:

$$10111111100000000000000000000000$$

- Analisando o número acima obtemos: S $= 1$, E $= 01111111 = 127$, F $= 10...0 = 1 \times 2^{-1} = 0{,}5$.
- Portanto

$$y = (-1)^1 \times (1 + 0,5) \times 2^{127-127} = -1,5$$

- Ex2: Encontre o valor decimal equivalente para o valor abaixo que está em ponto flutuante com precisão simples:

$$01000000101000000000000000000000$$

Representação por ponto flutuante - observações

- Na representação por ponto flutuante, a representação de números negativos é equivalente a representação sinal magnitude, ou seja, um número negativo tem exatamente os mesmos bits que sua contraparte positiva, com uma única diferença, que é o bit de sinal.
- As equações não produzem valor 0 (zero), portanto, uma representação especial se faz necessário para representar esse valor: S=0/1, E=0 e F=0 é equivalente a +0 e -0.
- Representação de infinito: S=0/1, E=max e F=0 é equivalente a $+\infty e -\infty$.
- Representação NaN (*not a number*) que são obtidas por números inválidos ou indeterminados (0/0, $\infty - \infty$ etc): S=0/1, E=0 e $F \neq 0$ é equivalente a NaN (erro).

Representação por ponto flutuante - exemplos

- Determinar a representação binária por ponto flutuante de precisão simples para -2,625.

Representação por ponto flutuante - exemplos

- Determinar a representação binária por ponto flutuante de precisão simples para -2,625.
  1. Encontrar o equivalente binário do número, conforme apresentado no slide 10: $-2,625_{10} = -10,101_2$
  2. Representar com notação científica normalizada: $-1,0101 \times 2^1$
  3. Extrair S, F e E: S=1, F=01010000000000000000000, E=128 (10000000), já que, $E = e + 127 ==> E = 1 + 127 ==> E = 128$.
  4. Portanto a sequência binária seria: 11000000001010000000000000000000.
- Determinar a representação binária por ponto flutuante de precisão simples para 0,75.

Representação por ponto flutuante - exemplos

- Determinar a representação binária por ponto flutuante de precisão simples para -2,625.
  1. Encontrar o equivalente binário do número, conforme apresentado no slide 10: $-2,625_{10} = -10,101_2$
  2. Representar com notação científica normalizada: $-1,0101 \times 2^1$
  3. Extrair S, F e E: S=1, F=01010000000000000000000, E=128 (10000000), já que, $E = e + 127 ==> E = 1 + 127 ==> E = 128$.
  4. Portanto a sequência binária seria: 11000000010100000000000000000000.
- Determinar a representação binária por ponto flutuante de precisão simples para 0,75.
  1. $0,75_{10} = 0,11_2 = 1,1 \times 2^{-1}$
  2. Portanto, S = 0, F = 10000000000000000000000 e E = 01111110 ($E = e + 127 = -1 + 127 = 126$)
  3. Portanto a sequência binária seria: 00111111010000000000000000000000.

Ponto flutuante versus inteiro

Ponto flutuante pode representar números muito grandes e números muito pequenos.

- Por exemplo, com os 32 bits pode-se representar:
  - de 2^{-126} até aproximadamente 2^{128} em ponto flutuante de representação simples
  - de 2^{-31} até $2^{+31} - 1$ na representação inteira, portanto, uma faixa numérica bem menor.
- Qual o preço a pagar?

Ponto flutuante versus inteiro

Ponto flutuante pode representar números muito grandes e números muito pequenos.

- Por exemplo, com os 32 bits pode-se representar:
  - de 2^{-126} até aproximadamente 2^{128} em ponto flutuante de representação simples
  - de 2^{-31} até $2^{+31} - 1$ na representação inteira, portanto, uma faixa numérica bem menor.
- Qual o preço a pagar?
- A precisão dos números em ponto flutuante pode ser menor.

Precisão – Exemplo prático

- A representação de 5.023.421.348,75 em IEEE 754 é:
  - Equivalente binário: 10010101101101011010100111010010 0,11
  - Notação científica normalizada: $1{,}00101011011010110101001110010011 \times 2^{32}$
  - $\therefore$ S=0, e=32 ==> E=32+127 ==> E=10011111 e F=00101011011010110101001
  - IEEE 754: $\boxed{01001111100101011011010110101001}$
- Revertendo: 01001111100101011011010110101001
  - S=0, E=10011111=159 e F=$00101011011010110101001_2 = 0,16960632801055908 2_{10}$
  - $\therefore y = (-1)^0 \times (1 + 0,16960632801055908 2) \times 2^{159-127}$
  - Decimal: $\boxed{5.023.420.928}$
- Conclusão: $5.023.421.348,75 \neq 5.023.420.928$