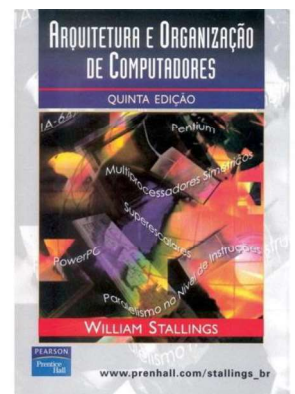


Aula 13 – Aritmética do Computador *Ponto Flutuante*

Prof. João Fernando Mari
joaof.mari@ufv.br

Referências

- STALLINGS, W. **Arquitetura e Organização de Computadores**, 8. Ed., Pearson, 2010.
 - **Capítulo 9**



Roteiro

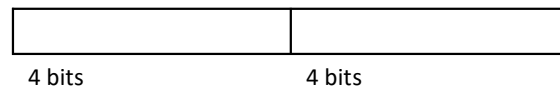
- Números reais
- Exemplo.: Ponto fixo
- Ponto flutuante
- Expoente polarizado
- Normalização
- Exemplo: Ponto flutuante
- Números representáveis
- IEEE 754
- Aritmética de ponto flutuante (+/-)
- Exemplo: Soma e subtração
- Aritmética de ponto flutuante (\times/\div)

Números reais

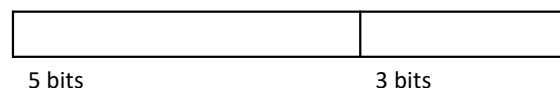
- Como representar os números reais no computador?
 - Poderia ser feito em binário puro.
 - $1001.1010 = 2^4 + 2^0 + 2^{-1} + 2^{-3} = 9,625$
- Onde está o ponto binário?
 - Fixo?
 - Muito limitado em termos de representação.
 - Móvel?
 - Como você determina onde está o ponto?

EXEMPLP: Ponto fixo

- Suponha palavras de 8 bits:
 - 4 bits para a parte inteira e 4 bits para a parte real



- Aumentar o intervalo de inteiros, diminui a precisão da parte real



- Aumentar a precisão da parte real, diminui o intervalo de inteiros



Ponto flutuante

- \pm Significando $\times 2^{\text{Exponente}}$



Obs.: O termo Mantissa, encontrado em muitos livros texto, é considerado obsoleto. Usaremos Significando.

Expoente polarizado

- Expoente está em **representação polarizada**.
 - Polarização: um valor fixo que é subtraído para obter o expoente.
 - Polarização: $2^{k-1} - 1$ (k é o número de bits do expoente)
 - EXEMPLO: Expoente com 8 bits (Intervalo de 0 a 255)
 - A polarização é $2^7 - 1 = 127$
 - Intervalo do expoente polarizado: -127 a +128.



Normalização

- Números em ponto flutuante geralmente são normalizados, ou seja
 - O expoente é ajustado de modo que bit inicial (MSB) da mantissa seja 1.
 - Por ser sempre 1, não é preciso armazená-lo.
- Ex.: $00100101.001 \times 2^0 = 1.00101001 \times 2^5$
- Obs.: Em notação científica, os números são normalizados para um único dígito antes do ponto decimal:
 - Ex.:
 - $3,123 \times 10^3 = 3123,0$
 - $3,123 \times 10^{-3} = 0,003123$

Exemplos de ponto flutuante



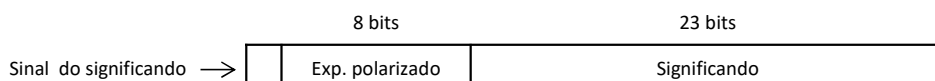
(a) Formato

$$\begin{array}{llll}
 1.1010001 \times 2^{10100} = 0\ 10010011\ 10100010000000000000000 & = 1.6328125 \times 2^{20} \\
 -1.1010001 \times 2^{10100} = 1\ 10010011\ 10100010000000000000000 & = -1.6328125 \times 2^{20} \\
 1.1010001 \times 2^{-10100} = 0\ 01101011\ 10100010000000000000000 & = 1.6328125 \times 2^{-2} \\
 -1.1010001 \times 2^{-10100} = 1\ 01101011\ 10100010000000000000000 & = -1.6328125 \times 2^{-2}
 \end{array}$$

(b) Exemplos

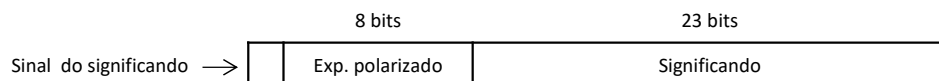
Exemplos de ponto flutuante

- Valor em binário:
 - 0 | 10010011 | 10100010000000000000000
- Resolver o expoente polarizado:
 - $147 - 127 = 20$
- Incluir o bit 1 mais significativo do significando
 - + | 10100 | 1.10100010000000000000000
- Converte os valores para base decimal:
 - + | 20 | 1.6328125
- Colocar em notação científica:
 - (+/-)significando $\times 2^{\text{expoente}}$
 - $+ 1.6328125 \times 2^{20} = 1.6328125 \times 1,048,576 =$
 - 1,712,128
- Valor em decimal:
 - + 1,712,128.0
- Valor em decimal:
 - + 1,712,128.0
- Valor em decimal:
 - + 1,712,128.0
- Converte para binário (a parte inteira e a parte decimal):
 - + 1 1010 0010 0000 0000 0000. 0
- Colocar em notação científica (base 2)
 - + 1 1010 0010 0000 0000 0000. 0 $\times 2^0$
- Mover o ponto até antes do bit 1 mais significativo:
 - + 1. 1010 0010 0000 0000 0000 $\times 2^{20}$
- Tornar o expoente polarizado e remover o 1 mais significativo:
 - $20 + 127 = 147$
 - + | 10010011 | 10100010000000000000000
- Valor em binário:
 - 0 | 10010011 | 10100010000000000000000



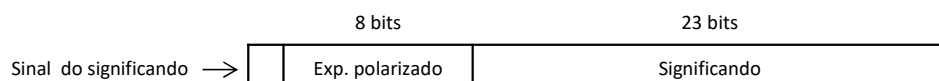
Exemplos de ponto flutuante (significando negativo)

- Valor em binário:
 - 1 | 10010011 | 101000100000000000000000
- Resolver o expoente polarizado:
 - $147 - 127 = 20$
- Incluir o bit 1 mais significativo do significando
 - - | 10100 | 1.101000100000000000000000
- Converte os valores para base decimal:
 - - | 20 | 1.6328125
- Colocar em notação científica:
 - (+/-)significando $\times 2^{\text{expoente}}$
 - $-1.6328125 \times 2^{20} = 1.6328125 \times 1,048,576 =$
 - $-1,712,128$
- Valor em decimal:
 - $-1,712,128.0$
- Valor em decimal:
 - $-1,712,128.0$
- Converte para binário (a parte inteira e a parte decimal):
 - $-1 \ 1010 \ 0010 \ 0000 \ 0000 \ 0000.0$
- Colocar em notação científica (base 2)
 - $-1 \ 1010 \ 0010 \ 0000 \ 0000 \ 0000.0 \times 2^0$
- Mover o ponto até antes do bit 1 mais significativo:
 - $-1. \ 1010 \ 0010 \ 0000 \ 0000 \ 0000 \times 2^{20}$
- Tornar o expoente polarizado e remover o 1 mais significativo:
 - $20 + 127 = 147$
 - - | 10010011 | 101000100000000000000000
- Valor em binário:
 - 1 | 10010011 | 101000100000000000000000



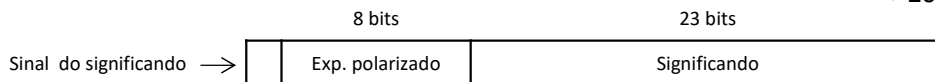
Exemplos de ponto flutuante (expoente negativo)

- Valor em binário:
 - 0 | 01101011 | 101000100000000000000000
- Resolver o expoente polarizado:
 - $107 - 127 = -20$
- Incluir o bit 1 mais significativo do significando
 - + | -10100 | 1.101000100000000000000000
- Converte os valores para base decimal:
 - + | -20 | 1.6328125
- Colocar em notação científica:
 - (+/-)significando $\times 2^{\text{expoente}}$
 - $+1.6328125 \times 2^{-20} =$
- Valor em decimal (um número muito pequeno):
 - $+0.000001557171344757080078125$



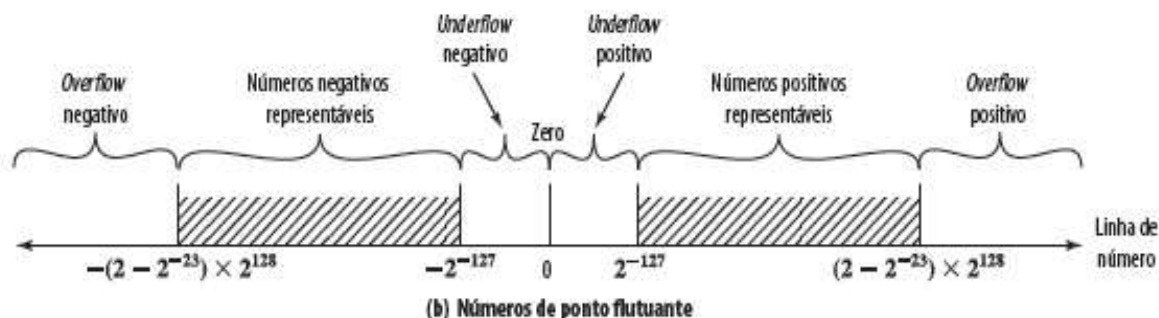
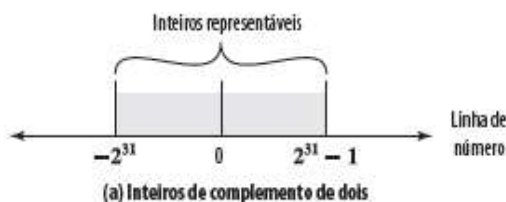
Exemplos de ponto flutuante (com valores reais)

- Valor em decimal:
 - + 20.3
- Converte para binário (a parte inteira e a parte decimal):
 - + 1 0100 . 0100 1100 1100 1100 1100 1100 (...)
- Colocar em notação científica (base 2)
 - + 1 0100 . 0100 1100 1100 1100 1100 1100 $\times 2^0$
- Mover o ponto até antes do bit 1 mais significativo:
 - + 1 . 0100 0100 1100 1100 1100 1100 1100 $\times 2^4$
- Tornar o expoente polarizado e remover o 1 mais significativo do significando (manter apenas 23 bits):
 - $4 + 127 = 131$
 - + | 1000 0011 | 0100 0100 1100 1100 1100 110
- Valor em binário:
 - 0 | 1000 0011 | 0100 0100 1100 1100 1100 110
- Valor em binário:
 - 0 | 1000 0011 | 0100 0100 1100 1100 1100 110
- Resolver o expoente polarizado:
 - $131 - 127 = 4$
- Incluir o bit 1 mais significativo do significando
 - + | 100 | 1. 0100 0100 1100 1100 1100 110
- Converte os valores para base decimal:
 - + | 4 | 1.268749952316284
- Colocar em notação científica:
 - (+/-)significando $\times 2^{\text{expoente}}$
 - + 1.268749952316284 $\times 2^4$
 - + 1.268749952316284 $\times 16 =$
 - + 20.29999924
- Valor em decimal (arredondando):
 - + 20.3



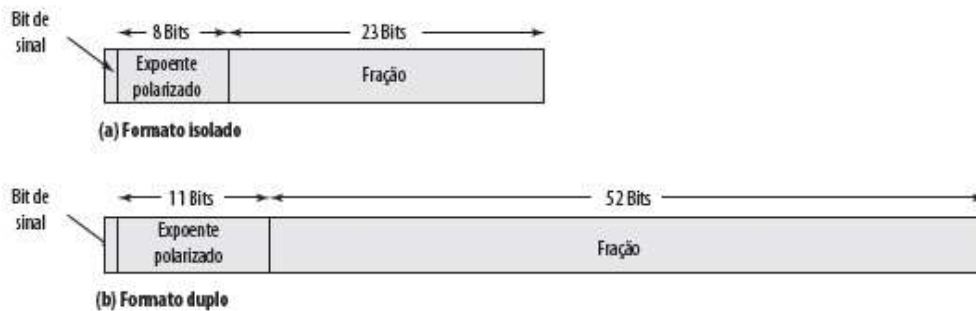
Números representáveis

- Intervalos de números expressos em 32 bits



IEEE 754

- Padrão para armazenamento de ponto flutuante.
- Padrões de 32 e 64 bits.
- Expoente de 8 e 11 bits, respectivamente.

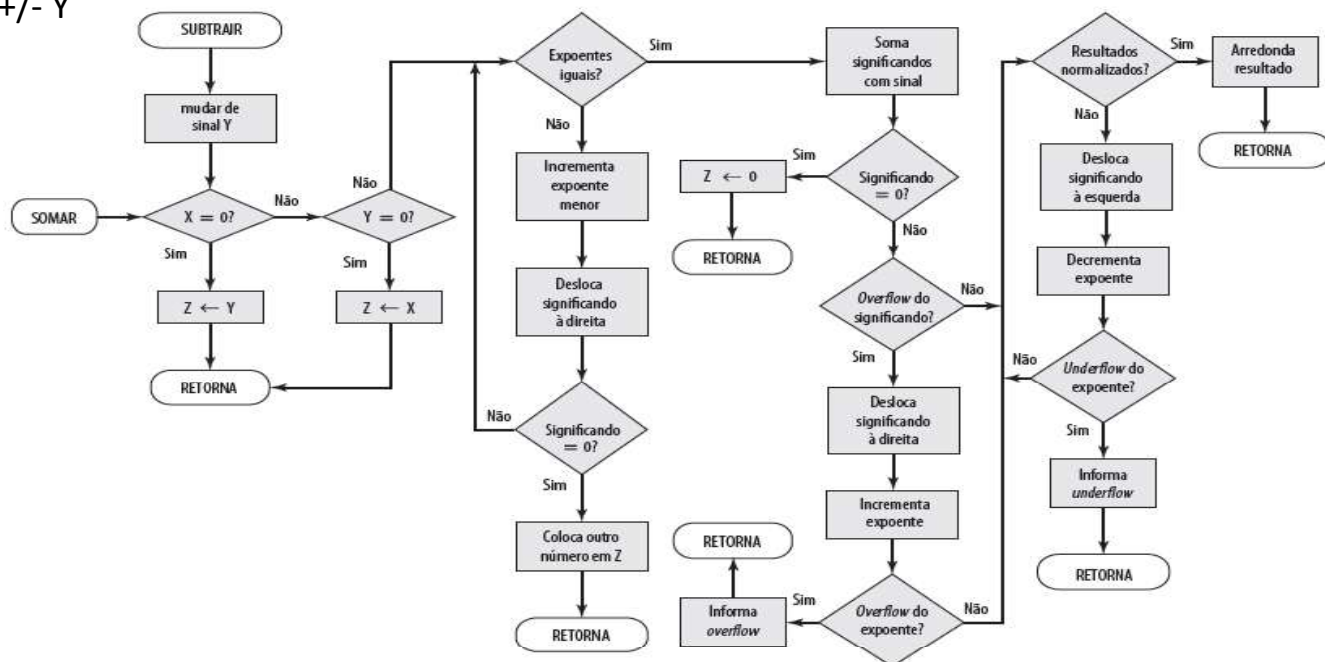


Aritmética de ponto flutuante (+/-)

- Algoritmo para soma e subtração de números binários em ponto-flutuante:
 - Verifique zero
 - Alinhe significandos (ajustando expoentes)
 - Some ou subtraia significandos
 - Verifique *overflow* ou *underflow*
 - Normalize o resultado
 - Arredonde o resultado

Fluxograma da adição e subtração de ponto flutuante

- $Z = X + / - Y$



Exemplo

IEEE FP16	5 bits		10 bits
Sinal do significando	Exp. polarizado	Significando	

20.3

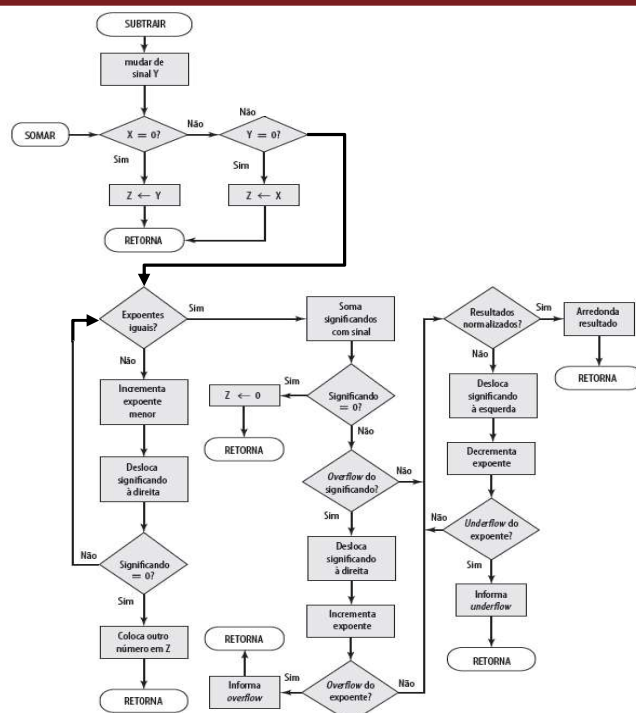
5.25

$$10100.01001100110011 \dots \times 2^0$$
$$101.01 \times 2^0$$
$$1.01000100110011 \times 2^4$$
$$1.0101 \times 2^2$$
$$4 + 15 = 19$$
$$2 + 15 = 17$$

0 | 10011 | 0100010011

0 | 10001 | 0101000000

Polarização: $k=5$, então $2^4 - 1 = 15$



Exemplo

IEEE FP16	5 bits	10 bits
Sinal do significando	Exp. polarizado	Significando

20.3 + 5.25

20.3

0 | 10011 | 1.0100010011
(19)

1.0100010011

0.0101010000 +

1.1001100011

0 | 10011 | 1001100011

Polarização: 19 - 15 = 4

1.5966796875 × 2⁴ =

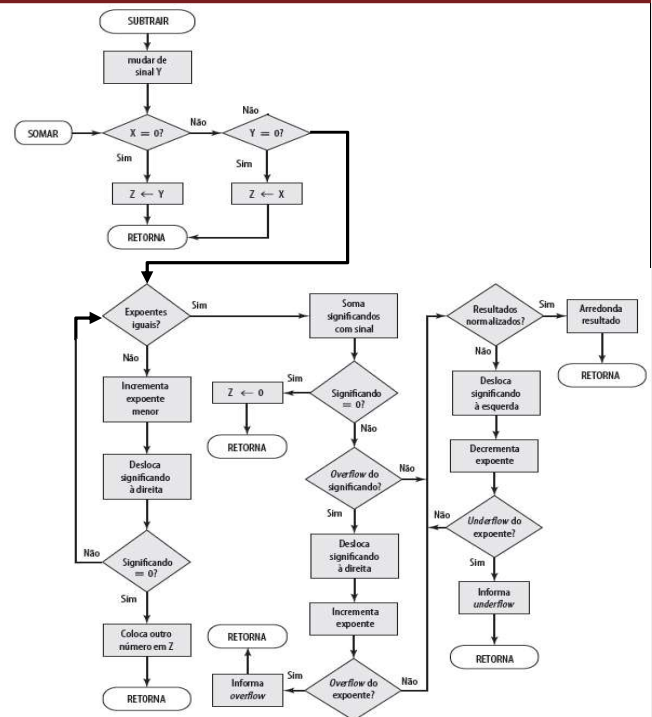
25.546875 = 25.55 (arredondando)

5.25

0 | 10001 | 1.0101000000
(17) + 1 = 18

0 | 10010 | 0.1010100000 | ⚠
(18) + 1 = 19

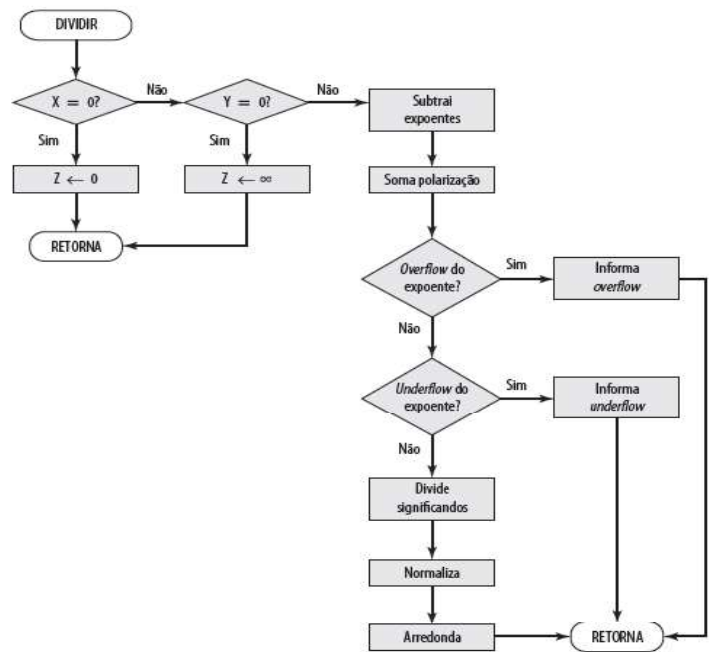
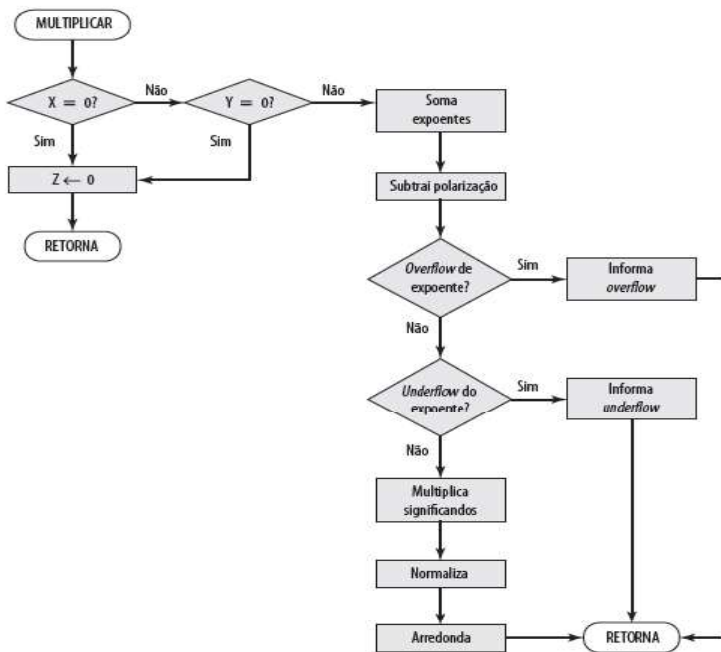
0 | 10011 | 0.0101010000 | ⚠
(19)



Aritmética de ponto flutuante (x/÷)

- Verifique zero.
- Some/subtraia expoentes.
- Multiplique/divida significandos (observe o sinal).
- Normalize.
- Arredonde.
- Resultados intermediários armazenados em tamanho duplo.

Multiplicação e divisão de ponto flutuante



FIM – Aula 13