Erros de representação

Na construção de um equipamento computacional, uma questão importante a ser considerada em sua arquitetura é a forma que será adotada para representar os dados numéricos.

Basicamente, na memória de um equipamento, cada número é armazenado em uma posição que consiste de um sinal que identifica se o número é positivo ou negativo e um número fixo e limitado de dígitos significativos.

Erros de representação

Sistema de ponto flutuante

Um número no sistema de ponto flutuante é caracterizado por uma base **b**, um número de dígitos significativos **n** e um expoente **exp**.

Dizemos que um número real **nr** está representado no sistema de ponto flutuante se for possível escrevê-lo da seguinte maneira:

$nr = m \times b^{exp}$

Onde, **m** é a mantissa do número,

Neste sistema de ponto flutuante, as seguintes condições devem ser verificadas:

$$m = \pm 0.d_1\, d_2 \cdots d_n \quad n \in N$$

Sendo n o número máximo do sistema de dígitos na mantissa d_1, d_2 , …, d_n, dígitos significativos da mantissa, do sistema de representação, com o primeiro digito satisfazendo a condição $1 \le d_1 \le (b-1)$ e os demais digitos da seguinte maneira:

$$0 \le d_i \le (b-1); i = 2,3,...,n.$$

Erros de representação

O expoente exp varia da seguinte maneira:

$$\exp_{min} \leq \exp \leq \exp_{max}$$

Sendo, $\exp_{min} \leq 0$ e $\exp_{max} \geq 1$ *com* $\exp_{min}$ *e* $\exp_{max}$ int*eiros*.

A união de todos os números em ponto flutuante, juntamente com a representação do zero, constitui o sistema de ponto flutuante normalizado, que indicamos por SPF (b,n, exp_{min},exp_{max}).

No sistema o zero é representado da seguinte maneira:

Zero: $(0.\underbrace{0000\ldots0}_{\text{n vezes}})b^{exp(min)}$

Considerando o SPF dado na fórmula genérica (b, n, exp_{min}, exp_{max}), temos:

a) O menor positivo exatamente representável, não nulo, é o real formado pela menor mantissa multiplicada pela base elevada ao menor expoente, isto é:

menor: $(0.1\underbrace{000\ldots0}_{\text{(n-1) vezes}})\ b^{exp(min)}$

Erros de representação

b) O maior positivo exatamente representável é o real formado pela maior mantissa multiplicada pela base elevada ao maior expoente, isto é:

maior: $(0 \cdot \underbrace{[b-1]\ [b-1] \ldots [b-1]}_{\text{n vezes}})\ b^{exp(max)}$

c) O número máximo de mantissas positivas possíveis é dado por:

$$Mantissas_{+} = (b-1)\ b^{n-1}$$

d) O número máximo de expoentes possíveis é dado por:

$$exp_{possíveis} = exp_{max} - exp_{min} + 1$$

e) O número de elementos positivos representáveis é dado pelo produto entre o número máximo de mantissas pelo máximo de expoentes, isto é:

$$NR_{+} = Mantissas_{+} \times exp_{possíveis}$$

Se considerarmos que dado um número real nr ϵ SPF temos que – nr ϵ SPF e a representação do zero, podemos concluir que o número total de elementos exatamente representáveis NR_t é dado por:

$$NR_t = 2 \times NR_{+} + 1$$

Exemplo

Considere o SPF (b, n, expmin, expmax)= SPF (3,2,-1,2) , isto é, de base 3, 2 dígitos de mantissa, menor expoente igual a -1 e maior expoente 2. Para este sistema temos:

a) O menor exatamente representável:

$$0.10\times 3^{-1} = \left(1\times 3^{-1} + 0\times 3^{-2}\right)\times 3^{-1} = \frac{1}{9}$$

b) O maior exatamente representável:

$$0.22 \times 3^{2} = \left(2 \times 3^{-1} + 2 \times 3^{-2}\right) \times 3^{2} = 8$$

c) A quantidade de reais positivos exatamente representáveis:

Temos que a quantidade de reais positivos exatamente representáveis é dada pelo produto entre todas as mantissas possíveis de 2 dígitos , formadas com os dígitos da base 3, isto é, 0.10, 0.11, 0.12, 0.20, 0.21, 0.22, e todas as possibilidades de expoentes, que no caso são -1, 0, 1, 2.

Desta forma, os 24 positivos exatamente representáveis estão listados a seguir

Exemplo

$$\exp = -1: \quad 0.10 \times 3^{-1} = \frac{1}{9}$$

$$\exp = 0: \quad 0.10 \times 3^{0} = \frac{1}{3}$$

$$\exp = 1: \quad 0.10 \times 3^{1} = 1$$

$$\exp = 2: \quad 0.10 \times 3^{2} = 3$$

Resolver para 0.11, 0.2, 0.21 e 0.22 ?

OBS: O menor real positivo representável é 1/9
o maior positivo representável é o real 8.

$$\exp = -1: \quad 0.12 \times 3^{-1} = 5/27$$

$$\exp = 0: \quad 0.12 \times 3^{0} = 5/9$$

$$\exp = 1: \quad 0.12 \times 3^{1} = 5/3$$

$$\exp = 2: \quad 0.12 \times 3^{2} = 5$$

Exemplo 2

Considere o SPF (b, n, expmin, expmax)= SPF (3,2,-1,2) , isto é, de base 3, 2 dígitos de mantissa, menor expoente igual a -1 e maior expoente 2. Para este sistema temos:

a) O menor exatamente representável:

$$x = (0.10)_3 \times 3^{-1} = \left(1\times 3^{-1} + 0\times 3^{-2}\right)\times 3^{-1} = \frac{1}{9} \quad \text{e} \quad y = 5 = (0.12)_3 \times 3^2$$

São exatamente representáveis, no entanto,

(x+y)= $(0.00010)_3 \times 3^2 + (0.12)_3 \times 3^2 = (0.1201)_3 \times 3^2$

Não é exatamente representável em SPF, uma vez que a mantissa é de 2 dígitos.

OBS: Pode ocorrer de outras propriedades consagradas no conjuntos dos números reais não serem verdadeiras, no sentido da exatidão da representação, no sistema de ponto flutuante normalizado, como as propriedades comutativa e associativa.

Exemplo 3

Dados x, y, z ϵ Reais e os Sistema de Ponto Flutuante normalizado SPF (3,2,-1,2) temos:

Se $x = \frac{5}{3} = (0.12)_3 \times 3^1$, $y = \frac{7}{27} = (0.21)_3 \times 3^{-1}$ *e*

$$z = \frac{8}{9} = (0.22)_3 \times 3^0$$

Temos:

e

$$x + (y + z) = 0.22 \times 3^1$$

$$(x + y) + z = 0.21 \times 3^1$$

Logo,

X+(y+z) ≠ (x+y) +z