Aritmética computacional - Aula5

Nicolas Chagas Souza

05/08/2022

Representação de números decimais em binário:

$5,5_{10} = 101,1_2$

Um número está representado em notação científica quando se encontra na forma $\pm F \times 10^E$, e encontra-se na forma **normalizada** quando $0 < F < 10$. O F denota a mantissa, também chamada de fração ou significando, e o E denota o expoente.

Para representar números binários utilizamos a base 2, logo, a representação de números binários em notação científica é da forma $\pm 1, zzzz... \times 2^{yyyy\cdots}$.

$5,5_{10} = 1,011 \times 2^2$

Representação IEEE754

A representação de números em pontos flutuantes segue o padrão IEEE754, que consiste em registrar os valores de:

- Mantissa (ou fração)
- Expoente
- Sinal

Que são utilizados para calcular o número da seguinte forma:

$(-1)^S \times (1 + F) \times 2^{E-bias}$

Um número de ponto flutuante é representado por uma palavra (32 bits) da seguinte forma:

31	30 ... 23	22 ... 0
Sinal (1 bit)	Expoente (8 bits)	Fração (22 bits)

Sendo o bit de sinal 0 para números positivos e 1 para números negativos.

Essa representação de ponto flutuante é chamada de precisão simples.

É possível representar em precisão dupla (double), que utiliza dois registradores para representação.

31	30 ... 19	18 ... 0 31 ... 0
Sinal (1 bit)	Expoente (11 bits)	Fração (52 bits)

O padrão IEEE754 estabelece os seguintes casos particulares:

Single precision		Double precision		Object represented
Exponent	Fraction	Exponent	Fraction	
0	0	0	0	0
0	Nonzero	0	Nonzero	± denormalized number
1–254	Anything	1–2046	Anything	± floating-point number
255	0	2047	0	± infinity
255	Nonzero	2047	Nonzero	NaN (Not a Number)

Exemplos

$5,5_{10} = 101,1_2 = 1,011 \times 2^2$

Sinal	Expoente	Fração
+	2	011

Sinal	Expoente (8 bits)	Fração (22 bits)
0	000 0001 0	011 0000 0000 0000 0000 0000

$9,25_{10} = 1001,01_2 = 1,00101 \times 2^3$

Sinal	Expoente	Fração
+	3	00101

Sinal	Expoente (8 bits)	Fração (22 bits)
0	000 0001 1	001 0100 0000 0000 0000 0000

Se usássemos complemento a 2 para representar os expoentes negativos, perderíamos a propriedade de que os binários ordenados implicam pontos flutuantes ordenados. Por esse motivo, o expoente é representado por excesso.

- Precisão simples: A representação dos expoentes válidos vai de 1 a 254, utilizamos um deslocamento (bias) de 127 unidades.
- Precisão dupla: A representação dos expoentes válidos vai de 1 a 2046, utilizamos um deslocamento (bias) de 1023 unidades.

$0,5_{10} = 0,1_2 = 1 \times 2^{-1}$

Sinal	Expoente	Fração
+	-1 → -1+127=126	0

Sinal	Expoente (8 bits)	Fração (22 bits)
0	011 1111 0	000 0000 0000 0000 0000 0000

Intervalos de representação

Precisão simples

Os menor e maior números representáveis em precisão simples são:

S	E + 127	F	Decimal
0	000 0000 0	000 ... 0000	$1,0... \times 2^{-126} \approx 1,2 \times 10^{-38}$
0	111 1111 0	111 ... 1111	$1,1... \times 2^{127} \approx 3,4 \times 10^{38}$

E em precisão dupla, temos:

S	E + 1023	F	Decimal
0	000 0000 0000	000 ... 0000	$1,0... \times 2^{-1022} \approx 2,2 \times 10^{-308}$
0	111 1111 0000	111 ... 1111	$1,1... \times 2^{1023} \approx 1,8 \times 10^{308}$