Cálculo Numérico

Representação Binária

Wellington José Corrêa

Universidade Tecnológica Federal do Paraná

17 de Junho de 2021

@correa.well

Representação Binária

A representação em ponto flutuante vista ate então é mais fácil de manusear pela sua familiaridade com a notação científica. No entanto, existe a Representação binária.

Representação Binária

Em 1985, o IEEE (Instituto de Engenheiros Elétricos e Eletrônicos) publicou um relatório chamado *Binary Floating Point Arithmetic Standard 754 - 1985*. Uma versão atualizada foi publicada em 2008 como *IEEE 754 - 2008*.

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Representação Binária

Em 1985, o IEEE (Instituto de Engenheiros Elétricos e Eletrônicos) publicou um relatório chamado *Binary Floating Point Arithmetic Standard 754 - 1985*. Uma versão atualizada foi publicada em 2008 como *IEEE 754 - 2008*.

Isso fornece padrões para números binários e decimais de ponto flutuante, formato para troca de dados, algoritmos para o arredondamento de operações aritméticas. Esses padrões são geralmente seguidos por todos os fabricantes de microcomputadores que usam hardware de ponto flutuante.

Representação Binária

É usada para um número real 64 bits (dígitos binários). Tem-se:

Representação Binária

É usada para um número real 64 bits (dígitos binários). Tem-se:

- O primeiro bit é um indicador de sinal (0 significa que o número é positivo e 1 para negativo);

Representação Binária

É usada para um número real 64 bits (dígitos binários). Tem-se:

- O primeiro bit é um indicador de sinal (0 significa que o número é positivo e 1 para negativo);
- 11 bits para o expoente c, chamado <u>característica</u>;

Representação Binária

É usada para um número real 64 bits (dígitos binários). Tem-se:

- O primeiro bit é um indicador de sinal (0 significa que o número é positivo e 1 para negativo);
- 11 bits para o expoente c, chamado característica;
- 52 bits, a saber, f chamada mantissa;

Representação Binária

É usada para um número real 64 bits (dígitos binários). Tem-se:

- O primeiro bit é um indicador de sinal (0 significa que o número é positivo e 1 para negativo);
- 11 bits para o expoente c, chamado <u>característica</u>;
- 52 bits, a saber, f chamada <u>mantissa</u>;
- A base para o expoente é 2.

Representação Binária

É usada para um número real 64 bits (dígitos binários). Tem-se:

- O primeiro bit é um indicador de sinal (0 significa que o número é positivo e 1 para negativo);
- 11 bits para o expoente c, chamado <u>característica</u>;
- 52 bits, a saber, f chamada <u>mantissa</u>;
- A base para o expoente é 2.

	10	9	...	0	-1	-2	...	-51	-52
s	c_{10}	c_9	...	c_0	f_1	f_2	...	f_{51}	f_{52}

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Representação Binária

Para economizar armazenamento e fornecer uma representação única de cada número em ponto flutuante, é imposta uma normalização da forma:

$$(-1)^s \, 2^{c-1023} \, (1+f) \, .$$

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Representação Binária

Exemplo (APS 1)

Considere o número de máquina

$$0\,{\color{blue}10000000011}\,{\color{red}1011100100010000000000000000000000000000000000000000}.$$

Exemplo (APS 1)

Considere o número de máquina

$$0\,{\color{blue}10000000011}\,{\color{red}1011100100010000000000000000000000000000000000000000}.$$

Solução: Note que o item mais à esquerda é $s = 0$, o que indica que o número é positivo.

Representação Binária

Exemplo (APS 1)

Considere o número de máquina

$$0\,10000000011\,1011100100010000000000000000000000000000000000000000.$$

Solução: Note que o item mais à esquerda é $s = 0$, o que indica que o número é positivo.

Os próximos 11 bits em azul, que fornecem a característica, são equivalentes ao número decimal

$$\begin{aligned} c &= 1 \cdot 2^{10} + 0 \cdot 2^{9} + \ldots + 0 \cdot 2^{2} + 1 \cdot 2^{1} + 1 \cdot 2^{0} \\ &= 1024 + 2 + 1 \\ &= 1027. \end{aligned}$$

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Representação Binária

Os 52 bits finais especificam que a mantissa é

$$f = 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} + 1 \cdot 2^{-4} + 1 \cdot 2^{-5} + 1 \cdot 2^{-8} + 1 \cdot 2^{-12}$$

.

Representação Binária

Os 52 bits finais especificam que a mantissa é

$$
\begin{aligned}
f &= 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} + 1 \cdot 2^{-4} + 1 \cdot 2^{-5} + 1 \cdot 2^{-8} + 1 \cdot 2^{-12} \\
&= \left(\frac{1}{2} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{256} + \frac{1}{4096} \right).
\end{aligned}
$$

.

Representação Binária

Os 52 bits finais especificam que a mantissa é

$$
\begin{aligned}
f &= 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} + 1 \cdot 2^{-4} + 1 \cdot 2^{-5} + 1 \cdot 2^{-8} + 1 \cdot 2^{-12} \\
&= \left(\frac{1}{2} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{256} + \frac{1}{4096} \right) .
\end{aligned}
$$

.

Representação Binária

Os 52 bits finais especificam que a mantissa é

$$
\begin{aligned}
f &= 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} + 1 \cdot 2^{-4} + 1 \cdot 2^{-5} + 1 \cdot 2^{-8} + 1 \cdot 2^{-12} \\
&= \left(\frac{1}{2} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{256} + \frac{1}{4096} \right).
\end{aligned}
$$

Portanto, esse número de máquina representa precisamente o número decimal

$$
\begin{aligned}
&(-1)^s \, 2^{c-1023} \, (1+f) \\
&= (-1)^0 \, 2^{1027-1023} \left[1 + \left(\frac{1}{2} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{256} + \frac{1}{4096} \right) \right] \\
&.
\end{aligned}
$$

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Representação Binária

Os 52 bits finais especificam que a mantissa é

$$
\begin{aligned}
f &= 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} + 1 \cdot 2^{-4} + 1 \cdot 2^{-5} + 1 \cdot 2^{-8} + 1 \cdot 2^{-12} \\
&= \left(\frac{1}{2} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{256} + \frac{1}{4096} \right) .
\end{aligned}
$$

Portanto, esse número de máquina representa precisamente o número decimal

$$
\begin{aligned}
&(-1)^s \, 2^{c-1023} \, (1 + f) \\
&= (-1)^0 \, 2^{1027-1023} \left[1 + \left(\frac{1}{2} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{256} + \frac{1}{4096} \right) \right] \\
&= 27,56640625 \, .
\end{aligned}
$$