Aluno: Natan Henrique Sanches 11795680 18/03/2020
ICMC/USP
Bacharelado em Ciências da Computação

O PADRÃO IEEE754

1 INTRODUÇÃO: O QUE É O PADRÃO IEEE754?

Do inglês “IEEE Standard for Floating-Point Arithmetic”, foi introduzido em 1985 por conta da intensa divergência que, até então, encontrava-se presente no método de computação de pontos flutuantes. Antes dele, cada fabricante de computadores adotava um formato diferente, causando grandes problemas na compatilidade entre as máquinas.

Dessa forma, o padrão, inicialmente definido pelo Instituto de Engenheiros Eletricistas e Eletrônicos (IEEE), atuou como uma padronização para o sistema de números binários em ponto flutuante, tanto em suas representações, quanto em suas operações.

Atualmente, o IEEE 754 é a representação mais comum para números reais nos computadores, até mesmo em Macs e computadores com plataforma Unix.

2 COMPOSIÇÃO DO PADRÃO IEEE 754

- **Sinal da mantissa:** representa o sinal do número real em questão. 0 para número positivo e 1 para número negativo. Representado pela letra S maiúscula.
- **Expoente:** expoente que acompanha uma base 2, representando tanto os expoentes positivos quanto negativos. O expoente será representado pela letra E maiúscula.
- **Mantissa:** trata-se da parte fracionada do número real em questão. Será representado pela letra M maiúscula.

A estrutura genérica de um número real no padrão IEEE 754 se constrói da forma:

$$(-1)^S \times 1,M \times 2^E$$

A representação também pode ser feita em bits (binário, mais comum), assim representando *floats* e *doubles*. Para precisão simples (*float*), aderindo a uma faixa de valores entre 2^{-126} e 2^{127} tem-se:

1 bit	8 bits	23 bits
Sinal (S)	Expoente em excesso (E)	Mantissa
Float (32 bits = 4 bytes)		

Como extensão, para precisão dobrada (*double*), entre uma faixa de valores de 2^{-1022} a 2^{1023}, tem-se:

1 bit	11 bits	52 bits
Sinal (S)	Expoente em excesso (E)	Mantissa
Double (64 bits = 8 bytes)		

A união desses três componentes em precisão simples formará um *float* de 32 bits, equivalente a 4 bytes. Analogamente, a união dos três componentes em precisão dobrada resultará em um *double*, com 8 bytes.

O excesso do expoente, ou BIAS, será importante para a representação do ponto flutuante em binário e será definido adiante, no tópico 4.

3 COMO CONSTRUIR UM NÚMERO REAL NO PADRÃO IEEE 754

Inicialmente, passaremos o número na base 10 para base 2 através do método de divisões sucessivas. Será transformada tanto a sua parte inteira quanto a sua parte fracionada.

Logo após, “andamos” com a vírgula até que o número tome uma forma semelhante à estrutura genérica do IEEE 754. O número de casas que foram ultrapassadas será o valor do expoente (E), considerando que o andamento para esquerda implica em um expoente de valor positivo, enquanto o andamento para direita implica em um expoente de valor negativo.

Feito isso, determinaremos o valor de S de acordo com o sinal do número em questão. Para tornar o método mais claro, usaremos alguns números de exemplo:

- Exemplo 1: Transforme o número real 17,25 para o padrão IEEE 754.

Como dito inicialmente, transformemos o número real de base 10 em base 2:

- Parte inteira: $17 = 10001_2$

Divisão	*Quocientes*				
2	*17*	*8*	*4*	*2*	*1*
	1	0	0	0	1
	Restos				

- Parte fracionária: $0,25 = 0,01_2$

Multiplicação	*0,25*	*0,5*	*1*
2	-	0	1
	Algarismo binário		

Depois de feitos os algoritmos, podemos escrever o número 17,25 como $10001,01_2$.

Logo após, para transformarmos o número $10001,01_2$ na forma de notação científica, andamos com a vírgula para a esquerda por quatro casas e, por consequência, determinamos nosso valor de E:

$17,25 = 10001,01_2 = 1,000101_2 * 2^4$.

Só resta, agora, determinar o sinal do número. Como vimos acima, um número positivo tem o equivalente zero em binário. Assim:

$(-1)^S * 1,000101_2 * 2^4 = (-1)^0 * 1,000101_2 * 2^4 = 1,000101_2 * 2^4$.

Finalizado, concluímos que uma das representações do número 17,25, no padrão IEEE 754, é $1,000101_2 * 2^4$. A representação final, na forma binária, será vista logo após.

4 REPRESENTAÇÃO DO PADRÃO IEEE 754 NA FORMA BINÁRIA

O número também pode ser descrito na forma binária, equivalente ao que é encontrado quando prosseguido com todos os passos citados anteriormente. O resultado da representação do real, o *float*, na forma binária, será formado de 32 dígitos, ou seja, 32 bits. Proporcionalmente, o *double* será formado por 64 dígitos (ou bits).

Na máquina, os números reais são lidos e processados na forma binária, portanto vale a pena reconhecer a importância de conhecer o processo de conversão para tal.

Basta que destrinchemos o número adquirido em três grupos: o sinal, o expoente em excesso e a mantissa, e depois agrupemos. Para conveniência, trabalharemos com a representação de um *float*, porém vale dizer que o processo para determinação de *doubles* é, praticamente, idêntico.

- **Sinal:** como foi visto anteriormente, o sinal (S) é representado por apenas um bit: 0 representará um número positivo e 1 representará um número negativo.
- **Expoente com repetição:** será representado por 8 bits para o *float* e 11 bits para o *double*, onde os mesmos serão compostos pela soma *BIAS* + *E* em binário. Para floats, o valor de BIAS é 127; para o tratamento de *doubles*, o valor de BIAS é 1023. O BIAS, de modo geral, é uma constante de balanceamento entre os números positivos e negativos.
- **Mantissa:** será representada pela parte fracionária do número encontrado no tópico anterior, ou seja, pelo o que foi representado por M.

Transformaremos o número encontrado no exemplo 1 na sua representação na forma binária:

- Exemplo 2: Transforme o número 17,25 em sua representação binária no padrão IEEE 754.

Nesse caso, 17,25 pode ser representado na forma $1{,}000101_2 * 2^4$ como visto no exemplo anterior. Faremos por partes e, logo após, o agrupamento.

- **Sinal:** Como 17,25 é um número positivo, representaremos seu bit por 0.
- **Expoente em excesso:** Nesse caso, em que E = 4 e BIAS = 127 (pois trata-se de um float), devemos converter o número 131 (127 + 4) para binário. Utilizando o mesmo procedimento de divisões sucessivas, teremos que $131 = 10000011_2$.
- **Mantissa:** Como a mantissa (parte fracionária) de 17,25 não possui 23 bits, preencheremos todos os bits, após os ocupados, por 0 até que se completem os 23 bits. Dessa forma, o valor da mantissa será: 00001010000000000000000.

Após tudo ter sido identificado, uniremos em sequência para que o *float* se torne completo, em binário:

Sinal | Expoente em excesso | Mantissa

0 | 10000011 | 00001010000000000000000

A representação final do *float* acaba sendo 01000001 10000101 00000000 00000000.

5 ARITMÉTICA NO PADRÃO IEEE 754 PARA PONTOS FLUTUANTES

Nesse tópico, abrangeremos algumas normas de aritméticas que valem para operações com os pontos flutuantes tomados. Muitas de tais relações já são conhecidas por meio da matemática elementar, mas algumas (as mais populares) valem a pena ser citadas.

Considerando um ponto flutuante X, são válidas as seguintes colocações:

- $X + \infty = \infty$
- $X - \infty = -\infty$
- $\infty - \infty =$ Indeterminação
- $X / 0 = \pm\infty$ (para $X \neq 0$)
- $X / \infty = 0$
- $\pm\infty \ / \ \pm\infty =$ Indeterminação
- $0 / 0 =$ Indeterminação
- $\pm\infty \times 0 =$ Indeterminação

Obs.: tratando-se de pontos flutuantes, assumimos a inexistência de números complexos.

Pode-se relatar, também, as quatro operações básicas da aritmética com pontos flutuantes: a adição, a subtração, a multiplicação e a divisão:

- **Adição e subtração:** para tais operações, os números devem ter o mesmo expoente (E). Após isso, as operações de adição e subtração serão feitas normalmente.
- **Multiplicação e divisão:** para essas, ambas as mantissas serão multiplicadas ou divididas, a depender da operação, e os expoentes serão somados (para o caso da multiplicação) ou subtraídos (para o caso da divisão).

6 CONCLUSÃO

Em suma, o padrão IEEE 754 estruturou uma maneira padronizada de representar números reais (ponto flutuante) nas diferentes CPUs da época e que, consequentemente, se estendeu até atualmente. Tal modelo foi proposto visto as dificuldades de compatibilidade geradas pelas diferentes maneiras de se representar pontos flutuantes.

O processo de conversão de um número real para o padrão citado pode ser feito de maneira intuitiva através de uma sucessão de passos, que inclui, na ordem:

Conversão do número em base 10 para base 2;

Representação do número na forma de notação científica na base 2;

Determinação do sinal do número;

Transformação da representação encontrada para binário, visto anteriormente no tópico 4.

Assim sendo, toma-se conhecimento da aritmética do padrão IEEE 754 ao analisar o processo de transformação que o número real é submetido, notando também sua semelhança com a aritmética de números reais da matemática elementar. Além disso, valida a importância para sua criação e bem como o motivo pelo qual foi criado.

7 REFERÊNCIAS BIBLIOGRÁFICAS

- Geeks for Geeks – IEEE Standard 754 Floating Point Numbers (encontrado em: https://www.geeksforgeeks.org/ieee-standard-754-floating-point-numbers/).
- PUC-Rio – Pontos Flutuantes (encontrado em http://www-di.inf.puc-rio.br/~endler/courses/inf1612/aula-9.pdf).
- Professor Pantoja – Arquitetura de Computadores: IEEE 754 (encontrado em https://www.youtube.com/watch?v=ealNNf7lGoU)
- Toda a Matemática - Padrão IEEE 754 (encontrado em https://www.youtube.com/watch?v=PDgT0T0Yodo)