A REPRESENTAÇÃO DECIMAL DE PONTO FLUTUANTE E A ARITMÉTICA PROPOSTAS PELA IEEE754

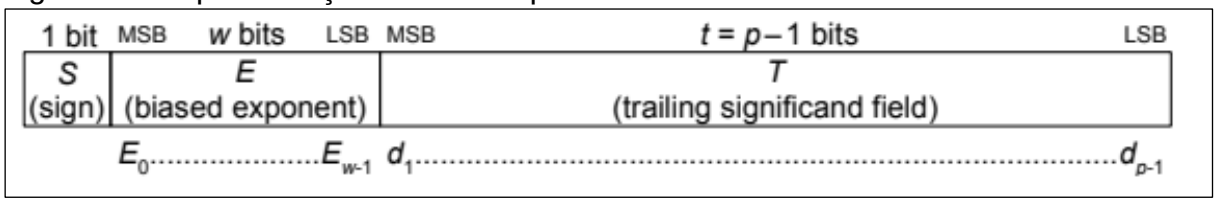
Matheus Henrique de Cerqueira Pinto
Nº USP: 11911104
Laboratório de Introdução à Ciência da Computação I (SCC0222)

Durante a década de 70, formou-se um grupo de trabalho nomeado IEEE p754, sob organização do Institute for Electrical and Electronics Engineers (IEEE), com o intuito de padronizar a representação numérica dos computadores produzidos à época para que, então, as máquinas obtivessem resultados únicos ao executar programas computacionais idênticos. Assim, liderado pelo professor William Kahan e composto por engenheiros de empresas como Apple, Zilog, DEC, Intel, Hewlett Packard, Motorola e National Semiconductor, o grupo desenvolveu o padrão IEEE754 (ROLIM, 2009).

Especificamente, de acordo com o padrão IEEE754 (IEEE, 2019, p. 8), definem-se “conjuntos de técnicas e maneiras comercialmente viáveis para executar a aritmética entre binários e pontos flutuantes decimais” que seguiram princípios, em linhas gerais, que priorizam a portabilidade, precisão e segurança numérica e, por fim, eficiência de aplicações computacionais (IEEE, 2019). Destarte, trataram-se problemas “mal condicionados”, produtores de resultados incorretos a partir do arredondamento usado nas operações aritméticas com números com grandezas distintas (ROLIM, 2009).

Logo, definiu-se a representação de um ponto flutuante no formato binário através de uma composição de três partes: a primeira, pelo sinal do ponto flutuante; a segunda, pelo expoente na base binária; por fim, a terceira, denominada mantissa ou significante, que descreve a parte binária na base dois. Faz-se necessário notar, além disso, que o padrão de ponto flutuante, sob a ótica da IEEE754, varia entre a precisão simples – ou *single* – (composta por 32 bits) e precisão dupla – ou *double* – (64 bits), o que confere uma extensão maior para representações numéricas (RODRIGUES, 2002).

Figura 1 – Representação binária de ponto flutuante



Fonte: IEEE (2019, p. 19)

Na Figura 1, demonstra-se graficamente a construção binária de um ponto flutuante em que “S” faz referência ao sinal do número (*sign*); “E”, ao expoente somado a uma constante numérica – 127, na precisão simples e 1023, na precisão dupla (IEEE, 2019) – (*biased exponent*); e “T”, ao significante (*trailing significand field*). Exemplificando: para transpor o número 22,2 para o padrão IEEE754, é necessário realizar inicialmente sua conversão para binário (tratando o resto através de sucessivas divisões da parte fracionária e considerando como resultado a inteira) e então, formatá-lo, conforme mostra a Figura 2.

Figura 2 – Conversão de 22,2 para IEEE754

$$N = 10110,0011... \times 2^0 \ (N \text{ em binário e não formatado})$$
$$S = 0 \ (\text{positivo}),\ E = 127 \ (127 + 0),\ T = ?$$

$$N = 1,01100011 \times 2^4 \ (\text{formatado})$$
$$S = 0 \ (\text{positivo}),\ E = 131 \ (127 + 4),\ T = 01100011...$$

$$\text{logo (com E em binário)},$$
$$S = 0,\ E = 10000011,\ T = 01100011... \text{ e, da junção, conclui-se que}$$
$$N = 01000001101100011001100110011001$$

Fonte: elaborada pelo autor (2020)

Faz-se necessário notar, entretanto, que os passos precedentes à formatação numérica se alteram caso o número a ser convertido não seja um decimal inteiro. Por exemplo, ao se tratar o número 2 a fim de obter um número no padrão IEEE754, ao se converter para binário, não é necessário utilizar o método das sucessivas divisões da parte fracionária, uma vez que esta é igual a 0. Logo, adotando os mesmos passos descritos pela Figura 2, tem se que $2_2 = 10$. Assim, formatando-se o número, obtém-se que $S = 0$, $E = 128\,(127 + 1) = 10000000\ (\text{em binário})$, $T = 00000000000000000000000$. Conclui-se que o número 2 equivale a $01000000000000000000000000000000$.

Com valores numericamente adaptados e armazenados, é possível realizar operações aritméticas básicas entre números no padrão IEEE 754 utilizando passos comuns tanto na divisão e multiplicação quanto para a soma e subtração. Nestas duas últimas, Vahid (2009) explicita que os passos necessários derivam da comparação dos expoentes, sendo que, quando divergentes, há a necessidade de adaptação do número com menor expoente para que este apresente expoente equivalente ao outro número da operação. Exemplifica-se, através da Figura 3, a referida operação:

Figura 3 – Soma e subtração entre 20 e 2

$$N_1 = 20 \ (N_1 \text{ em decimal e não formatado})$$
$$N_1 = 10100 \ (N_1 \text{ em binário})$$
$$S = 0 \ (\text{positivo}),\ E = 131 \ (127 + 4),\ T = 01000000000000000000000$$

$$N_2 = 2 \ (N_2 \text{ em decimal e não formatado})$$
$$N_2 = 10 \ (N_2 \text{ em binário})$$
$$S = 0 \ (\text{positivo}),\ E = 128 \ (127 + 1),\ T = 00000000000000000000000$$

$$\text{Assim, } N_1 = 2^4 \times 1,01 \text{ e } N_2 = 2^1 \times 1,0.$$
$$\text{Igualando-se os expoentes, tem-se } N_2 = 2^4 \times 0,001 \text{ (base igual a } N_1)$$

$$\text{Logo, } N_1 + N_2 = 1,011 \times 2^4 = 22 \text{ (decimal) e } N_1 - N_2 = 1,001 \times 2^4 = 18.$$

$$\text{Portanto, } N_1 + N_2 = 01000001101100000000000000000000 \text{ e}$$
$$N_1 - N_2 = 01000001100100000000000000000000$$

Fonte: elaborada pelo autor (2020)

Nas multiplicações e divisões, o processo de equivalência exponencial não é necessário, uma vez que, a partir da notação científica, trabalha-se com a multiplicação ou divisão das mantissas e a soma (na multiplicação) ou subtração (na divisão) dos expoentes na respectiva base (VAHID, 2009). Assim, exemplificam-se as operações na Figura 4 e Figura 5.

Figura 4 – Multiplicação e divisão entre 20 e 2

$$N_1 = 20$$
$$S = 0 \ (\text{positivo}),\ E = 131 \ (127 + 4),\ T = 01000000000000000000000$$

Fonte: elaborada pelo autor (2020)

Figura 5 – Multiplicação e divisão entre 20 e 2 (continuação)

$$N_2 = 2$$
$$S = 0 \text{ (positivo)}, E = 128 \ (127 + 1), T = 00000000000000000000000$$

$$N_1 = 2^4 \times 1,01$$
$$N_2 = 2^1 \times 1,0$$

$$N_1 \times N_2 = 2^{4+1} \times (1,01 \times 1,0) = 2^5 \times 1,01$$
$$S = 0 \text{ (positivo)}, E = 132 \ (127 + 5), T = 01000000000000000000000$$
$$N_1 \times N_2 = 01000010001000000000000000000000 \text{ (IEEE 754)}$$

$$N_1 \div N_2 = 2^{4-1} \times (1,01 \div 1,0) = 2^3 \times 1,01$$
$$S = 0 \text{ (positivo)}, E = 130 \ (127 + 3), T = 01000000000000000000000$$
$$N_1 \div N_2 = 01000001001000000000000000000000 \text{ (IEEE 754)}$$

Fonte: elaborada pelo autor (2020)

Vale ressaltar que, para além das operações apresentadas, também existem as de potenciação, radiciação e de arredondamento, com padrões e técnicas próprias. Ademais, também são tratadas pela IEEE 754 operações envolvendo infinito e dados não-numéricos (NaN), também com representações específicas. Há de se ressaltar, por fim, a importância do padrão que, ao unificar os métodos para calcular pontos flutuantes, traz confiabilidade e portabilidade operacional.

REFERENCIAL BIBLIOGRÁFICO

IEEE – INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS. **IEEE Standard for Floating-point Arithmetic**. 2019. Disponível em: https://standards.ieee.org/standard/754-2019.html. Acesso em: 18 mar. 2020.

RODRIGUES, M. I. **Projeto de uma unidade aritmética de ponto flutuante – Padrão IEEE 754 – implementada em computação reconfigurável**. 86 f. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) – Universidade de São Paulo, São Carlos, 2004.

ROLIM, A. U. A. **Desenvolvimento de uma FFT utilizando ponto flutuante para FPGA**. 83 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Pernambuco, Recife, 2009.

VAHID, F. **Sistemas Digitais**. Porto Alegre: Bookman Editora, 2009.