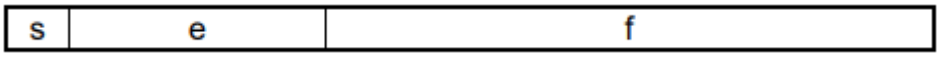
O padrão ieee funciona como uma formalidade aos fabricantes de compiladores, computadores, bibliotecas ao usar a aritmética binária de ponto flutuante. Essa padronização permite a portabilidade de softwares e criar novos tipos de dados.

Esse problema ainda ocorre, mas faz-se o uso de técnicas de verificação de resultados para contornar tal obstáculo. Essas técnicas envolvem arredondamentos, repetição de operações com mais precisão(precisão dupla), repetindo alterando alguns valores para verificar a estabilidade, uso de dígitos de guarda, aplicando o valor em expressões matemáticas a fim de se obter o resultado próximo, dentre outros.

Aritmética de ponto flutuante: Ela armazena o ponto flutuante em um conjunto de números binários divididos em: sinal, expoente e mantissa. O sinal pode ser 0 ou 1, sendo ele positivo ou negativo, respectivamente. O expoente é diretamente ligado a mantissa, uma vez que a mantissa é um número real que começa com 0. Ou seja, o expoente indica quantas vezes a mantissa será deslocada para retornar o valor desejado.



Formato da representação de números reais

Geralmente acaba-se arredondando para padronização do truncamento. Esse arredondamento varia de acordo com o projetista da linguagem ou do compilador. O arredondamento pode ser feito para o próximo número, para o menor possível ou retirando algumas casas decimais.

Padrão IEEE para ponto flutuante: O padrão recomenda um número de bits específicos para cada tipo de variável. Assim, percebe-se que há um limite para cada tipo, o qual pode ser arredondado quando ocorre overflow. Cada precisão de ponto flutuante possui um certo número de bits para expoente e mantissa. Dessa forma, há precisões que vão de 32 bits a mais de 79 bits, por exemplo.

Exemplos:

- Número 0.45:
  - Sinal: 0 (positivo)
  - Expoente: O expoente é -2, então fazemos 127-2 = 125 em binário é: 01111101
  - Mantissa: 110 porque é o número em binário (exclui o primeiro 1)
  - Resultado: 001111101110
- Número 0.8:
  - Sinal: 0 (positivo)
  - Expoente: O expoente é -1, então fazemos 127-1 = 126 em binário é: 01111110
  - Mantissa: 10011001100
  - Resultado: 00111111010011001100

Observação: Fazemos 127 + expoente para obter o expoente como um número binário sem sinal.

Fontes:
https://www.lia.ufc.br/~valdisio/download/ieee.pdf
https://learn.microsoft.com/pt-br/cpp/build/ieee-floating-point-representation?view=msvc-170
https://ufsj.edu.br/portal2-repositorio/File/nepomuceno/ca/03a-compieee.pdf