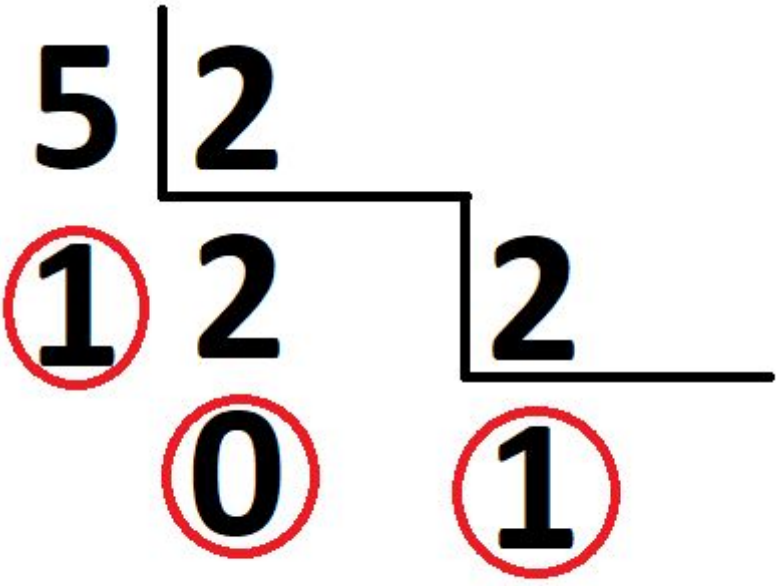
O padrão IEEE754 é um padrão de regras de normalização que define como podemos representar números reais como número binários com ponto flutuante. Esse padrão foi inventado, pois cada fabricante de computadores definia um formato de representação diferente. Em 1985 o IEEE (Instituto dos Engenheiros Elétricos e Eletrônicos) padronizou a ideia inventada por Willian Kahan em 1980.

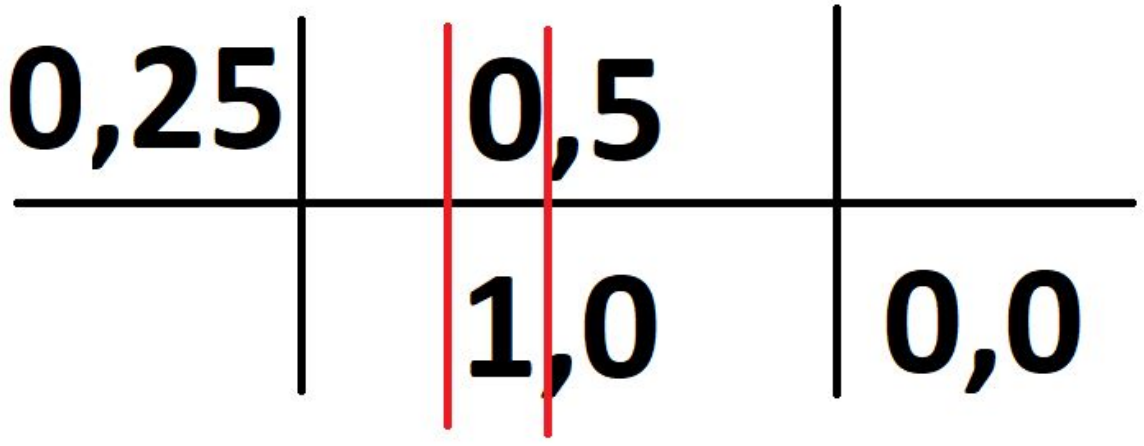
Podemos observar no exemplo a conversão de um número real para binário.

Escolheremos o número 5,25 e seguiremos alguns passos para defini-lo:

- 1° passo:
    - Transformar o número 5 em binário pelas divisões sucessivas por 2




    - Escolhemos então os números circulados em vermelho e escrevemos o número 5 em binário: 00000101
- 2° passo:
    - Transformar o número 0,25 pela multiplicação sucessiva por 2



    - Sempre que chegar em 1 ou mais subtraímos 1, até conseguir chegar em 0 e a conta acaba
    - Após terminar as multiplicações devemos pegar a parte inteira, assim o temos o número
    - Então 0,25 em binário é 0,01
- 3° passo:
    - Somar os números binários então o número real 5,25 em binário é igual a 00000101,01
- 4° passo:
- Para transformar um número binário no padrão IEEE754 o número tem que estar escrito na forma 1,m1m2m3… * 2^E
    - Então transformaremos nosso número 101,01 (5,25) para o formato IEEE754 logo transformaremos nosso número em algo parecido com uma notação científica, chamada notação normalizada, então o número ficará escrito como 1,0101 * 10^2
- 5° passo:
    - Agora podemos definir o número binário completo utilizando de uma máquina de precisão simples ou de precisão dupla
    - Para uma máquina de precisão simples (32 bits) dividimos esse 32 bits em:
        - 1 bit para o sinal
        - 8 bits para o E+BIAS:
            - O E é o número que o 2 está sendo elevado e o BIAS é predefinido para cada máquina de precisão, sendo a máquina de precisão simples o BIAS vale 127.
            - O BIAS é calculado a partir de uma fórmula, $2^{k-1} - 1$, onde k é o número número de bits no expoente, na máquina de precisão simples 8, então temos (2^8-1) - 1 = (2^7) - 1 = 128 - 1 = 127, dessa forma o expoente abrange de -126 até o 127.
        - 23 bits para a mantissa, a parte fracionária do número que vem após o 1, ao observamos o exemplo 1,m1m2m3… , os ms são as mantissas e no caso da máquina de precisão simples ele vai até o m23, sendo então o número, 1,m1m2...m23
    - Para uma máquina de precisão dupla (64 bits) dividimos esse 64 bits em:
        - 1 bit para o sinal
        - 11 bits para o E + BIAS, sendo o BIAS para máquinas de precisão dupla 1023
        - O BIAS é calculado a partir da fórmula, $2^{k-1} - 1$, onde k é o número número de bits no expoente, na máquina de precisão dupla 11, então temos (2^11-1) - 1 = (2^10) - 1 = 1024 - 1 = 1023, dessa forma o expoente abrange de -1022 até o 1023.
        - 52 bits para a mantissa, exemplificando, 1,m1m2...m52
- 6° passo:
    - Observar o número que eleva o 2, no nosso caso o número 2 e somá-lo ao BIAS

        - Para a máquina de precisão simples 127, logo 2+127 = 129 e transformar o 129 em um número binário. 129 em binário = 10000001 esse será o número salvo na parte do E+BIAS
        - Para a máquina de precisão dupla 1023, logo 2+1023 = 1025 e transformar o 1025 em um número binário. 1025 em binário = 1000000001 esse será o número salvo na parte do E+BIAS
- 7° passo:
    - Observar o sinal do número, se ele for positivo o número salvo será 0 e quando o número for negativo o número salvo será 1, logo nosso número salvo nesse exemplo será 0
- 8° passo:
    - Observemos a mantissa para definirmos seu número binário, no nosso caso o número real 5,25 nos gerou o número binário 1,0101 * 10^2, nossa mantissa será 0101 nos 4 primeiros bits e o resto será preenchido com 0 até chegamos a 23 bits, isso para máquinas de precisão simples
    - Para uma máquina de precisão dupla nossa mantissa continuaria sendo 0101 nos 4 primeiros bits e preencherá o resto com 0. porém até chegarmos a 52 bits.
- 9° passo:
    - Escreveremos nosso número na sequência: S E+BIAS M, importante lembrar que um número no padrão IEEE754 sempre deverá começar com 1, mesmo que ele seja um número, como por exemplo 0,25, que em binário seria 0,01, mas no padrão IEEE754 ele é escrito como 1 * 2^-2, com isso subentendesse que o primeiro número sempre será 1, portanto ele não precisa ser salvo em nenhum bit
    - Para o nosso exemplo 5,25 = 1,0101 * 2^2 o nosso número inteiro será salvo, na sequência S E+BIAS M, como:
    - S = 0, pois o número é positivo
    - E+BIAS = 1000 0001 (para máquinas de precisão simples) e 1000000001 (para máquinas de precisão dupla)
    - M = 0101 0000 0000 0000 0000 000 (para máquinas de precisão simples) e M = 0101 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0 (para máquinas de precisão dupla)
- Assim transformamos o número real 5,25 no número binário 0 1000 0001 0101 0000 0000 0000 0000 000 (na máquina de precisão simples) e no número binário 0 1000000001 0101 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0 (para máquinas de precisão dupla)

Resumindo:

- Para definirmos números reais como binários com pontos flutuantes utilizamos uma configuração definida por S M * 2^E. O S é o bit armazenado para o sinal do número, sendo 0 para números positivos e 1 para números negativos. O M é a mantissa, que é definida pela parte fracionária do número, em máquinas de precisão simples (32 bits) a mantissa consegue armazenar 23 bits enquanto em máquinas de precisão dupla (64 bits) a mantissa pode armazenar 52 bits. O E é o definido pelo

expoente que o 2 é elevado. Existe também o BIAS que é algo que é predefinido para a hora que formos escrever o real em binário.

As operações aritméticas se dão com os números em suas formas de padrão normalizado cada número tem sua mantissa e sua parte exponencial. As operações são: adição, subtração, multiplicação e divisão.

As operações se darão da mesma forma como as operações entre números em notação científica.

A adição ocorre e a subtração ocorrem quando os números tem o expoente elevado a potências iguais, a base desses números será sempre 2, logo casa exista o número 1,1 * 2^3 e o número 1,01 * 2^3, eles poderão ser somados ou subtraídos. Caso somemos tais números teremos como resultado o número 2,11 * 2^3. Se subtrairmos esses números, na ordem apresentada teremos 0,09 * 2^3 e transformaremos ele no número 9 * 2^1.

Caso existam números com mesma base, no caso 2, porém potências diferentes eles ainda poderão ser somados, precisamos apenas igualar suas potências. Como por exemplo 1,3 * 2^3 e 1,4 * 2^2, para somar ou subtraí-los precisamos transformar o 1,4 * 2^2 em um número com base 2 elevado ao cubo, então o número escrito elevado ao cubo será 0,14 * 2^3, ao somarmos esses números teremos como resultado 1,44 * 2^3. Caso subtrairmos, na ordem apresentada teremos o número, 1,16 * 2^3.

Para as operações de divisão e multiplicação o número não precisa ter o mesmo expoente elevando a base, pois utilizaremos a regra da multiplicação de mesma base e a regra da divisão de mesma base.

Para a multiplicação, se multiplicarmos dois números, por exemplo, x * 2^2 e y * 2^2, teremos o número (x*y) * 2^4, pois ao multiplicar dois números com a mesma base copiaremos a base após o número e somamos os expoentes, um exemplo complementar seria, (1,1 * 2^2)  * (1,1 * 2^2), a conta seguiria da seguinte maneira (1,1 * 1,1) * 2^(2+2), resultando em 1,21 * 2^4.

Para a divisão, se dividirmos dois números, por exemplo, x * 2^3 e y * 2^2. teremos o número (x/y) * 2^1, pois ao dividir dois números com a mesma base copiaremos a base após o número e subtraímos os expoentes, um exemplo para esclarecer seria, (1,6 * 2^4) / (0,8 * 2^2), passo a passo temos, (1,6/0,8) * 2^(4-2), resultando no número 2 * 2^2.

Utilizei os links abaixo como fonte de pesquisa:

- https://pt.wikipedia.org/wiki/IEEE_754
    - Utilizei o wikipédia para a parte histórica e para explicar a utilidade e o porque o padrão foi criado
- https://www.youtube.com/watch?v=PDgT0T0Yodo
    - Utilizei o vídeo do canal toda a matemática para explicar como transformar um número real em um número binário com ponto flutuante
- https://en.wikipedia.org/wiki/Exponent_bias

    - Utilizei o wikipédia para explicar de onde vem o BIAS
- https://en.wikipedia.org/wiki/Floating-point_arithmetic
    - Utilizei o wikipédia para explicar as operações aritméticas entre números com padrão normalizado