Ponto flutuante

Universidade Estadual de Mato Grosso do Sul – UEMS
Ciência da Computação
Linguagem de Montagem
Prf Dr Osvaldo Vargas Jaques
ojacques@comp.uems.br

Numero fracionário versos ponto flutuante

- Um número em ponto flutuante é escrito como
  - $\pm MxB^{\pm e}$

    M : mantissa, a parte fracionária

    B : a base

    e: expoente
- Para obter o número em ponto flutuante converte-se o número para a base na qual será armazenado, normaliza-o e por fim separa-se mantissa, expoente e sinais.
  - Exemplo:

    Assumindo:

    - 1 bit para o sinal do número
    - 1 bit para sinal do expoente
    - 4 bits para expoente
    - 10 bits para a mantissa

Numero fracionário x ponto flutuante

- Já que vamos Um número fracionário pode ter qualquer número inteiro que você quiser, como 101,32.
- Já um número em ponto flutuante, segue uma regra chamada **notação científica**. O padrão IEEE 754, tem o seguinte comportamento:

  A parte inteira é sempre 1.0_2 seguida da parte fracionária. E todo número binário fracionário é multiplicado por 2^n.
- Para entender, vamos representar 5.75_{10} em ponto flutuante, em um padrão onde a parte inteira seja 0.
  - Convertendo em binário temos 101.11_2
  - Podemos escrever 101.11_2 só com partes fracionárias (normalizar), como:
    - 0.10111×2^3
    - Assim, $10111_2 = 23_{10}$ é a mantissa
    - $3_{10} = 011_2$ é o expoente
    - Separando sinais, mantissa e expoente tem-se:
      - Sinal do número: (+) 0
      - Sinal do expoente: (+) 0
      - Expoente: 011 (2)
      - Mantissa: 10111
      - Portanto, tem-se $0\ 0\ 0011\ 0000010111_2$
      - O expoente 3 diz para ***"flutuar"*** o ponto 3 casas para a esquerda

Sinal número	Sinal expoente	Expoente	Mantissa
0	0	0011	0000010111

Ponto flutuante

- Assim, um valor fracionário é sempre descrito como:

$$valor = m.2^{e}$$

- Onde m é chamada de mantissa e contém o valor binário entre 0.0 e 1.0. E o expoente e, com base 2, desloca o ponto para um dos lados. Ou seja, faz o ponto "flutuar" (daí o nome). Por exemplo: O valor 0,5 pode ser escrito como $\mathbf{1.2^{-1}}$. Note que o expoente diz para deslocarmos (flutuarmos?) o ponto em uma casa para a esquerda, obtendo 0.1_2 .
- E já que estamos falando de "notação científica binária", esse valor inteiro '1' é implícito em toda codificação de um número em ponto flutuante. A mantissa m contém apenas a parte fracionária do número. Assim, a equação acima fica assim:

$$valor = (1 + m) \cdot 2^{e}$$

- Onde m é um valor no intervalo entre 0 e 1, excluindo o valor 1. Um *float* ou um *double* é codificado assim:

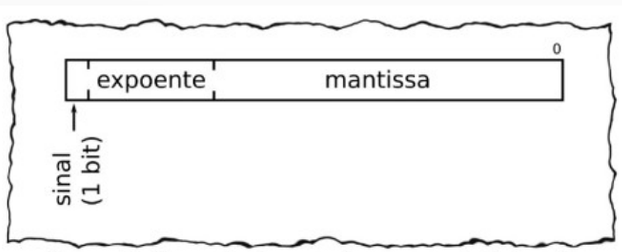

Figura 1: Formato de um valor em ponto flutuante

Ponto flutuante

- A diferença entre os tipos float e double é a quantidade de bits da mantissa e do expoente:

Tipo	Bits do expoente	Bits da mantissa
float	8	23
double	11	52

Table 1: Diferença entre float e double

- O expoente também é um valor binário (óbviamente), mas ele precisa poder conter um valor negativo. Infelizmente esse valor não usa a técnica do complemento 2. No caso de um float o valor de e é calculado assim:

$$e = E - 127$$

- Onde E é o valor do expoente contido num float. Com valores menores que 127 o valor de e torna-se negativo. No caso de um double o expoente tem 11 bits de tamanho e o valor constante é 1023.

Ponto flutuante

- Vimos, acima que 0.5 pode ser escrito, em ***"notação científica binária"*** como $1.0_2 \cdot 2^{-1}$. Assim, na codificação dos 32 bits de um ***float*** teríamos: *0 01111110 00000000000000000000000*$_2$ ou o valor 0x3f000000 (s=0, E=126 e m=0). Isso pode ser fácilmente comprovado com o fragmento de programa abaixo, em C:

```
float f = 0.5f;
unsigned int *p = (unsigned int *)&f;
printf("%#08X\n", *p);
```