

Organização de Computadores I

Ponto Flutuante

Sumário

- Padrão IEEE 754
- Versão 3 bits expoente 4 mantissa, positivos
- Conversão
- Arredondamento
- Adição e Multiplicação

IEEE 754

- Padrão para representação
- 32 bits
 - 1 bit Sinal
 - 8 expoente
 - 23 mantissa $\rightarrow 2^{23} = 2^{10} 2^{10} 2^3 \sim 8 \cdot 10^6$
 - Número de casas decimais = 6 a 7 dígitos = precisão
 - <https://www.h-schmidt.net/FloatConverter/IEEE754.html>
 - Novas tendências - <https://float.exposed/b0x431c>

IEEE 754

- Padrão para representação
- 64 bits
 - 1 bit Sinal
 - 11 expoente
 - 52 mantissa $\rightarrow 2^{52} \sim 4 \cdot 10^{15}$
 - 15 a 16 dígitos em decimal de precisão

Códigos reservados

- Representação de zero
- + infinito e – infinito
- NaN = Not a Number
- Números não normalizados

Valor	Sinal	Expoente	Mantissa
Zero	0	0s	0s
+ Infinito	0	1s	0s
- Infinito	1	1s	0s
NaN	0	1s	Diferente de 0s

No padrão IEEE 754, os NaN (Not a Number), possuem sinal 0, expoente 1 e mantissa com qualquer valor - exceto tudo 0s, pois isso caracteriza infinito- e representam exceções como divisão por zero, raiz de negativos etc.

Normalização

- $4,345 * 10^4$
- $43,45 * 10^3$
- $434,5 * 10^2$
- $4345,0 * 10^1$
- $43450,0 * 10^0$
- Todos representam o mesmo número
- Como padronizar ?

Em binário

- $1,1011 * 2^4$
- $11,011 * 2^3$
- $110,11 * 2^2$
- $1101,1 * 2^1$
- $11011,0 * 2^0$
- Como representar: expoente e mantissa
- Como normalizar ?

IEEE 754 simplificado

- Somente positivos
- Sem códigos especiais: infinito, zero, NaN
- 3 bits de expoente
- 4 bits de mantissa
- Facilitar a compreensão....estender para 32 e 64 bits...
- Normalização, arredondamento, adição e multiplicação

Formato

E

Mant

$$X = 2^{E-3} * (1 + \text{Mant})$$

Formato

E Mant

$$X = 2^{E-3} * (1 + \text{Mant})$$

Representa um número entre 1 e 2
Na mantissa

Formato

E Mant

$$X = 2^{E-3} * (1 + \text{Mant})$$

Mant = potências negativas de 2 ou frações

Formato

E

Mant

$2^{-1} 2^{-2} 2^{-3} 2^{-4}$

$$X = 2^{e-3} * (1 + \text{Mant})$$

Formato

E

Mant

2^{-1} 2^{-2} 2^{-3} 2^{-4}

1 1 1 1

— --- ---- ---

2 4 8 16

$$X = 2^{e-3} * (1 + \text{Mant})$$

Exemplo

E	Mant			
100	0	0	0	0
	2^{-1}	2^{-2}	2^{-3}	2^{-4}

$$X = 2^{e-3} * (1 + \text{Mant})$$

Exemplo

E	Mant			
100	0	0	0	0
	2^{-1}	2^{-2}	2^{-3}	2^{-4}

$$\begin{aligned} X &= 2^{e-3} * (1 + \text{Mant}) \\ &= 2^{4-3} * (1 + 0) = \\ &= 2^1 * 1 = 2 \end{aligned}$$

Exemplo


E	Mant
100	1 0 0 0
	2^{-1} 2^{-2} 2^{-3} 2^{-4}

$$\begin{aligned} X &= 2^{e-3} * (1 + \text{Mant}) \\ &= 2^{4-3} * (1 + 1/2) = \\ &= \end{aligned}$$

Exemplo

E
100

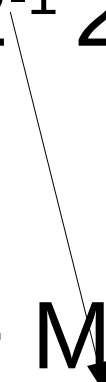
Mant
1 0 0 0
 2^{-1} 2^{-2} 2^{-3} 2^{-4}


$$\begin{aligned} X &= 2^{e-3} * (1 + \text{Mant}) \\ &= 2^{4-3} * (1 + 1/2) = \\ &= \end{aligned}$$

Exemplo

E
100

Mant
1 0 0 0
 2^{-1} 2^{-2} 2^{-3} 2^{-4}

$$\begin{aligned} X &= 2^{e-3} * (1 + \text{Mant}) \\ &= 2^{4-3} * (1 + 1/2) = \\ &= \end{aligned}$$


Exemplo

E	Mant
100	1 0 0 0
	$2^{-1} 2^{-2} 2^{-3} 2^{-4}$

$$\begin{aligned} X &= 2^{e-3} * (1 + \text{Mant}) \\ &= 2^{4-3} * (1 + 1/2) = \\ &= 2^1 * (3 / 2) = 3 \end{aligned}$$

Mudando o expoente

E	Mant
011	1 0 0 0
	2^{-1} 2^{-2} 2^{-3} 2^{-4}

$$\begin{aligned} X &= 2^{e-3} * (1 + \text{Mant}) \\ &= 2^{3-3} * (1 + 1/2) = \\ &= 2^0 * (3 / 2) = 1,5 \end{aligned}$$

Subtrair 1 no
Expoente
Divide por 2

Conversão real \rightarrow binário float

- Mantissa tem que ficar em 1 e 2
- Armazena $1 + M$, então $0 \leq M < 1$
- Três casos
 - Menor que 1
 - Entre 1 e 2
 - Maior que 2

Entre 1 e 2

- O expoente será 0 ou 2^0
- No formato $2^{e-3} = 2^{3-3}$, ou seja, $e = 011$
- Mantissa = $1 + M$
- M é formato por soma de fração $1/2$, $1/4$, $1/8$, $1/16$
- Em real frações = 0,5 0,25 0,125 0,0625
- Pesar como uma balança

Exemplo $x = 1,6$

- Entre 1 e 2, então expoente = 0 ou 2^{3-3}
- $1+M = 1,6 \rightarrow M = 0,6$

Exemplo $x = 1,6$

- Entre 1 e 2, então expoente = 0 ou 2^{3-3}
- $1+M = 1,6 \rightarrow M = 0,6$
- Pesos
 - 0,5 ok
 - 0,25 $\rightarrow 0,5 + 0,25 = 0,75 > 0,6$

Exemplo $x = 1,6$

- Entre 1 e 2, então expoente = 0 ou 2^{3-3}
- $1+M = 1,6 \rightarrow M = 0,6$
- Pesos
 - 0,5 ok
 - 0,25 $\rightarrow 0,5 + 0,25 = 0,75 > 0,6$
 - 0,125 $\rightarrow 0,5 + 0,125 = 0,625 > 0,6$
 - 0,0625 $\rightarrow 0,5 + 0,0625 = 0,5625 < 0,6$

Exemplo $x = 1,6$

- Entre 1 e 2, então expoente = 0 ou 2^{3-3}
 - $1+M = 1,6 \rightarrow M = 0,6$
 - Pesos
 - 0,5 ok
 - 0,25 $\rightarrow 0,5 + 0,25 = 0,75 > 0,6$
 - 0,125 $\rightarrow 0,5 + 0,125 = \mathbf{0,625 > 0,6} \leftarrow \mathbf{Mais Pr\u00f3ximo}$
 - 0,0625 $\rightarrow 0,5 + 0,0625 = 0,5625 < 0,6$
 - Em bin\u00e1rio $M = 1010$
- 011 1010 em 7 bits = $2^{3-3} * (1 + \frac{1}{2} + \frac{1}{8})$

$$X > 2$$

- Primeiro, dividir até ficar entre 1 e 2
- O número de divisões será o expoente positivo
- Armazenar no formato 2^{e-3}
- Calcular a mantissa, lembrar que $1+M$

Exemplo $x=2,75$

- $2,75 > 2$
- $2,75 / 2 = 1,375$
- $2^1 * 1,375 = 2,75$

Exemplo $x=2,75$

- $2,75 > 2$
- $2,75 / 2 = 1,375$
- $2^1 * 1,375 = 2,75$
- Expoente $2^{4-3} = 2^1$
- $1+M = 1,375 \rightarrow M = 0,375$

Exemplo $x=2,75$

- $2,75 > 2$
- $2,75 / 2 = 1,375$
- $2^1 * 1,375 = 2,75$
- Expoente $2^{4-3} = 2^1$
- $1+M = 1,375 \rightarrow M = 0,375$
- Pesos
 - 0,5 – muito
 - $M = 0,25 + 0,125 = 0,375$ exato !

Exemplo $x=2,75$

- $2,75 > 2$
- $2,75 / 2 = 1,375$
- $2^1 * 1,375 = 2,75$
- Expoente $2^{4-3} = 2^1$
- $1+M = 1,375 \rightarrow M = 0,375$
- Pesos
 - 0,5 – muito
 - $M = 0,25 + 0,125 = 0,375$ exato !
- $X = 100 \quad 0110 = 2^{4-3} * (1 + \frac{1}{4} + \frac{1}{8})$
 $= 2^1 * \frac{11}{8} = \frac{11}{4} = 2,75$

Menor que 1

- Multiplicar por 2 até ficar entre 1 e 2
- O número de multiplicações é o expoente negativo
- Gerar a mantissa $1+M$

Exemplo $X = 0,3$

- $0,3 * 2 = 0,6$
- $0,6 * 2 = 1,2$
- Então $2^{-2} * 1,2 = 0,3$ ou $2^{1-3} * (1 + 0,2)$

Exemplo $X = 0,3$

- $0,3 * 2 = 0,6$
- $0,6 * 2 = 1,2$
- Então $2^{-2} * 1,2 = 0,3$ ou $2^{1-3} * (1 + 0,2)$
- $0,2$
 - $0,5$ muito
 - $0,25$
 - Ou $0,125 + 0,0625 = 0,1875 \leftarrow$ mais próximo

Exemplo $X = 0,3$

- $0,3 * 2 = 0,6$
- $0,6 * 2 = 1,2$
- Então $2^{-2} * 1,2 = 0,3$ ou $2^{1-3} * (1 + 0,2)$
- $0,2$
 - $0,5$ muito
 - $0,25$
 - Ou $0,125 + 0,0625 = 0,1875 \leftarrow$ mais próximo
- $X = 001 \quad 0011 = 2^{1-3} * (1 + 1/8 + 1/16) =$
 $= 2^{-2} * 19/16 = 19/64 = 0,296875$

Formato

Soma

- Somar em ponto flutuante
- Em decimal
 - $10^2 * 1,23$
 - $10^{-2} * 1,45$

?

Ajustar expoentes

- Somar em ponto flutuante
- Em decimal

- $10^2 * 1,23$

- $10^2 * 0,000145$

1,230145

Arredondar para 2 digitos

- Somar em ponto flutuante
- Em decimal
 - $10^2 * 1,23$
 - $10^2 * 0,000145$

1,23