Ponto Flutuante

1980 com pesquisa de William Kahan, a Intel lançou uma solução para tratar números em ponto flutuante (proc matematico 8087)

Seguindo a mesma linha de estudo o IEEE em 1985, lançou um documento que serviria de referência para padronizar o tratamento destes números

A solução proposta passa pela normalização dos números, seguindo o padrão:
$+/- 1, m_1 m_2 m_3 \ldots \times 2^{E}$

Exemplo1:

$5,25_{(10)} = 101,01_{(2)}$

5 (vai pelas divisões sucessivas)

0,25 (vai pelas multiplicações sucessivas)
0,25 * 2 = 0,50 * 2 = 1,00 (01)

Normalizado, fica

$1,0101_{(2)} \times 2^{2}$ (trouxe duas casas para a esquerda, entao aumenta o expoente)

Exemplo2:

$0,375_{(10)} = 0,011_{(2)}$

0,375 (vai pelas multiplicações sucessivas)
0,375 * 2 = 0,75 * 2 = 1,50 (despreza o 1) * 2 = 1,00 (011)

Normalizado, fica

$1,1_{(2)} \times 2^{-2}$ (trouxe duas casas para a direita, entao diminui o expoente)
(explicar de onde veio o 2^{-2})

Convertendo binario com vírgulas para decimal

Vimos que a parte inteira se obtem por: digito x base ^{posicao}, como o expoente é negativa, causa o efeito da divisão, ou seja, 2^{-1}, é igual a ½, ou seja 0,5

4	2	1	0,5	0,25	0,125	0,0625	0,03125
2^2	2^1	2^0	2^{-1}	2^{-2}	2^{-3}	2^{-4}	2^{-5}

Assim sendo, 0,0011 ficaria

0,125 + 0,0625 = 0,1875

Deste modo, 0,2 ficaria melhor representado por 0,00110011 x 2^0 = 1,10011 x 2^{-3} = 0,19921875

O Padrão IEEE 754, tem a seguinte definição

1) Converter em binario
2) Normalizar o numero
3) Falaremos da precisão simples (32) e dupla (64)
4) Notar que um digito é suprimido devido a padrão

Sinal	E + BIAS	m_1 m_2 m_3 …
	Caracteristica	Significando ou Mantissa

BIAS = representação de excesso (viés = tendência)

Para precisão simples (32 bits)

No exemplo2:

$$+0{,}375 = +1{,}1_2 \cdot 2^{-2}$$

$$E + 127 = 125 = 111\,1101_2$$

0	0111 1101	100 0000 0000 0000 0000 0000

No exemplo1:

$$-5{,}25 = -1{,}0101_2 \cdot 2^2$$

$$E + 127 = 129 = 1000\ 0001_2$$

1	1000 0001	010 1000 0000 0000 0000 0000

Exemplo (precisão simples)

- Valor

  `float F = 15213.0;`

  15213_{10} = 11101101101101_2 = 1.1101101101101_2 X 2^{13}

- Mantissa

  M = **1.1101101101101_2**

  `frac` = $11011011011010000000000_2$

- Expoente

  E = 13

  exp = E + Bias = 13+ 127 = 140 = 10001100_2

Sinal:	**0**							
Hex:	**4**	**6**	**6**	**D**	**B**	**4**	**0**	**0**
Binário	**0100**	**0110**	**0110**	**1101**	**1011**	**0100**	**0000**	**0000**
140:		**100**	**0110**	**0**				
15213:				**1110**	**1101**	**1011**	**01**	

Para precisão Dupla (64 bits)

$$\pm 1, m_1 m_2 \ldots m_{52} \cdot 2^E$$

1 bit	11 bits	52 bits
SINAL	$E + 1023$	$m_1 m_2 \ldots m_{52}$

Exemplo (precisão dupla)

- Valor

  `double D = 178.125 = 128+32+16+2 + 0.125;`

  $178.125_{10} = 10110010.001_2$

  $1.78125_{10} = 1.0110010001_2 \times 2^7$

- Mantissa

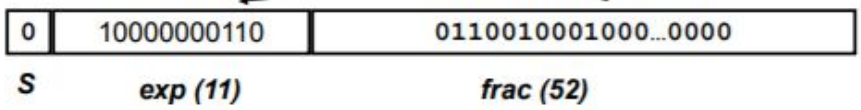
  M **=** **1.0110010001_2**

  `frac` = $0110010001 00\ldots0_2$

- Expoente

  E = 7

  exp = E + Bias = 7+ 1023 = 1030 = 10000000110_2

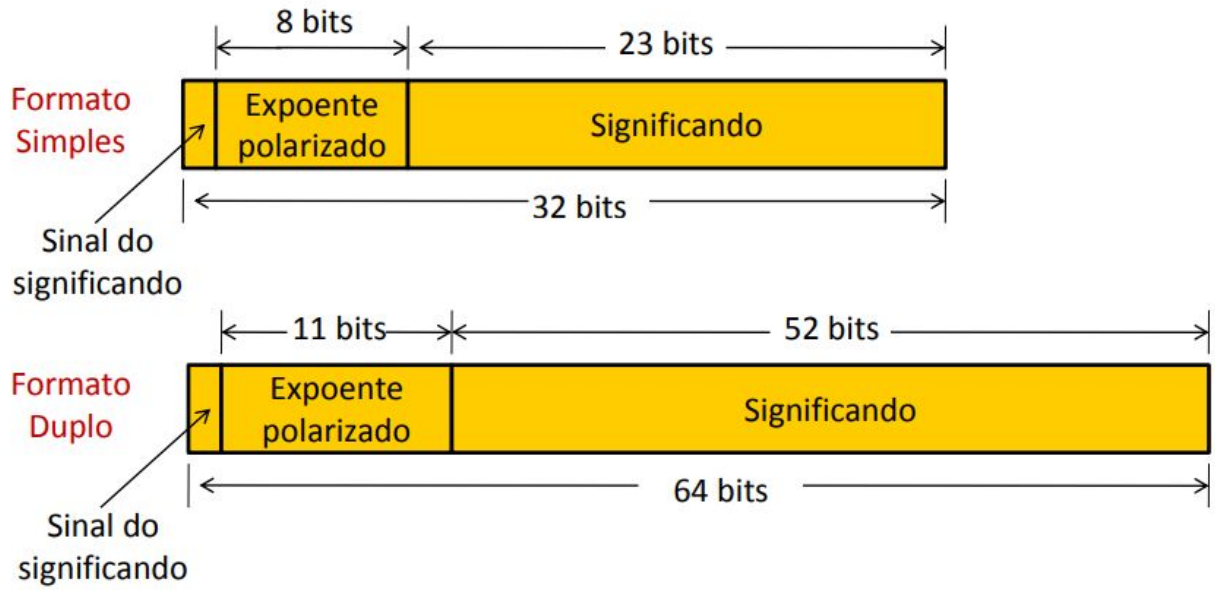


S ***exp (11)*** ***frac (52)***

IEEE 754: Precisões de Ponto Flutuante

- Tamanhos
  - **`float`**: `exp` = 8 bits, `frac` = 23 bits, `s` = 1 bit
    - Total: 32 bits
    - Faixa de valores: 2^{-126} até 2^{127}
  - **`double`**: `exp` =11 bits, `frac` = 52 bits, `s` = 1 bit
    - Total: 64 bits
    - Faixa de valores: 2^{-1022} até 2^{1023}
  - Precisão estendida: `exp` =15 bits, `frac` = 63 bits,`s` = 1 bit
    - Total: 80 bits
    - Faixa de valores: 2^{-16382} até 2^{16383}
    - 1 bit é desperdiçado

Resumão do IEEE 754



- Parâmetros do formato IEEE 754

Parâmetro	Formato Simples	Formato Duplo
Tamanho da palavra	32	64
Tamanho do expoente	8	11
Polarização do expoente	127	1023
Expoente máximo	127	1023
Expoente mínimo	-126	-1022
Tamanho da mantissa	23	52
Número de expoentes	254	2046
Número de mantissas	2^{23}	2^{52}
Número de valores	$1,98 \times 2^{31}$	$1,99 \times 2^{63}$

- Valores especiais definidos no IEEE 754

Sinal	Expoente Polarizado		Mantissa	Valor
	Formato Simples	Formato Duplo		
0	0	0	0	0
1	0	0	0	-0
0	255	2047	0	∞
1	255	2047	0	$-\infty$
0 ou 1	255	2047	$\neq 0$	NaN