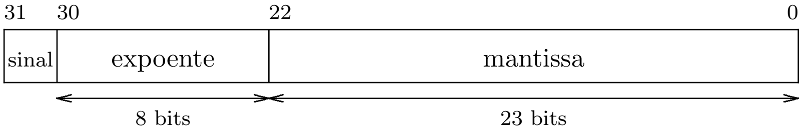
Representação posicional

$34.567_{10} = 3 \cdot 10 + 4 \cdot 1 + 5 \cdot 0.1 + 6 \cdot 0.01 + 7 \cdot 0.001$
$101.1001_2 = 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 + 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 0 \cdot 2^{-3} + 1 \cdot 2^{-4}$
$= 1 \cdot 4 + 0 \cdot 2 + 1 \cdot 1 + 1 \cdot 0.5 + 0 \cdot 0.25 + 1 \cdot 0.125 + 1 \cdot 0.0625$

Representação em ponto flutuante



e bits de *expoente*,
m bits de *mantissa* (fração)
$F = M \cdot \beta^E$
para mantissa M, exp E, base β

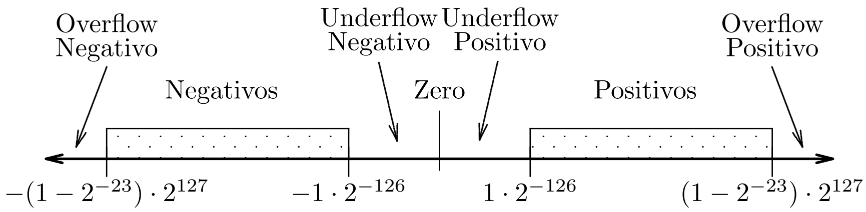
$F = (-1)^{sinal} \cdot (m_1 \cdot 2^{-1} + m_2 \cdot 2^{-2} + m_3 \cdot 2^{-3} + \dots) \cdot 2^E$

menor número: $\approx 2.0 \cdot 10^{-38}$
maior número: $\approx 2.0 \cdot 10^{+38}$
$\rightarrow$ precisão é menor que em ponto fixo

|expoente| $\rightsquigarrow$ faixa de representação
|fração| $\rightsquigarrow$ precisão na representação $0 <$ fração < 1
faixa enorme representada por 2^{32} padrões $\neq$s

Número é *normalizado* se não há 0s a direita do ponto binário
normalização: desloca fração para esquerda (aumentando precisão)
enquanto decrementa expoente: $0.\mathbf{00}101 \cdot 2^3 \overset{\text{norm}}{\leftrightarrow} 0.10100 \cdot 2^1$
Exemplo: $-0.75_{10} = -3/4 = -3/2^2 = 11.0_2/2^2 = -0.11_2 = -0.11 * 2^0$

Faixa dos PF positivos:
$M_{\min} \cdot 2^{E_{\min}} \leq F^+ \leq M_{\max} \cdot 2^{E_{\max}}$
$|F^+| = |F^-|$



overflow: expoente muito grande para representação $> +127$
underflow: expoente muito pequeno para representação < -126

Padrão IEEE 754 Padrão universal para representação em ponto flutuante. O primeiro dígito significativo do *significando* é implícito, à esq do ponto: `s eeee eeee 1.mmmm mmmm mmmm mmmm mmmm mmmm`

	sinal	exp	fraç
float	1	8	23
double	1	11	52

fração $\in [1, 2)$

Formato: $(-1)^s \cdot (1 + \text{fração}) \cdot 2^{(E-\text{deslocamento})}$
$(-1)^s \cdot (1 + m_1 \cdot 2^{-1} + m_2 \cdot 2^{-2} + m_3 \cdot 2^{-3} + \dots) \cdot 2^{(E-\text{desloc})}$
onde *deslocamento* é 127 ou 1023

Com expoente deslocado, número menor tem expoente menor; pode-se comparar floats e doubles com instruções para inteiros: $\rightarrow$ `beq` e `slt`

Zero é caso especial: expoente e mantissa são todos 0
A fração mais o **1** implícito é chamada de *significando*

Parâmetro IEEE 754	*float*	*double*
bits de precisão	24	53
Expoente máximo $E_{\max}$	127	1023
Expoente mínimo $E_{\min}$	-126	-1022
Deslocamento no exp.	127	1023

Exemplo 1: $-0.75_{10} = -3/4 = -3/2^2 = 11.0_2/2^2 = -0.11_2 = -0.11 * 2^0 \overset{\text{norm}}{\leftrightarrow} -1.1 * 2^{-1}$
representado em float: $(-1)^s \cdot (1 + \text{mantissa}) \cdot 2^{(\text{expoente}-127)} = (-1)^1 \cdot (1 + 0.1000\dots000) \cdot 2^{(126-127)}$
`1  0111 1110  1000 0000 0000 0000 0000 000`

Exemplo 2: $0.5_{10} = 0.1_2 \overset{\text{norm}}{\leftrightarrow} 1.0 \cdot 2^{-1} = (-1)^0 \cdot (1 + 0.0000\dots000) \cdot 2^{(126-127)}$
`0  0111 1110  0000 0000 0000 0000 0000 000`

Exemplo 3: $1.0_{10} = 1.0_2 \overset{\text{norm}}{\leftrightarrow} 1.0 \cdot 2^0 = (-1)^0 \cdot (1 + 0.0000\dots000) \cdot 2^{(127-127)}$
`0  0111 1111  0000 0000 0000 0000 0000 000`

Exemplo 4: $2.0_{10} = 10.0_2 \overset{\text{norm}}{\leftrightarrow} 1.0 \cdot 2^1 = (-1)^0 \cdot (1 + 0.0000\dots000) \cdot 2^{(128-127)}$
`0  1000 0000  0000 0000 0000 0000 0000 000`

Exemplo 5: sinal=1, expoente=129, mantissa=0100...000
`1  1000 0001  0100 0000 0000 0000 0000 000`
$(-1)^s \cdot (1 + \text{mantissa}) \cdot 2^{(\text{expoente}-127)} = (-1)^1 \cdot (1 + 0.0100\dots000) \cdot 2^{(129-127)} = -1 \cdot (1 + 0.25) \cdot 2^2 = -5.0_{10}$

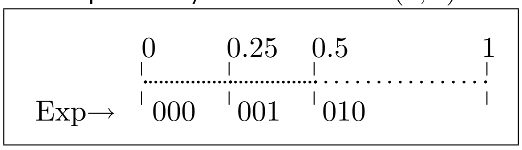
Representação de Ponto Flutuante em 8 bits Existem 256 valores diferentes que podem ser representados em ponto flutuante (PF) com 8 bits, e estes valores não são distribuídos uniformemente na reta dos Reais. A representação PF-8 é uma versão (muito) reduzida do Padrão IEEE 754.

Nesta representação, há seis intervalos (expoentes 001..110) com 16 pontos em cada intervalo (0000..1111). O intervalo com expoente 111 é usado para representar infinito e NaN (*not a number*). O intervalo com expoente 000 é dito denormalizado. O deslocamento do expoente é 3.

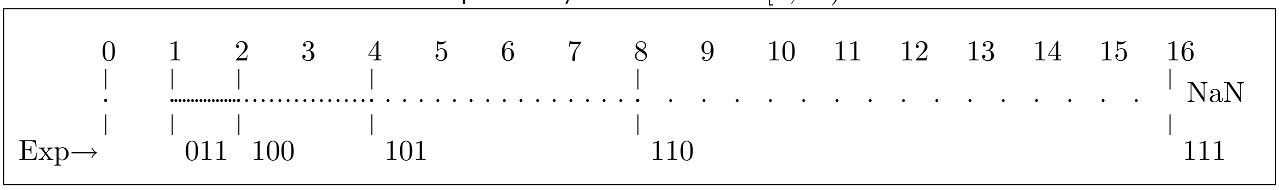
s	exp	fração
1	1 0 0	1 0 0 1

As figuras abaixo mostram a representação dos números reais em 8 bits. Note que as figuras mostram somente os Reais positivos, e que o intervalo com expoente 000 é denormalizado.

representação no intervalo $(0, 1)$



representação no intervalo $[1, 16)$



1) Preencha a tabela abaixo com os números representáveis em PF-8. Evidentemente, não é necessário representar todas as 256 possibilidades, embora os casos limites devam todos ser preenchidos. O campo 'expoente' deve conter o expoente deslocado (com o *bias* de 3). As 'magnitudes' são os números representados, o '*gap*' é o intervalo não-representável (vazio) entre dois números vizinhos na representação. Não esqueça das representações para $\pm\infty$ e NaN.

sinal	expoente	fração	núm $\beta = 2$	núm $\beta = 10$	gap	comentário
0/1	000	0000	0.000	zero		caso especial

2) Quais são os piores erros de arredondamento quando se faz cálculos nas faixas $[-8, 8]$ e $[-1, 1]$?

3) Prove que as propriedades aritméticas abaixo são válidas, ou dê um contra-exemplo. $\oplus$ e $\otimes$ são a adição e o produto de números representados em PF-8. Evidentemente, os casos de overflow e underflow não podem ser usados como contra-exemplos. *With thanks to Jeff Sanders!*

a) monotonicidade c.r.a soma: $x \leq y \Rightarrow x \oplus z \leq y \oplus z$

b) associatividade c.r.a multiplicação: $x \otimes (y \otimes z) = (x \otimes y) \otimes z$

c) distributividade: $x \otimes (y \oplus z) = (x \otimes y) \oplus (x \otimes z)$

4) Represente as potências de 2 entre 4 e 1024 no formato `float` IEEE 754.