

TPC2

Resultados dos exercícios propostos:

- (A) PEQUENO1: $V = (-1)^S * 1.F * 2^{E-7}$ (normalizado; e desnormalizado?)
 PEQUENO2: $V = (-1)^S * 1.F * 2^{E-3}$ (normalizado; e desnormalizado?)
- (A) Para ambos os formatos, apresente os seguintes valores em decimal:
 - O maior finito positivo: PEQ1 240 (0 1110 111) PEQ2 31/2 (0 110 1111)
 - O negativo normalizado +próx. 0 PEQ1 -1/64 (1 0001 000) PEQ2 -1/4 (1 001 0000)
 - O > n° positivo desnormalizado PEQ1 7/512 (0 0000 111) PEQ2 15/64 (0 000 1111)
 - O positivo desnormaliz +próx. 0 PEQ1 1/512 (0 0000 001) PEQ2 1/64 (0 000 0001)
- (A) Calcule os valores correspondentes ao formato PEQUENO1 (modelo de resposta em a)):
 - 10110011 Res: Valor normalizado, logo $V = (-1)^1 * 1.011_2 * 2^{-1} = -0.1011_2$
 - 01111010 Res: NaN (Não é um número real)
 - 10010001 Res: Valor normalizado, logo $V = (-1)^1 * 1.001_2 * 2^{-5} = -0.00001001_2$
 - 00000011 Res: Valor desnormalizado, logo $V = (-1)^0 * 0.011_2 * 2^{-6} = +0.00000011_2$
 - 11000001 Res: Valor normalizado, logo $V = (-1)^1 * 1.001_2 * 2^1 = -10.01_2$
- (R) Codifique os seguintes valores como números em vírgula flutuante no formato PEQUENO1
 - 111.01₃ Res: 1 1010 101 -> (-)1.101(000111)₂*2³, 3=E-7 -> E=10
 - 1/8 K Res: 0 1110 000 -> (+)1.0*2⁷, 7=E-7 -> E=14
 - 0x18C Res: 1 1111 000 -> (-)1.10001100₂*2⁸, 8=E-7 -> E=15 (-infinito)
 - 110.01 Res: 0 1101 101 -> (+)1.1011100...₂*2⁶, 6=E-7 -> E=13
 Res_a: 0 1101 110 -> Nota: Res_t (truncado), Res_a (arredondado, o correto)
 - 0.005₈ Res: 0 0000 101 -> (+)1.01₂*2⁻⁷, -7=E-7 -> E=0 (exceção: desnorm)
 -> (+)0.101₂*2⁻⁶
- (B) Converta os seguintes números PEQUENO1 em números PEQUENO2:

Limites para o campo E e para o expoente (normalizado à esquerda, e desnormalizado à direita):

PEQ1: E->[1,14], Exp->[-6,7] Exp=-6, F->]1,2⁻³], V->]2⁻⁶,2⁻⁹]
 PEQ2: E->[1,6], Exp->[-2,3] Exp=-2, F->]1,2⁻⁴], V->]2⁻²,2⁻⁶]

 - PEQ1: 0 0110 011 -> Exp=(6-7)=-1 PEQ2: -1= E-3, E= +2 -> 0 010 0110
 - PEQ1: 1 1101 001 -> Exp=(13-7)=6 PEQ2: Exp=+6 -> overflow -> 1 111 0000
 - PEQ1: 0 0010 000 -> Exp=(2-7)=-5 PEQ2: Exp=-5 -> desnorm -> 0 000 0010
 - PEQ1: 1 1001 110 -> Exp=(9-7)=+2 PEQ2: +2= E-3, E= +5 -> 1 101 1100
 - PEQ1: 1 0000 010 -> desnorm <2⁻⁶ PEQ2: Exp=-2 e F_a<2⁻⁴ -> underflow -> -0
 -> 1 000 0000

6. (B) É viável garantir 8 algarismos significativos na representação de variáveis do tipo `float`?

Para garantir que qualquer valor do tipo `real` em precisão simples tenha sempre pelo menos 8 algarismos significativos (na base 10), seria necessário que a sua codificação em binário (em IEEE 754) permitisse representar pelo menos 10^8 valores diferentes.

Sabendo que a codificação em precisão simples usa 32 bits, dos quais 23+1 (o "bit escondido" 1 na notação normalizada, 0 na desnormalizada) são usados para a mantissa, então com esta notação apenas se podem representar 2^{24} valores diferentes, $\sim 16 \cdot 10^6$, o que é claramente inferior a 10^8 .

7. (B) Qual o maior inteiro ímpar que é possível representar exatamente, como `float`, na norma IEEE 754?

Sendo este valor representável como normalizado, a parte fracionária deverá ser o máximo valor permitido (tudo 1s) e a máxima potência de 2 que se poderá usar no expoente deverá ser tal que desloque o ponto decimal 23 casas para a direita a partir do bit escondido, garantindo que o número resultante é um inteiro e não tem um zero no algarismo mais à direita, i.e., este valor é o $2^{24}-1$