Zad 3.

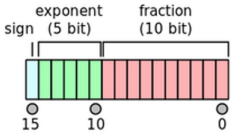
sobota, 18 marca 2023 10:26

Zadanie 3. Standard IEEE 754-2008 definiuje liczby zmiennopozycyjne o szerokości 16-bitów. Zapisz ciąg bitów reprezentujący liczbę $1.5625 \cdot 10^{-1}$. Porównaj zakres liczbowy i dokładność w stosunku do liczb zmiennopozycyjnych pojedynczej precyzji (`float`).

The IEEE 754 standard[9] specifies a **binary16** as having the following format:

- Sign bit: 1 bit
- Exponent width: 5 bits
- Significand precision: 11 bits (10 explicitly stored)

The format is laid out as follows:



zakres float16
max: $6,55 \cdot 10^{4}$
min(>0): $6,10 \cdot 10^{-5}$

float32
max: $3,4028 \cdot 10^{38}$
min: $1,1754 \cdot 10^{-38}$ znormalizowana

$1,5625 \cdot 10^{-1} = 0,15625_{10}$

$0,15625 \times 2 = 0,3125$ → 0
$0,3125 \times 2 = 0,625$ → 0
$0,625 \times 2 = 1,25$ → 1
$0,25 \times 2 = 0,5$ → 0
$0,5 \times 2 = 1$ → 1

} mantysa

$0,15625_{10} = 0.00101_{2} \times 2^{0} = 1.01 \cdot 2^{-3}$

$m = 0100000000$
$e = -3+15 = 12 = 01100$

The half-precision binary floating-point exponent is encoded using an offset-binary representation, with the zero offset being 15; also known as exponent bias in the IEEE 754 standard.

$s = 0$

$1,5625_{10} = \underbrace{0}_{s}\ \underbrace{01100}_{e}\ \underbrace{0100000000}_{m}{}_{2}$