Zadanie 2

March 9, 2021

Autor: Mateusz Chojnowski

Zadania:

- 1. Wyznaczyć najmniejszą dodatnią liczbę zmiennopozycyjną
- 2. Wyznaczyć najmniejszą dodatnią liczbę zmiennopozycyjną o znormalizowanej mantysie
- 3. Do powyższego zadania można wykorzystać jednostkę zaokrąglenia zdefiniowaną jako najmniejszą liczbę taką, że fl(1+e)>1
- 4. Wyznaczyć największą możliwą liczbę zmiennopozycyjną (mniejszą od ∞)
- 5. Obliczenia wykonać w pojedynczej i podwójnej precyzji (czy na pewno wykonano obliczenia w pojedynczej precyzji?)
- 6. Na wyjściu program wypisuje (trzy) wartości ekstremalne wyznaczone odpowiednio dla pojedynczej i podwójnej precyzji
- 7. Wyniki odnieść do wartości L, U, poziomu niedomiaru (underflow) i nadmiaru (overflow) zawartych w standardzie IEEE 754

1 Program

Na potrzebę zadania został przygotowany program w języku C. Kod programu znajduje się w pliku main.c załączonym do pracy.

2 Realizacja

2.1 Wyznaczyć najmniejszą dodatnią liczbę zmiennopozycyjną

najmniejszych dodatnich liczb zmiennopozycyjnych wyznaczenia użyta funkcja calculate_lowest_positive_float oraz calculate_lowest_positive_double. liczby przypadku najmniejsza możliwa typu float liczba wynosi wynik w takiej postaci). Jest on równy 2^{-149} co pokrywa się z postacią typu w standardzie IEEE 754 - 2^7 (jeden bit zarezerwowany na znak wykładnika) = 128 + 21 bitów z mantysy użytych do przechowania nadmiarowego wykładnika co daje w sumie 149.

Dla liczby typu double uzyskano liczbę o wartości około 2e-324. Alternatywną drogą do uzyskania minimalnej liczby double wykonania potęgowania $2^{-(1024+51)}$ ($1024 = 2^10$ - wykładnik, 51 - bity mantysy wykorzystane do przechowania wykładnika).

Należy zauważyć że wchodząc wykładnikiem w mantyse (łagodny niedomiar) stopniowo tracimy precyzje zawartą w najmniej znaczących bitach mantysy, sposowodwane jest to przesunięciem liczby w mantysie, aby zrobić miejsce na bity niosące informacje o wykładniku.

2.2 Wyznaczyć najmniejsz dodatnią liczbę zmiennopozycyjną o znormalizowanej mantysie

Do wyznaczenia najmniejszych dodatnich liczb zmiennopozycyjnych została użyta funkcja calculate_min_positive_float oraz calculate min positive double. W znormalizowanej formie najmniejsza wartość float okreslona w standardzie IEEE 754 (w kodzie zawarta jest asercja porównująca uzyskaną wartość ze stała).

W znormalizowanej formie najmniejsz wartość double około 5e-324, pokrywa się ona z wartością określoną w standardzie IEEE 754 (w kodzie zawarta jest asercja porównująca uzyskaną wartość ze stałą).

3 Wyznaczyć jednostkę zaokrąglenia e_{mach}

Do wyznaczenia jednostki zaokrąglenia e_{mach} dla typu float oraz double zostały użyte funkcje calculate_emach oraz calculate_eamch_double. Poszczególne kroki wyznaczania wartości zostały zapiane do plików załączonych do raportu (emach_double.csv oraz emach_float.csv).

Uzyskane wartości, 0.00000011920928955078125 dla typu float oraz 0.000000000000022204460492503131 dla typu double, pokrywają się z wartościami zawartymi w standardzie (w kodzie znajdują się odpowiednie asercje).

4 Wyznaczyć największą możliwą liczbę zmiennopozycyjną (mniejszą od ∞)

Do wyznaczenia maksymalnych wartości float i double zostały użyte funkcje calculate_highest_float oraz calculate_highest_double. Funkcje te najpierw wypełniają mantysę typu 1, a następnie multiplikują uzyskaną liczbę o 2 do czasu osiągnięcia wartości maksymalnej mniejszej od ∞ .