

Reprezentacja zmiennoprzecinkowa

Dariusz Borkowski

N

Wówczas będzie zadana liczba naturalna b . Długość liczby rzeczywistej $x \neq 0$ można przedstawić w postaci

$$x = s \cdot 2^{c-b} \cdot m,$$

gdzie $s \in \{-1, 1\}$ jest znakiem, liczba całkowita $(c - b)$ cechą, a liczba rzeczywista $m \in [1, 2)$ mantysą liczby x .

Twierdzenie 1. *Powyższy rozkład jest jednoznaczny.*

W komputerach osobistych mamy do czynienia z reprezentacją liczb rzeczywistych, w której do zapisania liczby używa się ściśle określonej liczby bitów t do zapisania mantysy i także określonej liczby bitów p do zapisania cechy danej liczby niezerowej x :

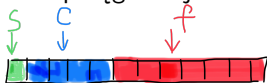
$$s\ c_1c_2 \dots c_p\ f_1f_2 \dots f_t$$

(łącznie $1 + p + t$ bitów). Liczby zapisane przy użyciu powyższej sekwencji bitów nazywa się liczbami maszynowymi. Są to jedyne dokładnie zapisywalne w komputerze liczby rzeczywiste, pozostałe będą musiały zostać wyrażone z wykorzystaniem liczb maszynowych.

Reprezentacją zmiennoprzecinkową niezerowej liczby x będziemy nazywać liczbę $rd_\nu(x)$ taką, że

$$rd_\nu(x) = (-1)^s \cdot (1 + f) \cdot 2^{c-b},$$

gdzie f jest liczbą dwójkową postaci $(0.f_1 \dots f_t)_2$, natomiast c jest liczbą naturalną postaci $(c_1 \dots c_p)_2$. Na znak liczby, s , przeznaczony jest jeden bit. Wartości c i f dobiera się tak, żeby $rd_\nu(x)$ była tak bliska x jak to możliwe. Stałą całkowitą b dobiera się tak, by uzyskać zbalansowany zakres cechy $c-b$ (mniej więcej tyle samo wartości ujemnych i dodatnich), a zysk z korzystania z niej jest taki, że nie marnujemy dodatkowego bitu na przechowywanie znaku wykładnika potęgi dwójki $c-b$.



Niech symbol $\circ$ oznacza jedno z czterech działań arytmetycznych. Niech x, y będą liczbami maszynowymi. Mamy obliczyć $x \circ y$. Przyjmujemy, że komputer działa tak, że po wykonaniu tego działania mantysa wyniku jest normalizowana (tzn. sprowadzana do właściwego przedziału, np. $[1, 2)$) i zaokrąglana, a cecha odpowiednio zmieniana.

Typ IEEE 754	Pojedynczej precyzji
Nazwa typu w C	float
Liczba bajtów dla typu w C	4
Liczba bitów cechy	8
Liczba bitów mantysy	23
Bias (liczba b powyżej)	127

Typ IEEE 754	Podwójnej precyzji
Nazwa typu w C	double
Liczba bajtów dla typu w C	8
Liczba bitów cechy	11
Liczba bitów mantysy	52
Bias (liczba b powyżej)	1023

Zadanie 1. *Napisz program w C++, który dla danej liczby float wypisuje jej zapis binarny w pamięci.*

Rozwiązanie: *bin.cpp*



Zadanie 2. *Dany jest binarny zapis pamięci liczby rzeczywistej typu float. Jaka to liczba ?*

a) 01000001001000000000000000000000

b) 11000000100000000000000000000000

c) 00111101110011001100110011001101

Zadanie 3. *Przedstaw w postaci binarnego zapisu pamięci dla typu float liczby:*

a) -2.5

b) 0.1

c) 1

Zadanie 4. *Przetestuj poniższy skrypt. Wyjaśnij otrzymany wynik.*

```
x=0;  
while x~=1  
    x=x+0.1  
end
```

Zadanie 5. *Przetestuj poniższy skrypt. Wyjaśnij otrzymany wynik.*

```
x=0;  
while x~=1  
    x=x+0.00390625  
end
```

Zadanie 6. *Przetestuj poniższy skrypt. Wyjaśnij otrzymany wynik.*

```
x=1;  
while x~=0  
    x=x/3  
end
```

Zadanie 7. *Przetestuj poniższy skrypt. Wyjaśnij otrzymany wynik.*

```
x=1;  
while x~=0  
    x=x*2/3  
end
```

Zadanie 8. *Przetestuj poniższy skrypt dla $n = -1, -2, \dots, -20$. Wyjaśnij otrzymany wynik.*

```
x=10^n;  
y=((1+x)-1)/x
```