УДК 519.7

Л. Мороз, А. Гринчишин, Я. Мірецька
Національний університет "Львівська політехніка",
кафедра безпеки інформаційних технологій

ПРОСТІ АЛГОРИТМИ ДІЛЕННЯ ЧИСЕЛ З ПЛАВАЮЧОЮ КОМОЮ

Подано опис простих алгоритмів швидкого ділення з використанням магічної константи зі зменшеними відносними похибками обчислень для чисел типу float.

Ключові слова: магічна константа, числа типу float, стандарт IEEE-754, відносна похибка обчислень, ділення чисел з плаваючою комою.

Description of simple fast division algorithms using magic constants with reduced errors in calculations for the float numbers has been provided

Key words: magic constant, floating-point numbers, standard IEEE-754, fast division.

Вступ

Ділення чисел з плаваючою комою є однією із чотирьох найуживаніших елементарних операцій [1–5].

У [4] описано теорію та практичну реалізацію швидких алгоритмів для обчислення оберненопропорційної залежності (англ. – *reciprocal*) $y_t = \frac{1}{x}$:

```
float reciprocal (float x)
{
int i = *(int*)&x; { перевід числа x з floating-point у integer }
i = 0x7ef311c3 -i; );{ціле початкове наближення для reciprocal, де 0x7ef311c3 – магічна константа }
float y = *(float*)&i; { перевід i з integer у floating-point для отримання початкового наближення}
y= y*(2.0f -x*y); {перша Ньютонівська ітерація}
y = y*(2.0f - x*y); {друга Ньютонівська ітерація}
return y;
}
```

Цей алгоритм має найнижчу точність – максимальна відносна похибка результатів обчислень y_t для усього діапазону чисел з плаваючою комою (режим single precision для стандарту IEEE-754 або для чисел типу float у мові С) на рівні $6.51 \cdot 10^{-4}$ %, що становить 17.2 коректних бітів результату.

Вищу точність має такий алгоритм:

```
float reciprocal (float x)
{
int i = *(int*)&x;
i = 0x7ef311c3 -i;
float y = *(float*)&i;
y = y*( 2.00130856f -x*y);
y = y*( 2.00000084 f - x*y);
return y;
        }
```

Його максимальна відносна похибка має приблизно $1.01 \cdot 10^{-4}\%$, що становить 19.9 коректних бітів результату для усього діапазону чисел з плаваючою комою, поданих у форматі одинарної точності (single precision) стандарту IEEE-754. Обидва алгоритми мають у своєму складі лише по чотири операції множення.

Вищенаведені алгоритми застосовуються тільки для обчислення обернено-пропорційної залежності та мають порівняно великі похибки – у цьому їх недолік.

Мета роботи – описати прості алгоритми для ділення чисел з плаваючою комою (типу float) зі зменшеними відносними похибками обчислень.

Опис алгоритмів

Для реалізації операції ділення $y_t = a/b$ пропонується такий алгоритм (назвемо його алгоритмом 1):

```
float division_ (float a, float b)
{
int i = *(int*)&b;
i = 0x7ef33409 -i;
float y = *(float*)&i;
y= y*( 2.00128159f -b*y);
y = a*y*( 2.00000082f - b*y);
return y;
}
```

На рис. 1 показана відносна похибка цього алгоритму.

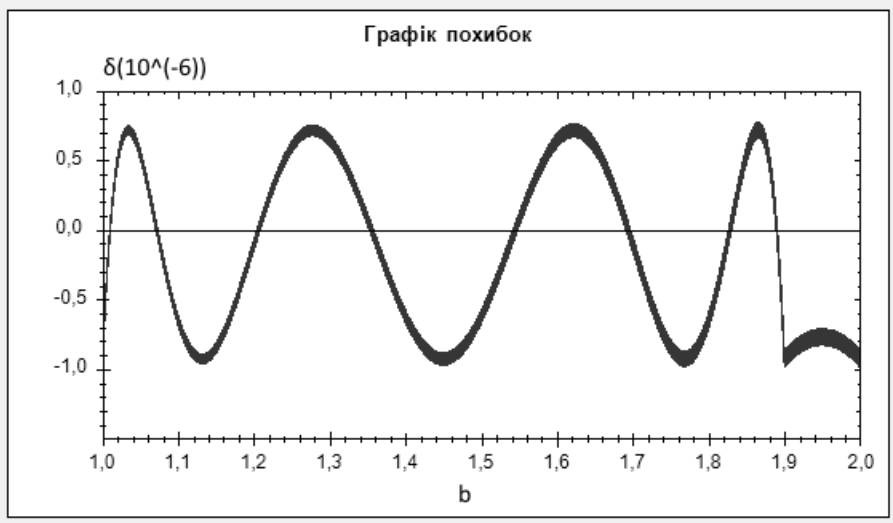

Рис. 1. Графік відносної похибки алгоритму 1 для $a=1, b\in[1,2)$

Цей алгоритм має у своєму складі лише п’ять операцій множення та забезпечує максимальну відносну похибку обчислень y_t для усього діапазону значень чисел типу float на рівні $9.84 \cdot 10^{-5}$ %, що становить 19.95 коректних бітів результату. Відмітимо, що абсолютна та відносна похибки результату (після завершення роботи алгоритму) визначаються за такими формулами:

$$D = y - y_t; \tag{1}$$

$$d = \frac{D}{y_t}. \tag{2}$$

Вищу точність має алгоритм 2:

```
float division_ (float a, float b)
{
int i = *(int*)&b;
i = 0x7EB504F3 -i;
float y = *(float*)&i;
y= y*( 2.82906784f -b*2*y);
y =a* y*( 2.0000001f - b*y);
return y;
}
```

Тут виконуються шість операцій множення та забезпечується максимальна відносна похибка обчислень $2.65 \cdot 10^{-5}$ % , що становить 21.85 коректних бітів результату.

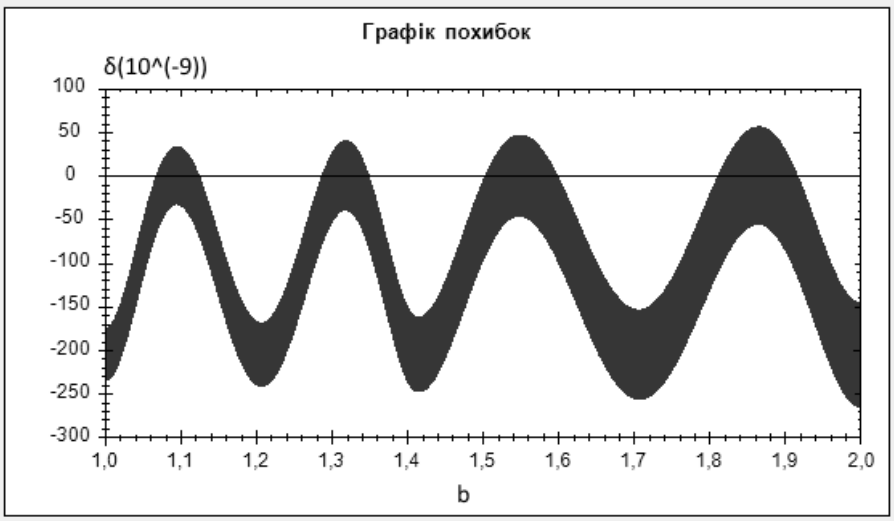

Рис. 2. Графік відносної похибки алгоритму 2 для $a=1, b\in[1,2)$

Найточнішим є алгоритм за номером 3:

```
float division_ (float a, float b)
{
int i = *(int*)&b;
i = 0x7eb504F3 -i;
float y = *(float*)&i;
y= 1.96875*y*( 1.4255685f -b*y);
y =a* y*( 2.0f - b*y);
return y;
}
```

Тут теж виконуються шість операцій множення, але максимальна відносна похибка – $1.18 \cdot 10^{-5}$ %, що становить 23.01 коректних бітів результату.

Висновки

Наведені прості алгоритми для ділення чисел з плаваючою комою (типу float) зі зменшеними похибками обчислень.

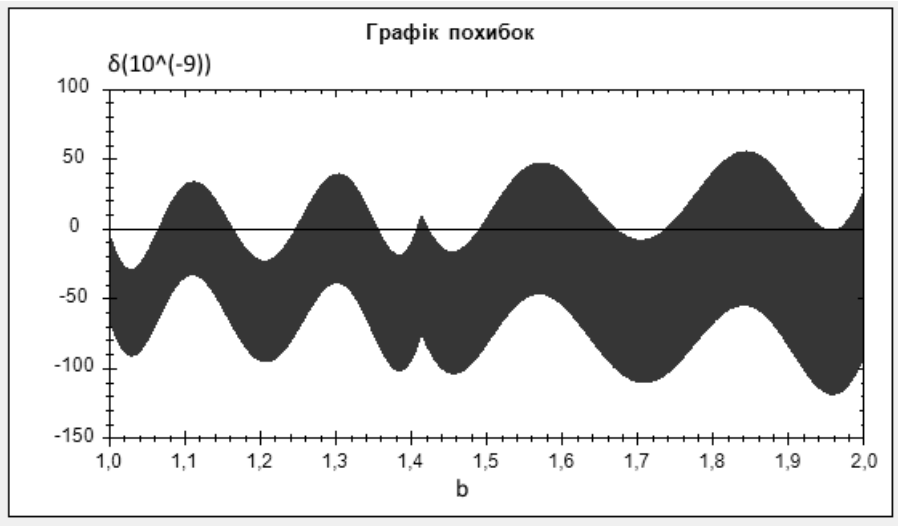

Рис. 3. Графік відносної похибки алгоритму 3 для $a=1, b\in[1,2)$

1. Minos Niu C., Sirish K. Nandyala, Won Joon Sohn, Terence D. Sanger. "Multi-scale hyper-time hardware emulation of human motor nervous system based on spiking neurons using FPGA." Advances in Neural Information Processing Systems. 2012. 2. Eric Papenhausen and Klaus Mueller. "Rapid Rabbit: Highly Optimized GPU Accelerated Cone-Beam CT Reconstruction". IEEE Medical Imaging Conference, Seoul, Korea, November 2013. 3. Jim Blinn. "Floating-point tricks". IEEE Computer Graphics and Applications, 17 (1997), no. 4. Page(s): 80 – 84. 4. Мороз Л., Гринчишин А. Швидке обчислення функції y=1/x з використанням магічної константи // Вісник НУ "Львівська політехніка" "Автоматика, вимірювання та керування". – 2015. – № 821. – С. 23–29. 5. Мороз Л. В. Теорія та швидкодіючі апаратно-програмні засоби ітераційних методів обчислення функцій // автореф. … д-ра техн. наук за спеціальністю 05.13.05 – комп'ютерні системи і компоненти. – Львів: Видавництво Львівської політехніки, 2013.