Львівський ордена Леніна державний університет імені Івана Франка
Випуск 9 Питання механіки і математики 1962

Л. Д. АДАМЕНКО, Є. С. ШОХ

ВАЖЛИВА ОСОБЛИВІСТЬ ВИКОНАННЯ АРИФМЕТИЧНИХ ОПЕРАЦІЙ У РЕЖИМІ ПЛАВАЮЧОЇ КОМИ НА МАШИНАХ ТИПУ «УРАЛ» (З ФІКСОВАНОЮ КОМОЮ)

Машини з фіксованою комою типу «Урал» працюють з числами, абсолютні величини яких менші одиниці. Тому при підготовці задачі для розв'язування на машині всі початкові дані масштабуються так, щоб проміжні і кінцеві результати обчислень були по абсолютній величині меншими одиниці. Але в практиці зустрічаються такі задачі, для яких масштабування є надзвичайно важким завданням, а в деяких випадках і зовсім неможливим.

Для розв'язування таких задач числовий матеріал перетворюють у нормальну форму (нормалізують), тобто приводять до вигляду:

$$a = q_a \cdot 10^{P_a},$$

де $\frac{1}{2} \leqslant |q_a| < 1$ — мантиса,

P_a — порядок числа a.

Арифметичні операції над числами в нормальній формі виконуються за такими алгоритмами:

1. Множення:

$$c = a \cdot b = q_a \cdot 10^{P_a} \cdot q_b \cdot 10^{P_b} = q_a \cdot q_b \cdot 10^{P_a + P_b} = q_c \cdot 10^{P_c},$$

де $\frac{1}{2} \leqslant |q_c| < 1$; $P_c = P_a + P_b + \delta$;

δ — поправка до порядку від нормалізації $q_a \cdot q_b$.

2. Додавання (віднімання):

$$c = a + b = q_a \cdot 10^{P_a} + q_b \cdot 10^{P_b} = (q_a \cdot 10^{P_a - P'} + q_b \cdot 10^{P_b - P'}) \cdot 10^{P'} =$$

$$= q_c \cdot 10^{P' + \delta} = q_c \cdot 10^{P_c},$$

де $P' = \max\{P_a, P_b\}$,

δ — поправка до порядку від нормалізації $q_a \cdot 10^{P_a - P'} + q_b \cdot 10^{P_b - P'}$.

3. Ділення:

$$c = \frac{a}{b} = \frac{q_a \cdot 10^{P_a}}{q_b \cdot 10^{P_b}} = \frac{q_a}{q_b} 10^{P_a - P_b} = q_c \cdot 10^{P_c},$$

де $P_c = P_a - P_b + \delta$,

δ — поправка до порядку від нормалізації $\frac{q_a}{q_b}$.

Кожне число, приведене до нормальної форми, записується в оперативній пам'яті (НМБ) в дві ячейки: одна ячейка (повна) — для мантиси числа, друга (неповна) — для порядку. Всі арифметичні операції над такими числами виконуються за вказаними вище алгоритмами окремо з порядками і окремо з мантисами.

Множення найпростіше реалізувати такою послідовністю команд:

$$k+1\ 02 < q_a >$$
$$k+2\ 06 < q_b >$$
$$k+3\ 15 < q_c >$$
$$k+4\ 01 < P_a >$$
$$k+5\ 01 < P_b >$$
$$k+6\ 16 < P_c >,$$

але якщо q_a або q_b рівне нулю, то і $q_c = 0$, в той час як P_c може бути, взагалі кажучи, відмінним від нуля.

Через це в обчисленнях, де є багато множень і серед початкових даних є нулі, нулі в кінцевих результатах фактично будуть мати нульові мантиси і, як правило, відмінні від нуля порядки (можуть бути навіть досить великими по модулю).

Те ж буде, якщо при діленні чисельник — нуль.

Якщо ж до такого нуля з порядком додавати якесь інше число, у якого порядок менший, ніж порядок нуля, результат може бути зовсім невірним.

Дійсно, нехай

$$a = 0 \cdot 10^{P_a},\ b = q_b \cdot 10^{P_b}$$

і нехай

$$P_a \gg P_b.$$

Тоді $a + b = 0 \cdot 10^{P_a} + q_b \cdot 10^{P_b} = (0 + q_b \cdot 10^{P_b - P_a}) \cdot 10^{P_a} = q_c \cdot 10^{P_c}$.

Якщо $|\Delta P| = |P_a - P_b| < 35$ (35 — кількість розрядів суматора «УРАЛ»-а), то

$$P_c = P_a - \Delta P = P_b,$$
$$q_c = q'_b,$$

де q'_b мантиса, у якої $|\Delta P|$ молодших розрядів будуть нулями, а $35-|\Delta P|$ старших розрядів співпадатимуть з відповідними розрядами q_b.

Якщо $|\Delta P| = |P_a - P_b| \geqslant 35$, то $P_c = P_a$, $q_c = 0$.

Коротше кажучи, коли проходить додавання якогось числа до нуля, відбувається непотрібний зсув мантиси з меншим порядком вправо на $|\Delta P|$ розрядів, додавання і потім нормалізація, внаслідок чого втрачається точність або одержується зовсім невірний результат.

Це стає очевидним, якщо прослідкувати роботу рекомендованих програм додавання в режимі плаваючої коми.

Програма наведена в книзі В. Н. Бондаренко і ін. «Программирование для цифровой вычислительной машины „Урал"»:

$$k+1\ 02 < P_a > \qquad k+\ 3\ 21\ k+15$$
$$k+2\ 03 < P_b > \qquad k+\ 4\ 10 < - >$$

$k+$ 5 17 $<0,5>$	$k+16$ 11 0000
$k+$ 6 11 0000	$k+17$ 06 $<q_a>$
$k+$ 7 06 $<q_b>$	$k+20$ 17 $<0,5>$
$k+10$ 17 $<0,5>$	$k+21$ 05 $<q_b>$
$k+11$ 05 $<q_a>$	$k+22$ 15 $<q_c>$
$k+12$ 15 $<q_c>$	$k+23$ 01 $<P_b>$
$k+13$ 01 $<P_a>$	$k+24$ 01 $<1\cdot 2^{-17}>$
$k+14$ 22 $k+24$	$k+25$ 16 $<P_c>$
$k+15$ 17 $<0,5>$	

Програма з арифметичним зсувом:

$k+$ 1 02 $<P_a>$	$k+11$ 01 $<P_a>$
$k+$ 2 03 $<P_b>$	$k+12$ 22 $k+20$
$k+$ 3 21 $k+13$	$k+13$ 17 $<q_a>$
$k+$ 4 10 $<->$	$k+14$ 11 4000
$k+$ 5 17 $<q_b>$	$k+15$ 01 $<q_b>$
$k+$ 6 11 4000	$k+16$ 15 $<q_c>$
$k+$ 7 01 $<q_a>$	$k+17$ 01 $<P_b>$
$k+10$ 15 $<q_c>$	$k+20$ 16 $<P_c>$

З метою уникнути такого неприємного явища при додаванні з нулем, що, як правило, матиме якийсь порядок, необхідно о б о в'я з к о в о аналізувати мантиси на нуль, і, якщо один з доданків має нульову мантису, за результат приймати другий доданок з його порядком. Якщо обидва доданки нулі, то, очевидно, за результат можна приймати будь-який з них.

З таким аналізом програма додавання матиме такий вигляд:

без арифметичного зсуву

$k+$ 1 02 $<q_a>$	$k+12$ 10 $<->$
$k+$ 2 15 0000	$k+13$ 17 $<0,5>$
$k+$ 3 21 $k+26$	$k+14$ 11 0000
$k+$ 4 02 $<q_b>$	$k+15$ 06 $<q_b>$
$k+$ 5 15 0000	$k+16$ 17 $<0,5>$
$k+$ 6 21 $k+16$	$k+17$ 05 $<q_a>$
$k+$ 7 02 $<P_a>$	$k+20$ 15 $<q_c>$
$k+10$ 03 $<P_b>$	$k+21$ 01 $<P_a>$
$k+11$ 21 $k+23$	$k+22$ 22 $k+32$

$k+23$ 17 $<0{,}5>$	$k+30$ 15 $<q_c>$
$k+24$ 11 0000	$k+31$ 01 $<P_b>$
$k+25$ 06 $<q_a>$	$k+32$ 01 $<1\cdot 2^{-17}>$
$k+26$ 17 $<0{,}5>$	$k+33$ 16 $<P_c>$
$k+27$ 05 $<q_b>$	

з використанням арифметичного зсуву

$k+$ 1 02 $<q_a>$	$k+14$ 11 4000
$k+$ 2 15 0000	$k+15$ 01 $<q_a>$
$k+$ 3 21 $k+23$	$k+16$ 15 $<q_c>$
$k+$ 4 02 $<q_b>$	$k+17$ 01 $<P_a>$
$k+$ 5 15 0000	$k+20$ 22 $k+26$
$k+$ 6 21 $k+15$	$k+21$ 17 $<q_a>$
$k+$ 7 02 $<P_a>$	$k+22$ 11 4000
$k+10$ 03 $<P_b>$	$k+23$ 01 $<q_b>$
$k+11$ 21 $k+21$	$k+24$ 15 $<q_c>$
$k+12$ 10 $<->$	$k+25$ 01 $<P_b>$
$k+13$ 17 $<q_b>$	$k+26$ 16 $<P_c>$

Розглянемо один з випадків, де зустрічається описана особливість додавання чисел в режимі плаваючої коми.

При розв'язуванні багатьох задач фізики і механіки, при розв'язуванні диференціальних і інтегральних рівнянь виникає потреба в знаходженні розв'язків лінійних систем алгебраїчних рівнянь. Нехай

$$AX = B,$$

де A — відома квадратна матриця порядку n,
B — відома прямокутна матриця розміром $n \times m$,
X — стовбцева матриця невідомих.

Одним з ефективних і точних методів розв'язування систем алгебраїчних рівнянь на ЦАМ є метод перехресного множення. Перетворення ведуться за формулами:

$$a^t_{ik} = \frac{a^{t-1}_{11} \cdot a^{t-1}_{i+1,\,k+1} - a^{t-1}_{1,\,k+1} \cdot a^{t-1}_{i+1,1}}{a^{t-1}_{11}}, \qquad (1)$$

$$a^t_{nk} = \frac{a^{t-1}_{1,\,k+1}}{a^{t-1}_{11}}. \qquad (2)$$

При $t = n$, $a^t_{i1} = x_i$;

$$i = 1,\ 2,\ldots n-1$$
$$t = 1,\ k = 1,\ 2,\ldots n;$$
$$t = 2,\ k = 1,\ 2,\ldots n-1;$$
$$\ldots\ldots\ldots\ldots$$
$$t = n,\ k = 1.$$

Вищевказана особливість виконання арифметичних операцій з плаваючою комою з найбільшою силою проявляється саме при розв'язуванні систем алгебраїчних рівнянь цим методом. Прямим наслідком цієї особливості є точність результатів, тобто розв'язків системи.

Матриця A може мати довільний характер, зокрема серед елементів a_{ij} матриці може бути багато нулів. Тоді багато з доданків у чисельнику формули (1) при перетвореннях можуть бути рівними нулю з великим додатним порядком і в результаті додавання результат буде невірний. В деяких випадках програма даного методу дає результати, але з невеликою точністю, якщо ж багато коефіцієнтів системи є нулями, то програма взагалі дає корені з нульовими мантисами і великими додатними порядками.

При використанні програм додавання з попередньою перевіркою мантис на нуль відносна похибка розв'язків не перевищує 0,000 001% (7—9 вірних знаків).

ЛІТЕРАТУРА

1. А. И. Китов и Н. А. Криницкий. Электронные цифровые машины и программирование. М., 1959.

2. Е. Л. Ющенко. Адресное программирование и особенности решения задач на машине «УРАЛ». К., 1960.

3. В. Н. Бондаренко и др. Программирование для цифровой вычислительной машины «УРАЛ». М., 1958.

Стаття надійшла 20. IX 1960.